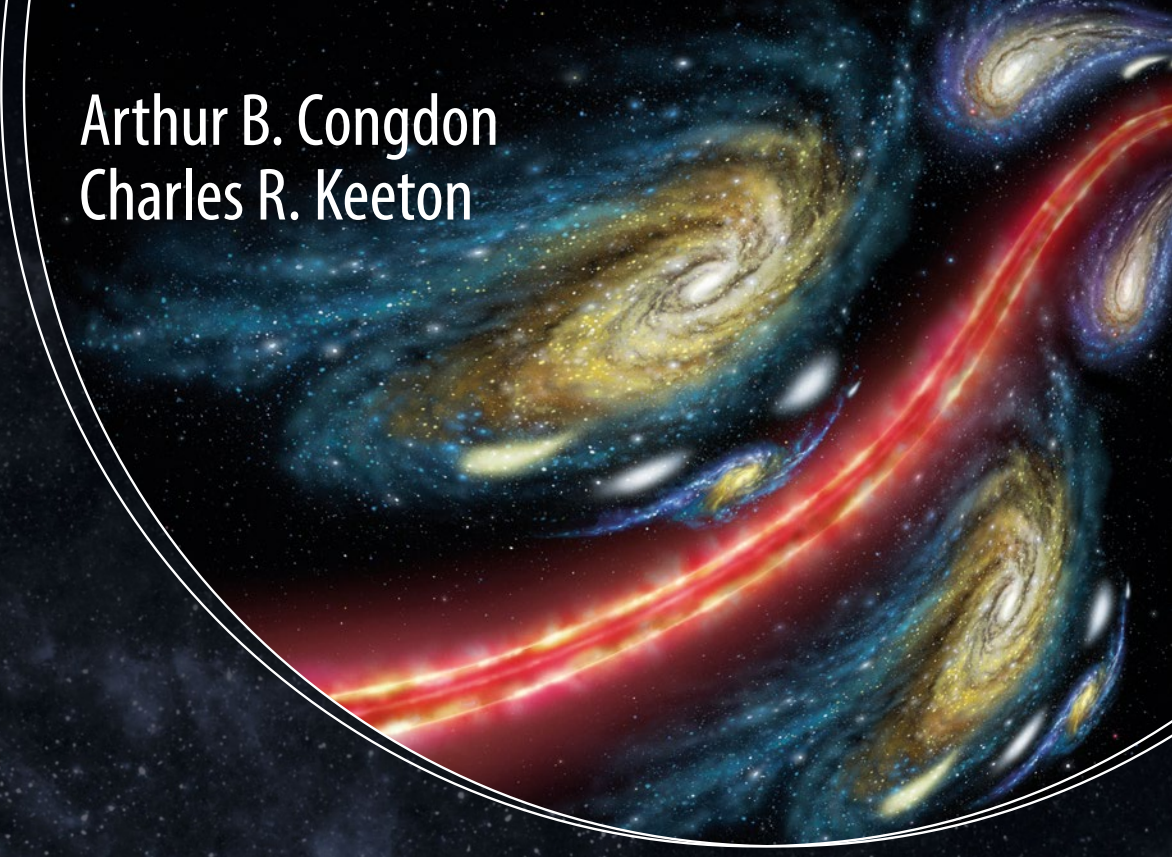
Arthur B. Congdon
Charles R. Keeton



Principles of Gravitational Lensing

Light Deflection as a Probe of Astrophysics and Cosmology

PRAXIS

Springer

Springer Praxis Books

Astronomy and Planetary Sciences

More information about this subseries at http://www.springer.com/series/4175

Arthur B. Congdon • Charles R. Keeton

Principles of Gravitational Lensing

Light Deflection as a Probe of Astrophysics and Cosmology

Springer

PRAXIS

Arthur B. Congdon
Monrovia, CA, USA

Charles R. Keeton
Physics & Astronomy Department
Rutgers University
Piscataway, NJ, USA

Springer Praxis Books
ISSN 2366-0082 ISSN 2366-0090 (electronic)
Astronomy and Planetary Sciences
ISBN 978-3-030-02121-4 ISBN 978-3-030-02122-1 (eBook)
https://doi.org/10.1007/978-3-030-02122-1

Library of Congress Control Number: 2018960262

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Gravitational lensing has matured into a thriving area of astrophysics, with applications ranging from detecting extrasolar planets (microlensing), to constraining the distribution of dark matter in galaxies (strong lensing), to determining cosmological parameters (weak lensing). While some problems are best suited to one of the three flavors of lensing (strong, weak, or micro), there are others for which it is possible to combine constraints from the different lensing regimes. The distribution of dark matter in clusters of galaxies offers one important example. For this reason, proficiency in the entire subject is indispensable to the student and researcher alike.

In their influential 1992 monograph, *Gravitational Lenses*, P. Schneider, J. Ehlers, and E.E. Falco presented the theory of gravitational lensing in a rigorous and systematic way and discussed the observations and applications then known. During the following quarter century, though, advances in instrumentation have made observations not possible at that time become almost routine today. The quality and the quantity of the data thus generated allow for sophisticated statistical analysis, making even the subtle distortions in the observed shapes of galaxies due to the large-scale structure of the universe detectable. Such weak lensing was in its infancy in the early 1990s, and microlensing had yet to be born.

A recent textbook that incorporates current research methods and applications of gravitational lensing at the undergraduate level is *Gravitational Lensing* by Scott Dodelson. His aim is to present enough gravitational lens theory so that the student can quickly confront the current literature and begin conducting research. We aim to take a complementary approach and present a thorough discussion of the *principles* of gravitational lensing. As the centenary of the observational discovery of gravitational lensing approaches in May 2019, we can only hope that the present effort is fitting tribute to researchers in the field: past, present, and future.

It is our hope that this book will prove useful to students and researchers alike. For those with prior experience in lensing, this book may serve as reference material or as a supplement for researchers who wish to explore aspects of lensing outside their own expertise. For the student, we envision this book as the basis for a one-semester course or seminar in lensing at the advanced undergraduate or beginning

graduate level. With this in mind, problems are included at the end of each chapter (apart from the Introduction) to build familiarity with lensing calculations and show how they connect with astrophysics research.

This book is organized as follow: We begin in Chap. 1 with a historical overview to offer context and background to the development of gravitational lensing. Gravitational lenses are introduced formally in Chap. 2. For the sake of gaining hands-on experience, we offer an intuitive, Newtonian presentation and make the necessary relativistic correction on the fly. The properties of gravitational lenses with circular symmetry, many of which generalize to lenses with asymmetry, are then discussed in some detail. Chapter 3 derives the fundamental equation for light bending in its full, relativistic glory and introduces the homogeneous, isotropic universe that is the backdrop for what follows in later chapters. We then turn to the theory of multiple imaging by arbitrary mass distributions in Chap. 4.

The remaining chapters explore various applications of gravitational lensing. Microlensing by stars and planets, in which multiple images cannot be spatially resolved, is presented using complex variables in Chap. 5. Strong lensing, where multiple images are resolved, is discussed in the context of galaxies in Chap. 6 and clusters of galaxies in Chap. 7. Weak lensing, in which the gravitational field is too weak to produce multiple images, is the subject of Chap. 7, where it is applied to clusters, and Chap. 8, where it is applied to large-scale structure. Chapter 9 gives an overview of lensing of the cosmic microwave background, in which we extend the methods developed in the preceding chapters.

Strictly speaking, we assume background only in multivariable calculus and introductory physics. Familiarity with intermediate classical mechanics, electrodynamics, and quantum mechanics would be useful, as much of the mathematics encountered in those subjects applies to lensing. For those without such experience, we include appendices on several topics that come up in the book: variational calculus (Appendix A), complex variables (Appendix B), orthogonal functions (Appendix C), Fourier analysis (Appendix D), and computational methods (Appendix E).

Monrovia, CA, USA — Arthur B. Congdon
Piscataway, NJ, USA — Charles R. Keeton
November 2018

Acknowledgments

This book ultimately owes its existence to the productive and enjoyable collaboration between PhD student, Arthur B. Congdon (ABC), and thesis advisor, Charles R. Keeton (CRK), a decade ago. As the influence on astrophysical research of the bending of light by gravity has only grown in the intervening years, we decided to rekindle our partnership to bring the principles of gravitational lensing to as broad an audience as possible. To this end, we have sought input from astronomers, physicists, mathematicians, and students alike.

Our own expertise has been supplemented in important ways by many people. In addition to being a friend and mentor to ABC for over 20 years, Allan Moser has offered many insights, inspired both by his work as a physicist and by his close interaction with the engineering world. Carl Droms continually reminded us that what is considered rigorous to a physicist is often hand-waving to a mathematician. His suggestions helped us to clarify our own thinking, which has translated into clearer, more concise prose throughout the text. He also contributed a number of figures that illustrate ideas presented in the text. Ted Burkhardt offered many helpful pointers on making our book accessible to physicists as well as astronomers and researchers as well as students. Several end-of-chapter problems he suggested appear in the following pages. We greatly appreciate their input throughout the duration of the writing process.

The following people read one or more chapter drafts and pointed out mathematical, scientific, stylistic, and typographical imperfections: Amir Babak Aazami, Sean Brennan, Jerod Caligiuri, John Callas, Tim Jones, Erik Nordgren, Sean Perry, Catie Raney, and Barnaby Rowe. Much of the material in this book served as the basis for a graduate seminar on gravitational lensing taught by CRK at Rutgers University. We are grateful to this inaugural group of students, who helped us translate our idea of what a lensing textbook should be into one that can actually work in practice.

ABC wishes to thank his family for their ongoing support and encouragement and Anna Göddeke for her steadfast friendship, both personal and intellectual, throughout the course of the project. CRK acknowledges support from the US National Science Foundation through grant AST-1716585.

Contents

Chapter 1
Introduction

We begin by summarizing the historical development of gravitational lensing. Our journey begins in the late eighteenth century, before the wave nature of light was definitively established (or so we thought) by Young in 1802. We then move to the early twentieth century, when the wave-particle debate resurfaced with the advent of the quantum. The particle interpretation of light is most useful in the context of lensing, since we intuitively picture gravity as affecting the motion of material objects, rather than waves. We will take this as our starting point and see where it leads. As it turns out, the wavelike properties of light can be ignored if the wavelength is much shorter than the distance scale over which a light ray is bent by gravity. This approximation holds throughout the following pages.

1.1 Light and Gravity in Newtonian Physics

Perhaps the most important idea to emerge from the work of Newton is that physical processes are governed by precise mathematical laws. Moreover, these laws apply equally to happenings on Earth and anywhere else in the universe. This led to the idea of "classical determinism," famously championed by Laplace, in which the present state of the universe determines its state at any other time, past or future. This idea naturally gives rise to a mechanistic conception of the world, where particles bump into other particles, keeping the machinery of the universe in working order.

It is against this backdrop that the effect of gravity on light was first considered. Michell (1784) and Laplace (1795) considered what would happen if a star were so dense that a particle emanating from its surface would be unable to escape. From classical mechanics, a particle will escape from a spherical body if its total energy is nonnegative, i.e.,

$$E \equiv \frac{1}{2}mv^2 - \frac{GmM}{R} \geq 0\,, \tag{1.1}$$

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_1

for a star of mass M and radius R and a test particle of mass m. If we take $v = c$, the smallest stellar radius for which a light ray can escape is

$$R_S \equiv \frac{2GM}{c^2}, \tag{1.2}$$

where c is the speed of light and R_S is what we now call the **Schwarzschild radius**. In eighteenth-century language, objects with $R < R_S$ were called "dark stars"; today, we call them black holes.

Now we turn to the bending of light by mass. Suppose that a light ray passes by a point mass with impact parameter (distance of closest approach) R. In the absence of the point mass, say, a star of mass M, the light ray will continue moving on a straight line until it reaches an observer. In the presence of the star, the light ray's trajectory will change, causing the ray to appear farther from the star than R. Since we can only measure angles on the sky rather than absolute distances, we characterize the deflection of light by the angle between its observed position on the sky and its position in the absence of gravity. We call this the deflection angle, denoted by $\hat{\alpha}$.[1]

To determine $\hat{\alpha}$, we consider what happens as R increases. We expect that $\hat{\alpha} \to 0$ as $R \to \infty$. It is natural, then, to suppose that $\hat{\alpha} \propto R^{-n}$, for some $n > 0$. For simplicity, we take $n = 1$. Our final task is to determine the proportionality constant. Since the deflection angle is dimensionless, this constant must have the dimension of length. From our earlier consideration of dark stars, the Schwarzschild radius seems like a good choice. This gives the expression

$$\hat{\alpha} = \frac{R_S}{R}. \tag{1.3}$$

As we will see in Chap. 2, this agrees with a more sophisticated Newtonian calculation, first published by Soldner (1804).

1.2 Light Bending in General Relativity

As plausible as our expression for the deflection angle may be, it relies on the assumption that light particles have nonzero mass. In arriving at the expression for the Schwarzschild radius (1.2), we canceled a factor of m in Eq. (1.1). This is incompatible with our current understanding of the photon as a massless particle. Our solution lies in the general theory of relativity, where all particles—massive or not—are affected by gravity. This is because gravity arises from spacetime curvature, to which all particles are subject, rather than action at a distance between

[1] Although a circumflex is often used to denote a unit vector, we do not use this convention here or in what follows.

massive bodies. We will show in Chap. 3 that in the weak-field limit (to be defined later), general relativity (GR) predicts a deflection angle that is *twice* the Newtonian value,[2] namely,

$$\hat{\alpha} = \frac{2R_S}{R}\,. \tag{1.4}$$

Deriving this expression theoretically was one thing, but confirming the prediction observationally was another. The main task was to find an object massive enough to bend light passing by and one or more sources of light behind this object. Since the effect would be difficult to detect, only a nearby star would work. The Sun was a natural candidate, but it is only possible to see stars behind it during a total solar eclipse. One complication is that only stars whose projected positions coincide with the solar "limb" (outer edge) can be measurably lensed. Finally, one needed to observe the relevant stars in the absence of the Sun. Comparing the positions of the lensed and unlensed stars would then yield the desired angle.

A German team, led by Erwin Freundlich, set out to Russia in 1914 to observe a total eclipse. Unfortunately, World War 1 broke out, and the team was captured behind enemy lines. They eventually returned to Germany in a prisoner exchange, but not until after the eclipse had occurred. Another team attempted to observe the same eclipse for signs of light bending was prevented by cloud cover.

Eddington, in part to avoid military service that would compromise his Quaker faith, proposed to undertake an expedition to observe the solar eclipse of 1919. Whatever result he found, its importance to science would presumably bring glory to the British Empire. This "reasoning" was rendered somewhat moot by the end of the war in 1918, but the eclipse was observed on May 29, 1919, as planned. Eddington found that Einstein's prediction was correct, although observational errors of roughly 30% were present.[3] His result was presented at a joint meeting of the Royal Society and Royal Astronomical Society late that year and published as Dyson et al. (1920). Although GR had explained Mercury's anomalous perihelion shift, light deflection was its first confirmed prediction; Einstein has been a household name ever since.

1.3 Consequences of Light Bending

The importance of light bending as a prediction that set GR apart from the prevailing theory of gravity cannot be understated. Just as the acceptance of theoretically attractive ideas like cosmic inflation and the multiverse has been hampered by a lack of testable predictions, a theory as mathematically abstruse as GR could truly be considered science only by connecting with observations in the real world. Yet

[2] Einstein's first published attempt to derive the deflection angle predicted the Newtonian value (Einstein 1911), but he detected the error before presenting GR in its final form (Einstein 1916).

[3] Subsequent measurements have lowered the uncertainty to about 0.01% (see Will 2014).

it was unclear at the time whether light bending would have a future in its own right. Indeed, Einstein (1936) dismissed the possibility that phenomena such as multiple images could actually be observed. He correctly noted that the two images characteristic of lensing by a point mass cannot be observationally resolved, at least in the stellar context he had in mind.

However, Zwicky (1937a) proposed that a *galaxy* could be observed as several distinct images if its light passed sufficiently close to a foreground cluster of galaxies. Not only do such clusters contain copious amounts of luminous matter, but Zwicky (1933, 2009) found that the Coma Cluster, for one, also required a significant amount of unseen "dark matter," in order to keep its member galaxies from flying apart. Although the vast distances to cluster lenses and background galaxies mean that only very bright sources could be seen, the magnification produced by extremely massive objects would make the probability of detecting light bending on this grand scale non-negligible (Zwicky 1937b). This insight would turn out to be prescient, though more than four decades would pass before such a cosmic lens was observed. Even individual galaxies can act as gravitational lenses since they also contain large quantities of dark matter, as discovered by Rubin and Ford (1970) for the Andromeda Galaxy and more generally for high-luminosity spiral galaxies by Rubin et al. (1978).

Zwicky's work on gravitational lensing was largely forgotten until Sjur Refsdal resurrected the field in the 1960s, though his Ph.D. thesis committee was initially skeptical that the task was worth the effort. After developing the necessary formalism (Refsdal 1964b), he proposed that the time delay between lensed images could be used to measure the Hubble constant independent of other methods (Refsdal 1964a, 1966b). This technique requires sources that vary in brightness. Since each lensed image corresponds to a particular trajectory of a light ray from the source to the observer, the associated travel time differs from one image to another. This means that a change in the flux of the source will be observed separately in each of the lensed images. Quasars and supernovae have the requisite brightness and variability and lie at cosmologically relevant distances. The first observed gravitational lens was a quasar, and a lensed supernova was eventually detected as well. These and other applications of *strong lensing* are presented in Chaps. 6 and 7. He also pointed out that observing lensed stars from widely separated locations would allow this phenomenon to be of practical use after all (Refsdal 1966a), notwithstanding the dismissal of the idea by Einstein (1936). We will have more to say on the subject of *microlensing* in Chap. 5.

1.4 Gravitational Lensing as an Observational Science

As with so many events, the historical significance of the detection of the double quasar Q0957+561 by Walsh et al. (1979) was not immediately obvious. The two "blue stellar objects" are at the same (cosmic) redshift, have remarkably similar spectra, and are separated by a mere 5.7 arcsec on the sky (see Fig. 1.1). While such

Fig. 1.1 *Hubble Space Telescope* image of the strong lens system Q0957+561. The two star-like objects in the center of the picture are lensed images of a single distant quasar, which occur because the light rays are bent by the gravity of the galaxy between the images. Credit: ESA/NASA

evidence would establish that this pair is indeed the result of a gravitational lens by today's standards, the passage of six decades from the 1919 eclipse without multiple images of the same source ever being seen led the authors to declare this as a "less conventional view." The subsequent discovery of other such systems made it clear that gravitational lensing had entered the realm of observational science. Over the years, three distinct flavors of lensing have developed into thriving subfields, which we summarize below.

1.4.1 Strong Lensing

The case in which a nearer object splits the light from a more distant source into multiple, resolved images is known as **strong lensing**. The lens and source must be closely aligned in order for this effect to occur, but strong lenses are straightforward to identify when candidates are found, at least for quasars. In the years immediately following the discovery of Q0957+561, other two-image and even four-image lenses were found, with the "Einstein cross" (Huchra et al. 1985) being a prominent example. In the following years, configurations with close pairs or triplets of images were discovered.

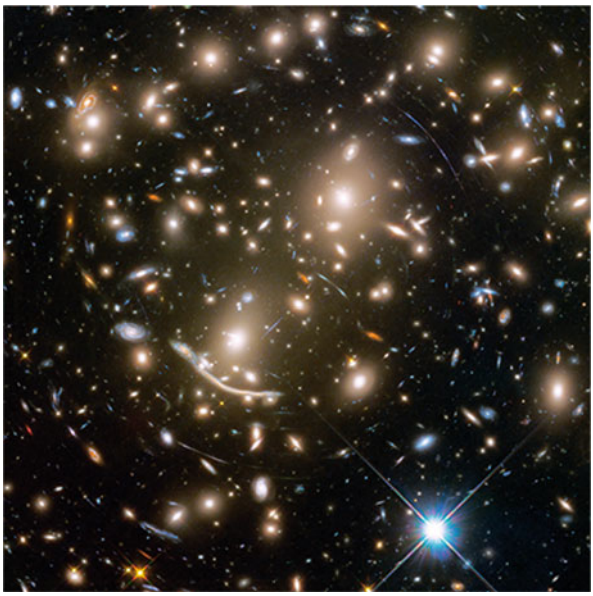


Fig. 1.2 *Hubble Space Telescope* image of the galaxy cluster Abell 370. The long structure to the lower left of the center was the first lensed arc to be discovered. Modern observations reveal a plethora of lensed images, which are discussed in Chap. 7. Credit: NASA, ESA, and J. Lotz and the HFF Team (STScI)

While lensed images of quasars are point-like, owing to their great distances from Earth, images of galaxies are extended. This means that separate lensed images can merge together, creating structures that are visually striking but may be hard to identify unambiguously as examples of lensing. A "giant luminous arc" was detected in the galaxy cluster Abell 370 (Lynds and Petrosian 1986; see Fig. 1.2), but its nature as a lensed image had to await detailed spectroscopy to be confirmed (Soucail et al. 1987a,b).

Strong lensing allows us to study the distribution of matter within galaxies and clusters of galaxies that split background sources into multiple images. Since lensing is sensitive to both dark and luminous matter, while other techniques principally constrain the luminous component, we can infer the distribution of the all-important yet equally mysterious dark matter by comparing results derived from complementary methods. It is also possible to constrain cosmological parameters by examining the lenses in their cosmological context. We discuss the theory of multiple imaging in Chaps. 2 and 4 and its applications to galaxies and galaxy clusters in Chaps. 6 and 7, respectively.

1.4.2 Microlensing

Even when several distinct images of a background source are produced by a foreground lens, they may not always be resolved. This is what led Einstein (1936) to dismiss such lensing from serious consideration. However, advances in

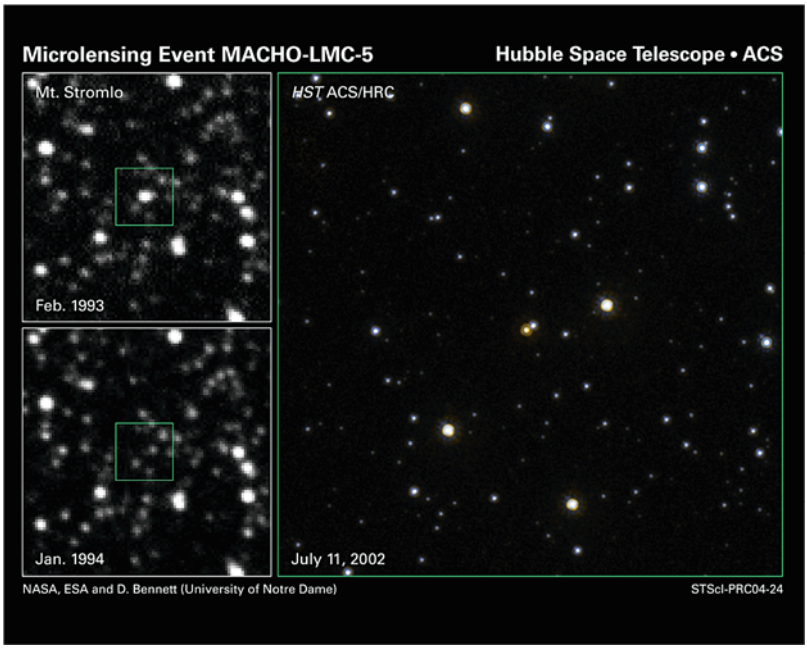


Fig. 1.3 A microlensing event. The two panels on the left show images of the same patch of sky taken from the ground at different times. The star in the middle of the small box was much brighter in Feb. 1993 than before or after, because its light was magnified when another star passed in front. The right panel shows a later *Hubble Space Telescope* observation that detected both the red lens star, which is in our Milky Way Galaxy, and the blue source star, which is in a nearby galaxy called the Large Magellanic Cloud. (The right panel spans the same area as the small boxes in the left panels.) Credit: NASA, ESA and D. Bennett (University of Notre Dame)

technology since the early days of the field opened the possibility of measuring the collective (or total) *magnification* of all the lensed images. This phenomenon is known as **microlensing**. In the case of stars, one observes the change in magnification resulting from the relative motion of the observer, lens, and source, which occurs on timescales of months or years (see Fig. 1.3).

Paczynski (1986b) suggested that dark matter in the form of compact objects in the halo surrounding the Milky Way would lead to microlensing of background stars in the Magellanic Clouds. While the probability that such a star is being lensed at a given time is quite small, there is a non-negligible probability of detecting microlensing if sufficiently many stars are regularly monitored. In fact, roughly a million stars would need to be observed for the probability to be close to unity. During the following decade, searches of this kind were undertaken, with a handful of events being detected in the Large Magellanic Cloud (Alcock et al. 1993; Aubourg et al. 1993) and in the bulge of the Milky Way (Udalski et al. 1993). These observations helped to demonstrate that *massive astrophysical compact halo objects* (MACHOs), including brown dwarfs and a variety of stellar remnants that

are invisible at large distances, exist in quantities too small to account for the motion of stars in orbit within galaxies (see Sect. 5.5.1).

Microlensing can also be used to search for extrasolar planets (exoplanets), complementing techniques based on the transit of planets in front of their host stars, or the Doppler shift of stellar light due to the gravitational pull of orbiting planets. Although planets are, of course, much less massive than the stars they orbit, they nevertheless alter the magnification in detectable and identifiable ways. Microlensing is most sensitive to planets with masses comparable to Jupiter or Saturn (e.g., Bond et al. 2004; see also Sect. 5.5.2).

In the case of lensed quasars, where it takes far longer to move a distance that is cosmologically significant, microlensing by stars within the lens galaxy can still perturb the magnification of each *macro*-image. The light rays for the different images pass through different star patterns, so microlensing of one image is uncorrelated with that of another image. Inspired by the discovery of Q0957+561, Chang and Refsdal (1979) considered perturbations to the magnification of a lensed image by a single star. Paczynski (1986a) also considered microlensing at large optical depth, where the collective effect of many stars is at play. With the help of numerical simulations, it is now possible to model the effect of many stars simultaneously. Such quasar microlensing is among the applications of strong lensing by galaxies discussed in Sect. 6.6.

1.4.3 Weak Lensing

The remaining possibility is that the shapes of background galaxies can be altered by lensing, without multiple images being formed. Unlike strong lensing and microlensing, such **weak lensing** cannot be inferred from a single object but is observable when many sources can be analyzed to yield a statistically significant signal. Objects behind the lens, even if they are significantly offset from its center, can still have their light bent, if the lens is massive enough. If the surface mass density of the lens is above a certain threshold, multiple images will form. Otherwise, a subtle distortion of the background source will be induced.

Tyson et al. (1990) observed the preferential alignment of galaxies tangential to the center of a lens cluster characteristic of weak lensing. Kaiser and Squires (1993) showed how the observed ellipticities of background galaxies, including both shape and orientation, can be used to reconstruct the surface mass density of a foreground lens cluster. Variations on this method have become the standard tools for the study of weak lensing by galaxy clusters. We describe the basic approach in Sect. 7.2

As it turns out, weak lensing can be used to probe not only individual mass concentrations but also the *large-scale structure* of the universe between the observer and source. The latter depends on cosmological parameters, such as the Hubble constant, the total energy density of the universe, and the relative contributions of radiation, matter, and dark energy. The effect of large-scale structure on the shapes of galaxies is called **cosmic shear** and was first detected about a decade

after weak lensing by clusters was discovered (Bacon et al. 2000; Van Waerbeke et al. 2000; Wittman et al. 2000). Interestingly, all of these early papers found that models *excluding* dark energy are inconsistent with the observed cosmic shear, neatly following the discovery of dark energy just a couple years earlier (Riess et al. 1998; Perlmutter et al. 1999). The fusion of advanced instrumentation and sophisticated statistical techniques has made cosmic shear an integral part of the current cosmological research.

We devote the entirety of Chap. 8 to lensing by large-scale structure, including the relevant cosmological concepts. The effect of large-scale structure on the cosmic microwave background is the subject of Chap. 9. The themes of these final chapters are related, but there are notable differences as well.

References

Alcock, C., Akerlof, C. W., Allsman, R. A., Axelrod, T. S., Bennett, D. P., Chan, S., et al. (1993). Possible gravitational microlensing of a star in the Large Magellanic Cloud. *Nature, 365*, 621.

Aubourg, E., Bareyre, P., Bréhin, S., Gros, M., Lachièze-Rey, M., Laurent, B., et al. (1993). Evidence for gravitational microlensing by dark objects in the Galactic halo. *Nature, 365*, 623.

Bacon, D. J., Refregier, A. R., & Ellis, R. S. (2000). Detection of weak gravitational lensing by large-scale structure. *Monthly Notices of the Royal Astronomical Society, 318*, 625.

Bond, I. A., Udalski, A., Jaroszyński, M., Rattenbury, N. J., Paczyński, B., Soszyński, I., et al. (2004). OGLE 2003-BLG-235/MOA 2003-BLG-53: A planetary microlensing event. *Astrophysical Journal Letters, 606*, L155.

Chang, K., & Refsdal, S. (1979). Flux variations of QSO 0957+561 A, B and image splitting by stars near the light path. *Nature, 282*, 561.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London Series A, 220*, 291.

Einstein, A. (1911). Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes. *Annalen der Physik (Berlin), 340*, 898.

Einstein, A. (1916). Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik (Berlin), 354*, 769.

Einstein, A. (1936). Lens-like action of a star by the deviation of light in the gravitational field. *Science, 84*, 506.

Huchra, J., Gorenstein, M., Kent, S., Shapiro, I., Smith, G., Horine, E., et al. (1985). 2237 + 0305: A new and unusual gravitational lens. *Astronomical Journal, 90*, 691.

Kaiser, N., & Squires, G. (1993). Mapping the dark matter with weak gravitational lensing. *Astrophysical Journal, 404*, 441.

Laplace, P. (1795). *Exposition Du Systême Du Monde*. Paris: Cercle-Social.

Lynds, R., & Petrosian, V. (1986). Giant luminous arcs in galaxy clusters. *Bulletin of the American Astronomical Society, 18*, 1014.

Michell, J. (1784). On the means of discovering the distance, magnitude, &c. of the fixed stars, in consequence of the diminution of the velocity of their light, in case such a diminution should be found to take place in any of them, and such other data should be procured from observations, as would be farther necessary for that purpose. By the Rev. John Michell, B. D. F. R. S. In a Letter to Henry Cavendish, Esq. F. R. S. and A. S. *Philosophical Transactions of the Royal Society London, 74*, 35.

Paczynski, B. (1986a). Gravitational microlensing at large optical depth. *Astrophysical Journal, 301*, 503.

Paczynski, B. (1986b). Gravitational microlensing by the galactic halo. *Astrophysical Journal, 304*, 1.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., et al. (1999). Measurements of Ω and Λ from 42 high-redshift supernovae. *Astrophysical Journal, 517*, 565.

Refsdal, S. (1964a). On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society, 128*, 307.

Refsdal, S. (1964b). The gravitational lens effect. *Monthly Notices of the Royal Astronomical Society, 128*, 295.

Refsdal, S. (1966a). On the possibility of determining the distances and masses of stars from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society, 134*, 315.

Refsdal, S. (1966b). On the possibility of testing cosmological theories from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society, 132*, 101.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., et al. (1998). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astronomical Journal, 116*, 1009.

Rubin, V. C., & Ford, W. K. Jr. (1970). Rotation of the Andromeda Nebula from a spectroscopic survey of emission regions. *Astrophysical Journal, 159*, 379.

Rubin, V. C., Ford, W. K. Jr., & Thonnard, N. (1978). Extended rotation curves of high-luminosity spiral galaxies. IV - Systematic dynamical properties, SA through SC. *Astrophysical Journal Letters, 225*, L107.

Soldner, J. (1804). Über die Ablenkung eines Lichtstrahls von seiner geradlinigen Bewegung, durch die Attraktion eines Weltkörpers, an welchem er nahe vorbei geht. *Berliner Astronomisches Jahrbuch, 1804*, 161.

Soucail, G., Fort, B., Mellier, Y., & Picat, J. P. (1987a). A blue ring-like structure, in the center of the A 370 cluster of galaxies. *Astronomy & Astrophysics, 172*, L14.

Soucail, G., Mellier, Y., Fort, B., Mathez, G., & Cailloux, M. (1987b). Discovery of the first gravitational Einstein ring - The luminous arc in Abell 370. *The Messenger, 50*, 5.

Tyson, J. A., Valdes, F., & Wenk, R. A. (1990). Detection of systematic gravitational lens galaxy image alignments - Mapping dark matter in galaxy clusters. *Astrophysical Journal Letters, 349*, L1.

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., Krzeminski, W., Mateo, M., et al. (1993). The optical gravitational lensing experiment. Discovery of the first candidate microlensing event in the direction of the galactic bulge. *Acta Astronautica, 43*, 289.

Van Waerbeke, L., Mellier, Y., Erben, T., Cuillandre, J. C., Bernardeau, F., Maoli, R., et al. (2000). Detection of correlated galaxy ellipticities from CFHT data: First evidence for gravitational lensing by large-scale structures. *Astronomy & Astrophysics, 358*, 30.

Walsh, D., Carswell, R. F., & Weymann, R. J. (1979). 0957 + 561 A, B - Twin quasistellar objects or gravitational lens. *Nature, 279*, 381.

Will, C. M. (2014). The confrontation between general relativity and experiment. *Living Reviews in Relativity, 17*, 4.

Wittman, D. M., Tyson, J. A., Kirkman, D., Dell'Antonio, I., & Bernstein, G. (2000). Detection of weak gravitational lensing distortions of distant galaxies by cosmic dark matter at large scales. *Nature, 405*, 143.

Zwicky, F. (1933). Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta, 6*, 110.

Zwicky, F. (1937a). Nebulae as gravitational lenses. *Physical Review, 51*, 290.

Zwicky, F. (1937b). On the probability of detecting nebulae which act as gravitational lenses. *Physical Review, 51*, 679.

Zwicky, F. (2009). Republication of: The redshift of extragalactic nebulae. *General Relativity and Gravitation, 41*, 207.

Chapter 2
Gravitational Lenses with Circular Symmetry

We have seen the beauty and power of gravitational lensing in both the local and distant universe, but exploiting this rich phenomenon requires a quantitative treatment. Many of the most important concepts in lensing can be gleaned from the axisymmetric case, where the mass distribution of the lens has either spherical or cylindrical symmetry in three dimensions, corresponding to circular symmetry in two dimensions. Strictly speaking, we must work within the confines of general relativity, but a Newtonian derivation of the deflection angle is nevertheless instructive (Sect. 2.1) and requires only a slight mathematical tweak to bring the result in line with relativity.

Because the radial extent of a lens is much smaller than either the distance from the observer to the lens or from the lens to the source, we can project along the line of sight to arrive at a two-dimensional problem (Sect. 2.2). Using this "thin lens" approximation, we derive the equation that a deflected light ray must obey and discuss how to determine the positions and magnifications of lensed images. We then use this framework to analyze a few simple yet representative lens models (Sect. 2.3). The final three sections derive conditions for multiple imaging by axisymmetric lenses and discuss how to extend our analysis to non-axisymmetric lenses.

2.1 Deflection Angle: Newtonian Derivation

We begin by deriving the deflection angle according to Newtonian gravity (Soldner 1804). In Chap. 1, we used a dimensional argument to suggest that the deflection angle of a point mass is given by $\hat{\alpha}(R) = R_S/R$, where $R_S = 2GM/c^2$ and R is the impact parameter of an incoming light ray. Consider a mass distribution centered at the origin of cylindrical coordinates (R, ϕ, z). The special case of a point mass is shown in Fig. 2.1. Suppose that the position of the observer is $(0, 0, D_l)$, where

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_2

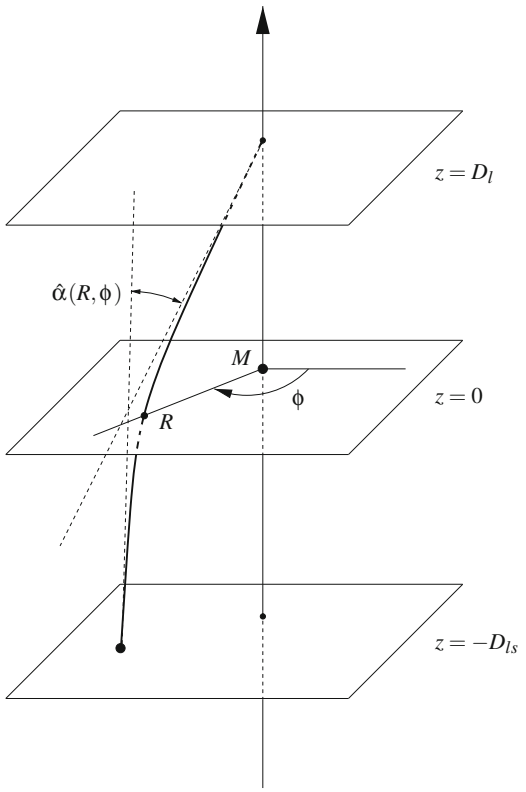


Fig. 2.1 Illustration of lensing by a point mass. A light ray is emitted by a source at $z = -D_{ls}$, is deflected by a lens of mass M at $z = 0$, and is received by an observer at $z = D_l$. Note that the azimuthal angle ϕ remains constant. In the thin lens approximation of Sect. 2.2.1, the deflection through the angle $\hat{\alpha}$ is taken to occur instantaneously at the point where the light ray crosses the lens plane. For an extended mass distribution, M refers to the mass inside the radius R

$D_l > 0$ denotes the distance between the observer and the lens. Suppose that a light ray comes in from a source located at $z = -D_{ls}$ (with $D_{ls} > 0$) and crosses the plane of the lens at a distance R from the z-axis. We take its initial velocity to be $c\mathbf{e}_z \equiv \mathbf{v}_0$, where c is the speed of light and $\mathbf{e}_z$ is a unit vector in the z-direction. Note that z increases toward the observer. The light ray will experience a gravitational acceleration due to the lens, $\mathbf{a} = -\nabla\Phi(\mathbf{r})$, where Φ is the gravitational potential (potential energy per unit mass) at position $\mathbf{r}$. This acceleration gives rise to a change in velocity as the light ray travels from the source to the observer:

$$\Delta\mathbf{v} = \int_{t_S}^{t_O} \mathbf{a}\,\mathrm{d}t \approx \frac{1}{c}\int_{-D_{ls}}^{D_l} \mathbf{a}\,\mathrm{d}z\,, \tag{2.1}$$

where the second integral is only approximate, since the path of integration is the straight line from the source to the observer rather than the actual trajectory. This is analogous to the first-order Born approximation[1] of quantum mechanics.

[1]This approximation was introduced by Lord Rayleigh (Strutt 1881) in the context of classical scattering.

In principle, we could compute a second-order correction using the first-order trajectory as input in the same way that the first-order correction is determined from the zeroth-order (unperturbed) path.

Writing $\Delta\mathbf{v} \equiv \Delta v_\parallel \mathbf{e}_z + \Delta v_\perp \mathbf{e}_R$, we can express Eq. (2.1) as separate equations for the components of velocity parallel and perpendicular to the line of sight.[2] For the parallel component, we have

$$\begin{aligned}\Delta v_\parallel &= \frac{1}{c}\int_{-D_{ls}}^{D_l} \nabla_\parallel \Phi(\mathbf{r})\,\mathrm{d}z = \frac{1}{c}\int_{-D_{ls}}^{D_l} \frac{\partial\Phi(\mathbf{r})}{\partial z}\,\mathrm{d}z \\ &= \frac{1}{c}\left[\Phi(\boldsymbol{\xi}, D_l) - \Phi(\boldsymbol{\xi}, -D_{ls})\right], \end{aligned} \tag{2.2}$$

where $\boldsymbol{\xi} \equiv (R, \phi)$. Since D_l and D_{ls} are large compared to the radial extent of the lens, and since $\Phi(\mathbf{r}) \to 0$ as $|\mathbf{r}| \to \infty$, we find $\Delta v_\parallel = 0$.

The perpendicular component of the gravitational acceleration will impart a change in velocity toward the z-axis:

$$\Delta v_\perp = \frac{1}{c}\int_{-D_{ls}}^{D_l} \nabla_\perp \Phi(\mathbf{r})\,\mathrm{d}z = \frac{1}{c}\int_{-D_{ls}}^{D_l} \frac{\partial\Phi(\mathbf{r})}{\partial R}\,\mathrm{d}z\,. \tag{2.3}$$

Thus, the light ray's final velocity will be given by

$$\mathbf{v} = \mathbf{v}_0 + \Delta\mathbf{v} = c\mathbf{e}_z + \Delta v_\perp \mathbf{e}_R \equiv v_\parallel \mathbf{e}_z + v_\perp \mathbf{e}_R\,, \tag{2.4}$$

where we assume that $v_\perp \ll v_\parallel$, so that $|\mathbf{v}| \approx c$. In fact, $|\mathbf{v}| = c\sqrt{1+\beta_\perp^2} > c$, where $\beta_\perp \equiv v_\perp/c$. Since $|\beta_\perp| \ll 1$, we have $|\mathbf{v}| = c\left[1 + \mathscr{O}\left(\beta_\perp^2\right)\right]$. Assuming the light ray's path deviates only slightly from a straight line, we can ignore terms beyond the first order in $\beta_\perp$. This conclusion admittedly involves a bit of hand-waving, but we offer a proper derivation in Chap. 3.

The relevant quantity for gravitational lensing is the angle $\hat{\alpha}$ by which a light ray is bent by a mass distribution. In terms of the velocity components $v_\parallel$ and $v_\perp$, this angle is the solution of $\tan\hat{\alpha} = \left|v_\perp/v_\parallel\right|$. Since $|v_\perp| \ll \left|v_\parallel\right|$ by assumption, we have

$$\hat{\alpha} \approx \tan\hat{\alpha} = \frac{1}{c^2}\left|\int_{-D_{ls}}^{D_l} \nabla_\perp \Phi(\mathbf{r})\,\mathrm{d}z\right|\,. \tag{2.5}$$

Since a light ray is deflected in a particular direction, we define the deflection vector $\hat{\boldsymbol{\alpha}}$. Its magnitude is twice that of the deflection angle $\hat{\alpha}$ we just computed, as

[2] Since gravity is a central force, it produces no azimuthal acceleration.

required by general relativity (see Chap. 3). The direction of $\hat{\boldsymbol{\alpha}}$ is opposite to that of $\mathbf{v}_\perp$, because a deflected light ray appears farther from the line of sight to the lens than an undeflected ray. Assuming that the lens is localized around $z = 0$, we can let D_l, $D_{ls} \to \infty$ in Eq. (2.5). The deflection vector is then

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{2}{c^2} \int_{-\infty}^{\infty} \nabla_\perp \Phi(\boldsymbol{\xi}, z)\, \mathrm{d}z\,, \tag{2.6}$$

where we write the position vector as $\mathbf{r} \equiv (\boldsymbol{\xi}, z)$, so that $\Phi(\mathbf{r}) = \Phi(\boldsymbol{\xi}, z)$.

We now specialize to the point-mass lens, for which Eq. (2.6) can be written as

$$\hat{\boldsymbol{\alpha}}(R, \phi) = \mathbf{e}_R(\phi)\, \frac{2}{c^2} \int_{-\infty}^{\infty} \left[\frac{\partial \Phi(R, z)}{\partial R} \right] \mathrm{d}z\,. \tag{2.7}$$

By writing $\boldsymbol{\xi}$ in terms of polar coordinates (R, ϕ), we can write the deflection vector as a product of its magnitude, which depends only on R, and its direction, which depends only on ϕ. We use the notation $\mathbf{e}_R(\phi)$ to indicate a unit vector in the R-direction at azimuthal angle ϕ. For a point mass M, the potential is given by

$$\Phi(R, z) = -\frac{GM}{\sqrt{R^2 + z^2}}\,. \tag{2.8}$$

We then have

$$\begin{aligned} \frac{\partial \Phi(R, z)}{\partial R} &= -GM \frac{\partial}{\partial R} \left(R^2 + z^2 \right)^{-1/2} \\ &= \frac{GMR}{\left(R^2 + z^2 \right)^{3/2}}\,. \end{aligned} \tag{2.9}$$

The deflection vector is then

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(R, \phi) &= \mathbf{e}_R(\phi)\, \frac{2GMR}{c^2} \int_{-\infty}^{\infty} \left(R^2 + z^2 \right)^{-3/2} \mathrm{d}z \\ &= \mathbf{e}_R(\phi)\, \frac{2GM}{c^2 R} \int_{-\pi/2}^{\pi/2} \cos u\, \mathrm{d}u \\ &= \mathbf{e}_R(\phi)\, \frac{4GM}{c^2 R}\,, \end{aligned} \tag{2.10}$$

where $z/R \equiv \tan(u)$. Note that this result differs from that of Chap. 1 only by the factor of 2 introduced to make our conclusion consistent with the relativistic result.

2.2 Theory of Axisymmetric Lenses

We saw in the previous section that the deflection angle does not depend on the z-coordinate, i.e., it is a vector field defined on the (x, y) plane. This result furnishes a useful physical interpretation, which is the subject of Sect. 2.2.1. We use the framework developed there to obtain an expression for the deflection angle of an axisymmetric lens in terms of a specified three-dimensional density profile. The next step is to derive the lens equation (Sect. 2.2.2), whose solutions specify where lensed images will form. A lensed image will be magnified relative to the unlensed source. We explore this topic in Sect. 2.2.3.

2.2.1 Thin Lens Approximation

Unless stated otherwise, we will treat the path of a lensed light ray as a straight line from the source to a point in the **lens plane** followed by a second straight line from this point to the observer. This is known as the **thin-lens approximation** and is equivalent to the first-order Born approximation discussed in Sect. 2.1. Starting from a specified lens model, our task is to compute the corresponding deflection angle.

In three dimensions, the gravitational field at a distance r from the center of a spherically symmetric mass distribution is proportional to the mass $M_{3\mathrm{D}}(r)$ interior to radius r. In the thin lens approximation, the mass distribution is confined to a disk, so that the deflection is affected by the mass $M_{2\mathrm{D}}(R)$ within radius R. This reasoning also applies to the deflection angle given in Eq. (2.10). Thus we replace the mass M by $M(R) \equiv M_{2\mathrm{D}}(R)$ in Eq. (2.10) for the deflection angle. We write

$$\hat{\boldsymbol{\alpha}}(R, \phi) = \frac{4G}{c^2} \frac{M(R)}{R} \mathbf{e}_R(\phi) \,, \tag{2.11}$$

where

$$M(R) = \int_0^R 2\pi R' \Sigma(R') \, \mathrm{d}R' \,. \tag{2.12}$$

(See Problem 2.1 for an alternative derivation of $M(R)$.) The surface mass density, $\Sigma(R)$, is given in terms of the volume mass density, $\rho(r)$, by

$$\Sigma(R) = \int_{-\infty}^{\infty} \rho(r) \, \mathrm{d}z \,, \tag{2.13}$$

where $r = \sqrt{R^2 + z^2}$. While the foregoing discussion says nothing new about the point-mass lens, it allows us to generalize to extended mass distributions (Sect. 2.3).

2.2.2 *Lens Equation: Geometric Derivation*

With Eq. (2.11) for the deflection angle of an axisymmetric lens in hand, we wish to determine where lensed images will form for a given source position. The result will be the lens equation. Before deriving it, let us recast Eq. (2.11) in a more useful form. Since we restrict our attention to axisymmetric lenses in this chapter, we would like to replace the deflection vector by some appropriate scalar quantity. A natural choice would be the magnitude of the deflection vector, but this would exclude the possibility that images may form on opposite sides of the lens. To see this, suppose that an image forms at an angle on the sky $\theta_+ > 0$ relative to the lens. By symmetry, any additional images must form along the line connecting the lens and the image at θ_+. If such an image appears on the other side of the lens with respect to the image at θ_+, its deflection vector will point in the opposite direction. This corresponds to the transformation $\phi \to \phi + \pi$ in Eq. (2.7), which causes the deflection vector to change sign.

We therefore define a **deflection angle**, $\hat{\alpha}$, such that $\hat{\alpha} > 0$ when $\theta > 0$ and $\hat{\alpha} < 0$ when $\theta < 0$. Specifically, we write

$$\hat{\alpha}(x) = \frac{4G}{c^2}\frac{M(|x|)}{x}\,, \tag{2.14}$$

where $x \equiv D_l\theta$. Note that $\hat{\alpha} \neq \left|\hat{\boldsymbol{\alpha}}\right|$, since $\hat{\alpha}(-x) = -\hat{\alpha}(x)$. The trade-off of switching from $\hat{\boldsymbol{\alpha}}$ to $\hat{\alpha}$ is that we suppress the azimuthal dependence in Eq. (2.11), giving the false impression that lensing by an axisymmetric mass distribution occurs on a line rather than in a plane.

For a given deflection angle $\hat{\alpha}$ and angular source position β, at what values of θ will images form? From Fig. 2.2, we have the geometric relation

$$D_s \tan\beta = D_s \tan\theta - D_{ls} \tan\hat{\alpha}\,. \tag{2.15}$$

In the small-angle approximation ($\tan\vartheta \approx \sin\vartheta \approx \vartheta$), this reduces to

$$D_s\beta = D_s\theta - D_{ls}\hat{\alpha}\,, \tag{2.16}$$

where D_{ls} and D_s are the distances[3] from lens to source and observer to source, respectively. Defining the **reduced deflection angle** by

$$\alpha(\theta) \equiv \frac{D_{ls}}{D_s}\,\hat{\alpha}(D_l\theta)\,, \tag{2.17}$$

[3]More precisely, these quantities measure angular diameter distances (see Chap. 3).

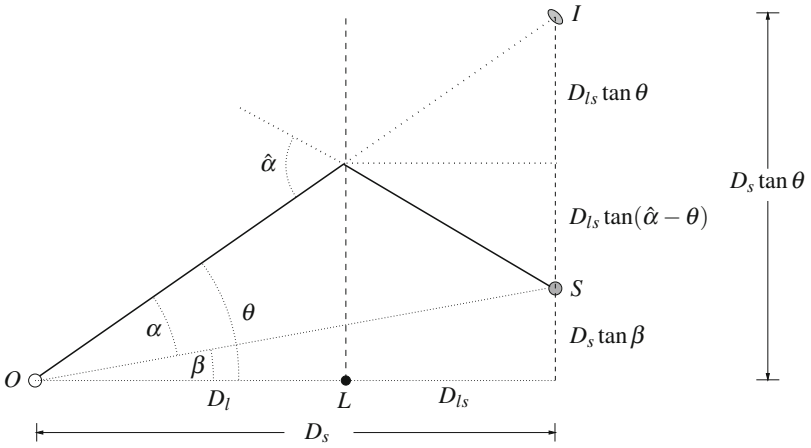


Fig. 2.2 Trajectory of a light ray subject to a weak gravitational field (solid line). The light ray emitted by source S is bent by lens L, producing an image I seen by observer O. Relative to the line of sight from O to L, the true angular position of S is β, while that of I is θ. Note that a given lens may produce multiple images and that these images are distorted as shown in Sect. 2.2.3. The angle between I and S as measured by O is denoted by α, and the angle through which the light ray is bent by L is $\hat{\alpha}$. They are related by Eq. (2.17). The distances D_l, D_s, and D_{ls} are angular diameter distances, so that $D_{ls} \neq D_s - D_l$ in general. This distinction becomes important in cosmological settings (see Chap. 3)

we finally obtain the **lens equation**:

$$\beta = \theta - \alpha(\theta)\,. \tag{2.18}$$

Images are observed at positions θ that solve the lens equation for a given source position β and deflection angle $\alpha(\theta)$. As we will see, there are often multiple solutions, implying the existence of more than one image of a single background source. We define a dimensionless mass profile:

$$m(\theta) \equiv \frac{4G}{c^2}\frac{D_{ls}}{D_l D_s} M(D_l\,|\theta|)\,, \tag{2.19}$$

so that

$$\alpha(\theta) = \frac{m(\theta)}{\theta} \tag{2.20}$$

(cf. Eqs. (2.14) and (2.17)). The lens equation can then be written as

$$\beta = \theta - \frac{m(\theta)}{\theta}\,. \tag{2.21}$$

2.2.3 *Image Magnification*

An image with position θ will be magnified by an amount $\mu(\theta)$ which satisfies

$$|\mu| = \left|\frac{\mathrm{d}\Omega_I}{\mathrm{d}\Omega_S}\right| = \left|\frac{\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi_I}{\beta\,\mathrm{d}\beta\,\mathrm{d}\varphi_S}\right| = \left|\frac{\theta\,\mathrm{d}\theta}{\beta\,\mathrm{d}\beta}\right|, \tag{2.22}$$

where $\mathrm{d}\Omega_I$ and $\mathrm{d}\Omega_S$ are elements of solid angle in the image and source planes, respectively. The sign of μ is determined by the orientation of the lensed image relative to the unlensed source: if they are oriented in the same way, then $\mu > 0$, and we say that the image has positive parity; if the orientations are opposite, then $\mu < 0$, and we say that the image has negative parity. The meaning of **parity** is a bit more involved for non-axisymmetric lenses, because the image can be inverted along either or both of the two orthogonal directions. We will return to this point in Chap. 4. Note that $\mathrm{d}\varphi_I = \mathrm{d}\varphi_S$ for an axisymmetric lens since the deflection in that case is purely radial (cf. Fig. 2.3). Because the lens equation tells us how β depends on θ, it is easier to work with the inverse magnification:

$$\mu^{-1}(\theta) = \frac{\beta}{\theta}\frac{\mathrm{d}\beta}{\mathrm{d}\theta} = \left(1 - \frac{\alpha}{\theta}\right)\left(1 - \frac{\mathrm{d}\alpha}{\mathrm{d}\theta}\right). \tag{2.23}$$

As illustrated in Fig. 2.3, we can interpret the factors in Eq. (2.23) as follows. Although β and θ measure angles on the sky, we can think of them as radial coordinates in the source and image planes, respectively. Consider a source that subtends an angle $\Delta\varphi$. Its arc length is then $\Delta s = \beta\Delta\varphi$. A lensed image of this source will have an arc length of $\Delta s' = \theta\Delta\varphi$. The magnification of this image is given by $\mu_t = \Delta s'/\Delta s = \theta/\beta$, which is the reciprocal of the first factor in Eq. (2.23). In other words, θ/β represents a tangential magnification. Now consider a source with radial extent $\Delta\beta$. The radial extent of a lensed image will then be $\Delta\theta = (\mathrm{d}\theta/\mathrm{d}\beta)\Delta\beta \equiv \mu_r\Delta\beta$, assuming that $\Delta\beta$ and $\Delta\theta$ are small. We see that μ_r is the reciprocal of the second factor in Eq. (2.23) and represents a radial magnification. The total magnification of the image is just the product $\mu_r\mu_t$.

Another feature of Eq. (2.23) that is worth noting is that the magnification diverges if either (or both) of the factors vanishes, i.e., if $\alpha = \theta$ or $\mathrm{d}\alpha/\mathrm{d}\theta = 1$. Values of θ that satisfy these relations are said to lie on **critical curves**. For axisymmetric lenses, critical curves are circles in the image plane. Under the lens equation, critical curves correspond to curves in the source plane called **caustics**. Axisymmetric lenses produce caustics that are points ($\mu_t^{-1} = 0$) or circles ($\mu_r^{-1} = 0$). Critical curves and caustics that satisfy $\alpha(\theta) = \theta$ or $\mathrm{d}\alpha(\theta)/\mathrm{d}\theta = 1$ are termed *tangential* or *radial*, respectively. As we will argue in Sect. 2.5.1 for axisymmetric lenses, and extend in Chap. 4 to more general lenses, the number of images of a given lens increases or decreases by two when a source crosses a caustic.

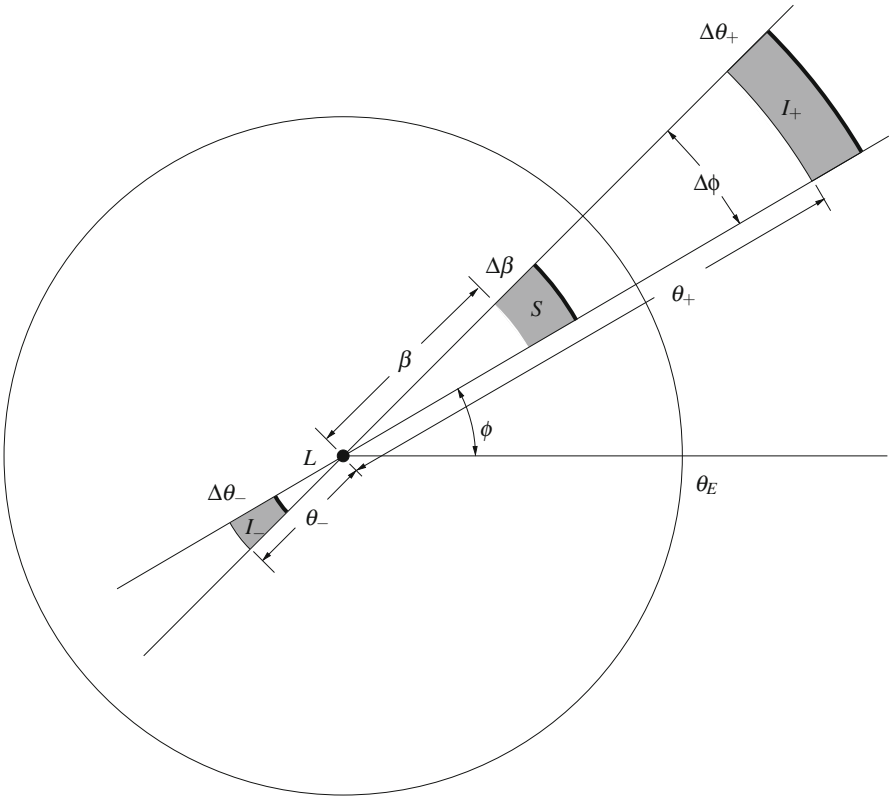


Fig. 2.3 Magnification of lensed images. In the plane of the sky, a source S at angular distance β from lens L appears as images $I_\pm$ at distances $\theta_\pm$. Note that the angle ϕ is arbitrary. The (infinitesimal) area of S is $\beta\Delta\beta\Delta\phi$, while that of $I_\pm$ is $\theta_\pm\Delta\theta_\pm\Delta\phi$. This generally gives rise to tangential and radial magnification as discussed in Sect. 2.2.3. However, the singular isothermal sphere (Sect. 2.3.2), from which the present figure is generated, only produces tangential magnification. When $\beta = 0$, an "Einstein ring" of angular radius θ_E is formed

Before considering a few specific lens models, we note that β is an odd[4] function of θ, so that β/θ and $\mathrm{d}\beta/\mathrm{d}\theta$ are even functions of θ. Thus, $\mu(\theta)$ is also even. This means that the magnification of an image produced by an axisymmetric lens depends only on its distance from the center of the lens, as we would expect. Equivalently, magnification of a source depends only on its (projected) distance from the lens. This is in contrast to the deflection angle, which points away from the lens and hence is an odd function of image position.

[4]Recall that a function $f(x)$ is said to be odd if $f(-x) = -f(x)$ and even if $f(-x) = f(x)$.

2.3 Axisymmetric Lens Models

We now turn to a few specific lens models to illustrate how one solves the lens equation and computes image magnification in practice. In the axisymmetric case we are considering, the lens equation is effectively one-dimensional. This restriction notwithstanding, the models we present demonstrate the variety of lens systems that even the simplest mass distributions can produce. Allowing for asymmetry is essential for understanding observed lenses, but including such complexity comes at a high algebraic cost. We will confront this painful reality in due time, but our focus in this section is on the qualitative behavior common to axisymmetric and non-axisymmetric lenses alike.

2.3.1 Point Mass Lens

The simplest lens consists of a single point mass, whose deflection angle is given by

$$\alpha = \frac{D_{ls}}{D_s}\frac{4GM}{c^2 R} = \frac{D_{ls}}{D_l D_s}\frac{4GM}{c^2\theta} \tag{2.24}$$

(cf. Eqs. (2.19) and (2.20)). Defining the Einstein radius as

$$\theta_E \equiv \sqrt{\frac{D_{ls}}{D_l D_s}\frac{4GM}{c^2}}\,, \tag{2.25}$$

the lens equation takes the form

$$\beta = \theta - \frac{\theta_E^2}{\theta}\,. \tag{2.26}$$

For a given β, there are two solutions:

$$\theta_\pm = \frac{\beta \pm \sqrt{\beta^2 + 4\theta_E^2}}{2}\,. \tag{2.27}$$

Notice that $\theta_+ > 0$ and $\theta_- < 0$ for all $\beta > 0$. This means that one image will form on the same side of the lens as the source, and another will form on the opposite side of the lens. This conclusion remains valid if $\beta < 0$; only the signs of $\theta_\pm$ will be reversed. For convenience, we take $\beta > 0$ in what follows, unless stated otherwise. Since we do not know the true source position, β, how can we determine which image is which? The answer lies in the magnification, given for a point-mass lens by

$$\mu^{-1}(\theta) = \left(1 - \frac{\theta_E^2}{\theta^2}\right)\left(1 + \frac{\theta_E^2}{\theta^2}\right) = \left(1 - \frac{\theta_E^4}{\theta^4}\right) \tag{2.28}$$

(cf. Eq. (2.23)). As we will see in Problem 2.2, $|\theta_-| < \theta_E$ and $|\theta_+| > \theta_E$. This implies that $\mu_- \equiv \mu(\theta_-) < 0$ and $\mu_+ \equiv \mu(\theta_+) > 0$ for all $\beta > 0$. In other words, the image at θ_+ has positive parity while that at θ_- has negative parity.

In the case of Galactic[5] microlensing, to be discussed in Chap. 5, the individual images cannot be spatially resolved. It is therefore useful to write down the total magnification produced by both images, since that can be observed. Let us first write the magnification in terms of the image position in units of the Einstein radius, $x \equiv \theta/\theta_E$:

$$\mu(x) = \frac{x^4}{x^4 - 1}\,. \tag{2.29}$$

The total magnification is then

$$\begin{aligned} \mu_{\rm tot} &\equiv |\mu_+| + |\mu_-| \\ &= \frac{x_+^4}{x_+^4 - 1} - \frac{x_-^4}{x_-^4 - 1}\,. \end{aligned} \tag{2.30}$$

In terms of the source position in units of the Einstein radius, $u \equiv \beta/\theta_E$, we have:

$$\mu_{\rm tot} = \frac{u^2 + 2}{u\sqrt{u^2 + 4}}. \tag{2.31}$$

Note that $\mu_{\rm tot}$ represents a change in brightness of a source caused by lensing, so it is not strictly a magnification; the latter would also contain parity information.

The **centroid** (center of light) of the two images is located at

$$\begin{aligned} \bar{x} &= \frac{x_+\,|\mu_+| + x_-\,|\mu_-|}{\mu_{\rm tot}} \\ &= u\left(1 + \frac{1}{u^2 + 2}\right). \end{aligned} \tag{2.32}$$

Thus, the shift of the centroid position due to lensing is

$$\Delta\bar{x} = \bar{x} - u = \frac{u}{u^2 + 2} \tag{2.33}$$

[5] We write Galaxy with a capital "G" when referring to the Milky Way galaxy.

(Paczyński 1998 and references therein), which vanishes in the limit $u \to \infty$, where there is no lensing. The centroid will play a role in our discussion of microlensing in Chap. 5.

Let us now consider the case of a source that is directly behind the lens. Reverting to angular positions β and θ, this corresponds to $\beta = 0$, $\theta_\pm = \pm\theta_E$, and $\mu_\pm \to \infty$. What does it mean for images to form at $\pm\theta_E$? Do they form to the left and right of the lens, above or below it, or on some other line? Although we have been treating the lens equation for an axisymmetric lens as one-dimensional, this is not quite correct. The proper interpretation is that the azimuthal angular position of the source is not affected by purely radial light deflection. This means that the source and lensed images are collinear, with the orientation of the line being determined by the azimuthal position of the source. For a source at $\beta = 0$, all lines passing through β and $\theta_\pm$ are equivalent, so that a ring of angular radius θ_E, known as an **Einstein ring**, is formed. Note that $\beta = 0$ is a caustic point, whose (tangential) critical curve is the Einstein ring of the lens.

2.3.2 *Singular Isothermal Sphere*

A point mass is a reasonable approximation for a star but is too crude to describe galaxies, which are an important class of lenses. The simplest plausible such model is known as the **singular isothermal sphere** (SIS). Its key virtue is that it predicts a flat rotation curve (Fig. 2.4), which is observed at large radii in disks of spiral

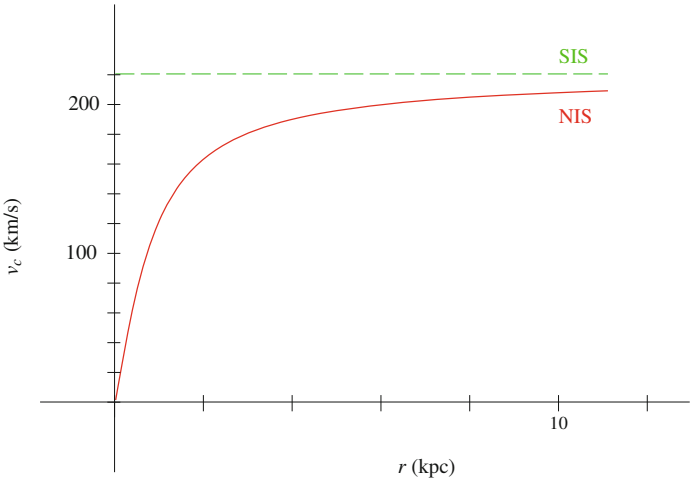


Fig. 2.4 Galactic rotation curves for two common density profiles. An SIS produces constant rotational speed as a function of galactocentric distance. An NIS has a flat rotation curve at large radii, but deviations from its asymptotic value are seen at small radii. We have chosen a typical maximum rotation speed of $v_{\rm rot} = 220\,{\rm km/s}$

galaxies. One drawback of the SIS, however, is that the density is infinite at its center. This singularity can be removed by introducing a small, finite-density core, as described in Sect. 2.3.3.

Physical Motivation

It is often useful to picture a galaxy as a ball of gas where the pressure arising from particle collisions is replaced by the orbital motion of stars and **dark matter**. This motion balances the galaxy's gravitational field at all radii. Assuming our gas sphere (galaxy) can be described locally by the ideal gas law, we have

$$P(r) = n(r)kT = \frac{kT}{m}\rho(r)\,, \tag{2.34}$$

where P, n, and ρ are the pressure, number density, and mass density, respectively. A galaxy with a flat rotation curve is analogous to a gas of constant temperature, T. For simplicity, we assume all particles (stars) have the same mass, m.

Of course, collisionless stars and particles of dark matter in galaxies do not really have a temperature but are instead described by a characteristic speed, known as the velocity dispersion, σ. In terms of the temperature, this quantity is given by $\sigma = \sqrt{kT/m}$. Hence, the ideal gas law takes the form

$$P(r) = \sigma^2\rho(r)\,. \tag{2.35}$$

A lens model entails a density profile, so our task is to find $\rho(r)$. To eliminate P, we invoke the equation of **hydrostatic equilibrium**, which says that pressure balances gravity. It reads

$$\boldsymbol{\nabla} P = -\rho\boldsymbol{\nabla}\Phi\,, \tag{2.36}$$

which reduces to

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\rho\frac{\mathrm{d}\Phi}{\mathrm{d}r} \tag{2.37}$$

under the assumption of spherical symmetry. Finally, we use Poisson's equation to write the density in terms of the potential:

$$4\pi G\rho(r) = \nabla^2\Phi(r) = \frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\Phi}{\mathrm{d}r}\right)\,. \tag{2.38}$$

Combining Eq. (2.37) with the derivative of Eq. (2.35) yields

$$\frac{\mathrm{d}\Phi}{\mathrm{d}r} = -\sigma^2\frac{\mathrm{d}\ln\rho}{\mathrm{d}r}\,. \tag{2.39}$$

Plugging this result into Poisson's equation (2.38) gives

$$-\frac{4\pi G\rho}{\sigma^2} = \frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\ln\rho}{\mathrm{d}r}\right) . \tag{2.40}$$

Following Binney and Tremaine (2008), we hypothesize that $\rho(r) = Ar^{-\alpha}$, and thus

$$\frac{4\pi G A r^{-\alpha}}{\sigma^2} = \alpha r^{-2} . \tag{2.41}$$

In order for this equation to hold at all radii, we require $\alpha = 2$. It is then trivial to solve for A. The SIS profile is therefore

$$\rho(r) = \frac{\sigma^2}{2\pi G}\frac{1}{r^2} . \tag{2.42}$$

A spiral galaxy consists of a rotating disk embedded in a halo of dark matter. We obtain the rotation speed of the disk by equating centripetal and gravitational acceleration, i.e., $v_{\mathrm{rot}}^2(r) = r\Phi'(r)$. For the SIS, we find $v_{\mathrm{rot}} = \sqrt{2}\,\sigma$, a constant!

In the case of an elliptical galaxy, however, stellar orbits are randomly oriented, so the quantity v_{rot} is not meaningful. Instead, we measure the distribution of the component of velocity parallel to the line of sight to the galaxy (Fig. 2.5). For an ideal gas, this distribution is Gaussian, and its standard deviation is the velocity

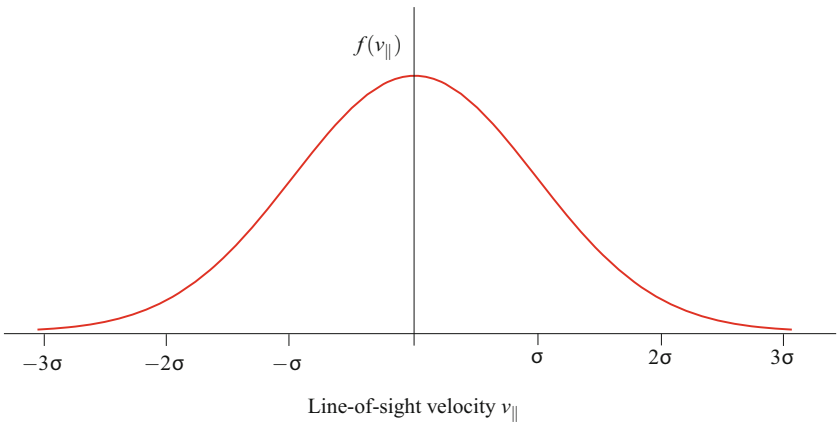


Fig. 2.5 Line-of-sight velocity distribution $f(v_\parallel)$. Stars moving away (toward) the observer have positive (negative) values of $v_\parallel$. Assuming isotropic stellar motion, we find $\langle v_\parallel \rangle = 0$. A Gaussian with mean zero and velocity dispersion (standard deviation) $\sigma \sim 150$ km/s, corresponding to $v_{\mathrm{rot}} \sim 220$ km/s, provides a good fit to observed spiral galaxies. Elliptical galaxies are typically more massive than spirals, with correspondingly greater values of σ

dispersion σ. Specifically, if $f(v_\parallel)\mathrm{d}v_\parallel$ is the fraction of stars with line-of-sight velocity between $v_\parallel$ and $v_\parallel + \mathrm{d}v_\parallel$, we have

$$f(v_\parallel) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-v_\parallel^2}{2\sigma^2}\right) . \tag{2.43}$$

Even for a spiral galaxy, where the stars are concentrated in a disk, the function $f(v_\parallel)$ has physical significance: Ostriker and Peebles (1973) showed that disks are subject to large-scale instability if they are not embedded in a three-dimensional halo. The absence of a visible halo is further evidence for dark matter in galaxies. The velocity dispersion is a measure of a galaxy's total mass, so knowing its value and the total amount of luminous material allows us to infer the abundance of dark matter. As we will now see, the SIS gives rise to multiple images of a sufficiently aligned background source, offering an additional probe of galactic structure.

Lensing Properties

Regarding the SIS as a lens, our first task is to compute the deflection angle. As we saw in Sect. 2.2.1, this requires specification of a surface mass density profile. Integrating Eq. (2.42) along the line of sight yields

$$\begin{aligned} \Sigma(R) &= \frac{\sigma^2}{2\pi G} \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{R^2 + z^2} \\ &= \frac{\sigma^2}{2\pi G R} \int_{-\pi/2}^{\pi/2} \mathrm{d}u \\ &= \frac{\sigma^2}{2G} \frac{1}{R} , \end{aligned} \tag{2.44}$$

where $z \equiv R \tan u$. From Eq. (2.12), the mass enclosed within radius R is then

$$M(R) = \frac{\sigma^2 \pi R}{G} . \tag{2.45}$$

Substituting this into Eq. (2.11), the deflection vector for an SIS is given by

$$\hat{\boldsymbol{\alpha}}(R, \phi) = 4\pi \frac{\sigma^2}{c^2} \mathbf{e}_R(\phi) . \tag{2.46}$$

Note that the magnitude of the deflection angle is independent of R. This property greatly simplifies the computation of the positions and magnifications of the images produced by an SIS. An alternative derivation of Eq. (2.46) is considered

in Problem 2.3. In terms of image position θ, the reduced deflection angle of the SIS is given by

$$\alpha(\theta) = \theta_E \frac{|\theta|}{\theta}\,, \tag{2.47}$$

where

$$\theta_E \equiv 4\pi \frac{\sigma^2}{c^2} \frac{D_{ls}}{D_s}\,. \tag{2.48}$$

The lens equation therefore takes the form

$$\beta = \theta - \theta_E \frac{|\theta|}{\theta}\,. \tag{2.49}$$

Solving the lens equation takes some care because absolute values are involved. If $\theta > 0$, we find that

$$\theta = \beta + \theta_E\,. \tag{2.50}$$

For $\theta < 0$, we have

$$\theta = \beta - \theta_E\,, \tag{2.51}$$

which gives $\theta < 0$ only if $\beta < \theta_E$. When $\beta = 0$, an Einstein ring with angular radius θ_E is formed.

Like the point mass, the SIS produces a caustic point at $\beta = 0$, which maps to a critical curve (Einstein ring) in the image plane. For both models, the Einstein ring is parametrized by θ_E, but it is important to note that the specific value of θ_E depends on the assumed model. One qualitative difference between the two models is that the point mass always produces two images of a background source, while the number of images produced by an SIS depends on the source position. In the latter case, a source will be doubly imaged only if it is within the Einstein radius of the lens.

In the general situation, the number of images changes by two as the source crosses a curve along which the magnification diverges. For the SIS, on the other hand, the number of images changes by one as the source crosses the curve along which the magnification vanishes:

$$\begin{aligned} \mu^{-1}(\theta) &= 1 - \theta_E \frac{|\theta|}{\theta^2} \\ &= 1 - \frac{\theta_E}{|\theta|}\,, \end{aligned} \tag{2.52}$$

so that $\mu = 0$ when $\theta = 0$ or, equivalently, when $\beta = \theta_E$. The circle defined by $\beta = \theta_E$ is known as a **pseudocaustic**.

In addition to the change in brightness of an image, it is of interest to ascertain its change in shape caused by lensing. Since $\mathrm{d}\alpha/\mathrm{d}\theta \equiv \alpha'(\theta) = 0$ for $\theta \neq 0$, the SIS does not produce a radial caustic or critical curve. The only way for an image to appear radially elongated is if the tangential magnification is small. The SIS does produce tangential arcs for sources near $\beta = 0$, corresponding to $|\theta| \sim \theta_E$.

2.3.3 Nonsingular Isothermal Sphere

While the SIS has the desirable property of producing a flat rotation curve, it is far from perfect. For one thing, it is spherical, which is only a rough approximation to real galaxies. We will deal with this issue in Chap. 4. We focus here on the more fundamental problem that the SIS has infinite density at its center. We can address this shortcoming by introducing a finite-density core with radius r_c. The result is the **nonsingular isothermal sphere**[6] (NIS), with density

$$\rho(r) \equiv \frac{\sigma^2}{2\pi G} \frac{1}{r^2 + r_c^2}. \tag{2.53}$$

Note that $\rho \propto r^{-2}$ when $r \gg r_c$ and is roughly constant for $r \ll r_c$. This means that the NIS yields a flat rotation curve only for $r \gg r_c$, which is consistent with observation.

Projecting along the line of sight, we have

$$\Sigma(R) = \frac{\sigma^2}{2G} \frac{1}{\sqrt{R^2 + r_c^2}}. \tag{2.54}$$

The mass within (projected) radius R is then

$$\begin{aligned} M(R) &= \frac{\pi\sigma^2}{G} \int_0^R \frac{R'\mathrm{d}R'}{\sqrt{R'^2 + r_c^2}} \\ &= \frac{\pi\sigma^2}{G} \left(\sqrt{R^2 + r_c^2} - r_c \right). \end{aligned} \tag{2.55}$$

[6] This model is also referred to as the *softened isothermal sphere*, but this leads to an ambiguous abbreviation.

The dimensionless mass profile (2.19) is consequently

$$m(\theta) = \frac{4G}{c^2}\frac{D_{ls}}{D_l D_s} M(D_l\,|\theta|) = 4\pi\,\frac{\sigma^2}{c^2}\,\frac{D_{ls}}{D_s}\left(\sqrt{\theta^2+\theta_c^2}-\theta_c\right), \tag{2.56}$$

where $\theta_c \equiv r_c/D_l$. The deflection angle can then be written as

$$\alpha(\theta) = \frac{m(\theta)}{\theta} = \frac{\theta_E^{\rm SIS}}{\theta}\left(\sqrt{\theta^2+\theta_c^2}-\theta_c\right), \tag{2.57}$$

where $\theta_E^{\rm SIS} \equiv 4\pi\sigma^2 D_{ls}/(c^2 D_s)$ is the Einstein radius of the corresponding SIS (i.e., the NIS with $\theta_c = 0$). Given this deflection angle, the lens equation for the NIS takes the form

$$\beta = \theta - \frac{\theta_E^{\rm SIS}}{\theta}\left(\sqrt{\theta^2+\theta_c^2}-\theta_c\right). \tag{2.58}$$

Defining $u = \beta/\theta_E^{\rm SIS}$, $x = \theta/\theta_E^{\rm SIS}$, and $a = \theta_c/\theta_E^{\rm SIS}$, we can write

$$u = x - \frac{1}{x}\left(\sqrt{x^2+a^2}-a\right). \tag{2.59}$$

We replace the reduced deflection angle α by the function

$$\tilde{\alpha}(x) \equiv \frac{1}{x}\left(\sqrt{x^2+a^2}-a\right), \tag{2.60}$$

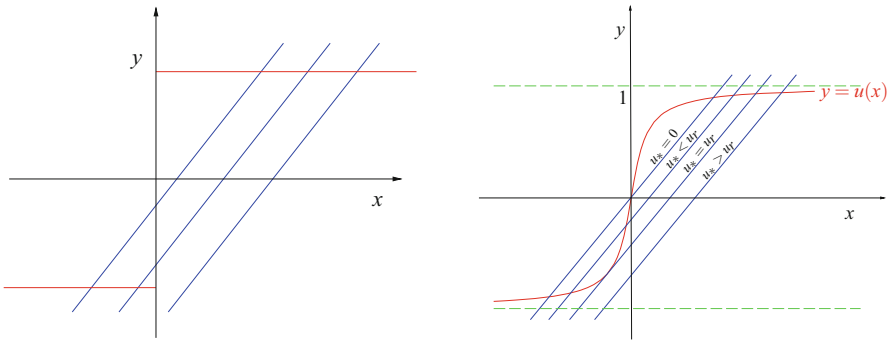
so that Eq. (2.59) can be written as $u = x - \tilde{\alpha}(x)$.

Graphical Solution of the Lens Equation

Unlike the point mass and SIS, it is not so simple to invert the lens equation for the NIS to find the image positions x as a function of the source position, u. We could solve for x numerically for particular values of u and a. However, our purpose is to highlight the qualitative differences and similarities between the NIS and the other models we have considered so far. To this end, we rewrite the lens equation in the more suggestive form

$$\tilde{\alpha}(x) = x - u, \tag{2.61}$$

so that the solutions can be seen as the intersections of the curve $y = \tilde{\alpha}(x)$ and the lines $y = x - u$ for different values of u (Figs. 2.6 and 2.7).



(a) SIS (b) NIS-src

Fig. 2.6 Deflection curves for the singular isothermal sphere (**a**) and the nonsingular isothermal sphere with core radius $a = 0.2$ (**b**), together with lines $y = x - u$ for various values of u

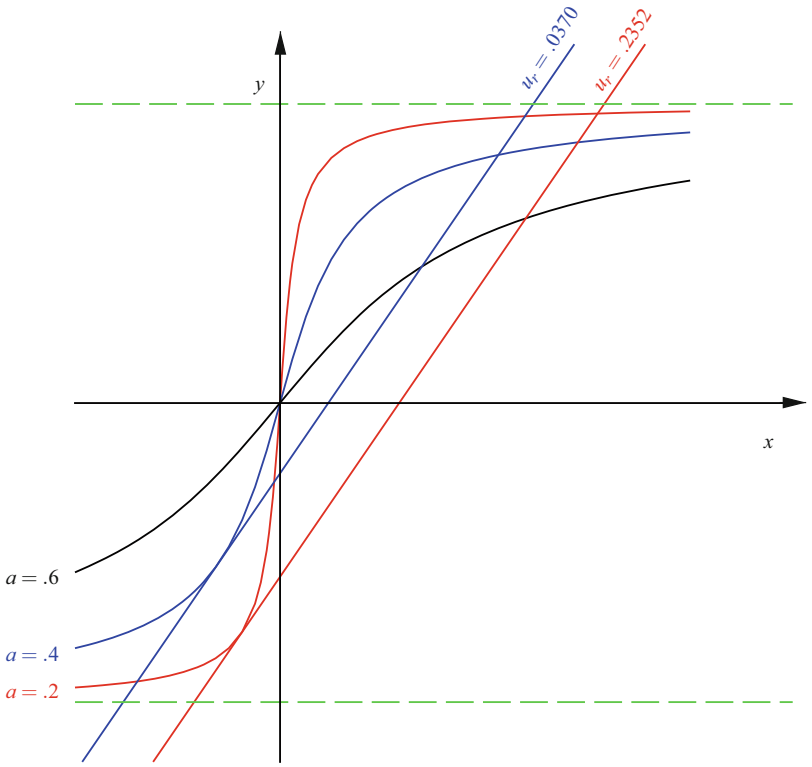


Fig. 2.7 Deflection curves $y = \tilde{\alpha}(x)$ for several values of a, together with the corresponding lines $y = x - u_r$. Note that if $a > 0.5$, the slope of the curve $y = \tilde{\alpha}(x)$ is less than one everywhere, so there are never three intersections and consequently no solution for u_r

Comparing the two panels of Fig. 2.6 shows that the jump discontinuity at $x = 0$ in the deflection curve for the SIS does not occur on the curve for the NIS. This opens the possibility of three lensed images forming for suitable values of u. Note that, as before, we consider only positive values for the source position. For the SIS ($a = 0$), $u = 1$ marks the boundary beyond which there is only one image. For the NIS, however, the location of this boundary, denoted by u_r, will depend of the value of a. Finding this boundary analytically is the focus of Problem 2.5.

We have established that introducing a finite-density core results in a number of qualitative differences with the SIS model, but we have yet to determine the effect of a particular choice of core radius a. We show three illustrative examples in Fig. 2.7. As a increases, the radial caustic shrinks, i.e., u_r decreases. Notice that there is no solution for u_r when $a > 0.5$. Put another way, the slope of the curve $y = \tilde{\alpha}(x)$ must exceed unity at the origin in order for $u_r > 0$ to exist (cf. Problem 2.4).

The size of the central core affects not only the location of the radial caustic but also the positions of the lensed images. For instance, a source at $u = 0$ will produce an Einstein ring whose radius, x_E, satisfies $x_E = \tilde{\alpha}(x_E)$. From Eq. (2.60), we find

$$x_E = \sqrt{1 - 2a}\,. \tag{2.62}$$

Notice that x_E becomes imaginary for $a > 1/2$, which is consistent with our earlier finding that only one image can form for such large cores. After all, an Einstein ring would be a two-image system were it not for the axial symmetry of the NIS lens. Moreover, we will see in Sect. 2.5 that the ability of a lens to produce an Einstein ring is a necessary and sufficient condition for the creation of multiple images. For a source with position $u \neq 0$, introducing a central core produces a third lensed image and alters the positions of the two images formed by an SIS (see Problem 2.6 for more).

Analytic Approach

Solving the lens equation graphically is intuitive and straightforward to implement computationally, but it is of interest to approach the problem more formally. To that end, we recast the lens equation in the form

$$x^3 - 2ux^2 + (u^2 + 2a - 1)x - 2ua = 0\,. \tag{2.63}$$

Since the coefficients of this cubic equation are real, it has either three real solutions, or one real solution and two complex solutions that are conjugate. Since complex image positions are not physically meaningful, the NIS produces one or three images, depending on the specific values of u and a. We know from the graphical solution of the lens equation described above that for $a < 1/2$, there is a critical value u_r that separates the source plane into regions with one ($|u| > |u_r|$) or three ($|u| < |u_r|$) images. The number of real solutions of a cubic equation is determined

by the sign of its discriminant: there are three real roots if the discriminant is positive and one if it is negative. Thus, u_r is the value of u for which the discriminant vanishes. Unfortunately, in transforming the lens equation (2.59) into the cubic equation (2.63), we introduced a spurious solution for u_r. This issue is explored in Problem 2.7.

To avoid the difficulty of spurious image positions, we return to lens equation (2.59) to locate the radial caustic. An exact solution can be obtained, but an approximate expression is often both convenient and sufficient. One such example is an NIS with a small central core ($a \ll 1$). The first step is to find the critical curve. To do this, we determine the value of x that satisfies

$$\frac{\mathrm{d}u}{\mathrm{d}x} = 0\,, \tag{2.64}$$

which we denote by x_r. The critical curve then has radius $|x_r|$. Expanding to first order in a, we have $\sqrt{x^2 + a^2} = x + \mathscr{O}\left(a^2\right)$. Substituting this into Eq. (2.59) gives

$$u \approx x - 1 + \frac{a}{x}\,, \tag{2.65}$$

whence $\mathrm{d}u/\mathrm{d}x \approx 1 - a/x^2$. Therefore, the critical curve has radius $|x_r| \approx \sqrt{a}$. The radial caustic then has radius $|u_r| = |u(x_r)| \approx |1 - 2x_r| = 1 - 2\sqrt{a}$.

Thus, the location of the radial caustic for the NIS differs from the location of the pseudocaustic for the SIS by $2x_r$. It is worth noting at this point that nonsingular lenses capable of producing multiple images give rise to true caustics, where the magnification of a source is formally infinite; pseudocaustics can only arise for lenses with divergent central surface density.

2.4 Einstein Rings

A common feature of the lens models presented in Sect. 2.3 is that they all produce both multiple images and Einstein rings for suitable parameter values. Thus, one may suspect a more general connection between multiple images and Einstein rings. Indeed, any axisymmetric lens that has an Einstein radius produces multiple images when the lens and source are sufficiently aligned. In this section, we derive the condition under which Einstein rings form. We connect this result to the existence of multiple images in Sect. 2.5.

Setting $\beta = 0$ in Eq. (2.21), we see that there is an Einstein ring if the condition

$$\frac{m(\theta)}{\theta^2} = 1 \tag{2.66}$$

is satisfied for some $\theta \equiv \theta_E > 0$. Writing the left-hand side of Eq. (2.66) in terms of the surface mass density using Eqs. (2.12) and (2.19), we obtain

$$\begin{aligned}\frac{m(\theta)}{\theta^2} &= \frac{4G}{c^2}\frac{D_{ls}}{D_l D_s}\frac{M(D_l\,|\theta|)}{\theta^2} \\ &= \frac{4G}{\theta^2 c^2}\frac{D_{ls}}{D_l D_s}\int_0^{D_l|\theta|} 2\pi R\,\Sigma(R)\mathrm{d}R \\ &= \frac{4\pi G}{c^2}\frac{D_l D_{ls}}{D_s}\left(\frac{1}{\pi\theta^2}\int_0^{|\theta|} 2\pi\theta'\,\Sigma(D_l\theta')\mathrm{d}\theta'\right), \end{aligned} \tag{2.67}$$

where $R \equiv D_l\theta'$. On defining the **critical density for lensing** by

$$\Sigma_{\mathrm{crit}} \equiv \frac{c^2}{4\pi G}\frac{D_s}{D_l D_{ls}} \tag{2.68}$$

and the **convergence** by

$$\kappa(\theta) \equiv \frac{\Sigma(D_l\,|\theta|)}{\Sigma_{\mathrm{crit}}}, \tag{2.69}$$

Eq. (2.67) can be written as

$$\frac{m(\theta)}{\theta^2} = \frac{1}{\pi\theta^2}\int_0^{|\theta|} 2\pi\theta'\kappa(\theta')\mathrm{d}\theta' \equiv \bar{\kappa}(\theta), \tag{2.70}$$

where $\bar{\kappa}(\theta)$ is the **mean convergence** interior to $|\theta|$. The condition (2.66) for an Einstein ring then becomes

$$\bar{\kappa}(\theta_E) = 1. \tag{2.71}$$

Thus, a source directly behind a lens appears as an Einstein ring if and only if there is some angle θ_E that encloses a disk with a mean convergence of unity. Note that the convergence depends not only on the assumed mass distribution of the lens but also on the distances between lens, source, and observer. The same is true for quantities derived from the convergence, including the value of θ_E.

For realistic lens models, both κ and $\bar{\kappa}$ decrease as $|\theta|$ increases and approach zero for $|\theta| \to \infty$. Moreover, one can show in Problem 2.9 that $\bar{\kappa} > \kappa$ for $\theta \neq 0$. We then conclude (see Fig. 2.8) that $\bar{\kappa}(\theta_E) = 1$ for some $\theta_E > 0$ if and only if $\kappa(0) > 1$. A lens with $\kappa(0) > 1$ is said to be **supercritical**. Thus, an Einstein ring can form if and only if the lens is supercritical.

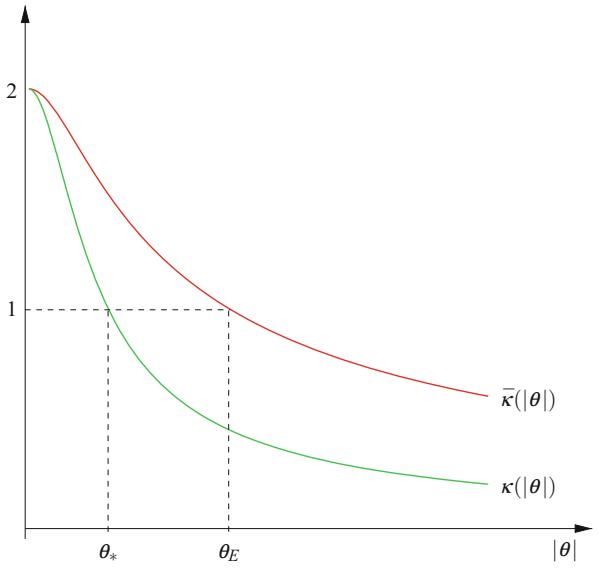


Fig. 2.8 Local convergence κ and mean convergence $\bar{\kappa}$ as functions of angular radius $|\theta|$ for a nonsingular isothermal sphere with $\kappa(0) = 2$. Note that $\kappa(\theta_*) = 1$ and $\bar{\kappa}(\theta_E) = 1$

2.5 Supercriticality and Strong Lensing

Having demonstrated that an Einstein ring forms if and only if a lens is supercritical, we now show that supercriticality is a necessary and sufficient condition for an axisymmetric lens to produce multiple images. This condition remains sufficient in the non-axisymmetric case, but it is not necessary (see Chap. 4).

For convenience we define the function

$$f(\theta) \equiv \theta - \alpha(\theta)\,, \tag{2.72}$$

which is known as the **lens mapping**. Using Eqs. (2.20) and (2.70), we can express the deflection angle in terms of the mean convergence as

$$\alpha(\theta) = \bar{\kappa}(\theta)\theta\,, \tag{2.73}$$

and hence the lens mapping takes the form

$$f(\theta) = [1 - \bar{\kappa}(\theta)]\theta\,. \tag{2.74}$$

Another important quantity is the derivative of the lens mapping

$$\begin{aligned} f'(\theta) &= 1 - \bar{\kappa}(\theta) - \theta\bar{\kappa}'(\theta) \\ &= 1 - 2\kappa(\theta) + \bar{\kappa}(\theta)\,, \end{aligned} \tag{2.75}$$

where we have used Eq. (2.94) in Problem 2.9. One simple but useful consequence of this relation is that

$$\alpha'(\theta) = 2\kappa(\theta) - \bar{\kappa}(\theta)\,. \tag{2.76}$$

The lens equation is solved for a given source position β_0 at values of θ where the horizontal line $\beta = \beta_0$ intersects the lens mapping $\beta = f(\theta)$. Since $f(\theta)$ is an odd function, we may assume $\beta_0 \geq 0$ without loss of generality. We will see in Sect. 2.5.1 that the number of images changes at extremal points of $f(\theta)$. Multiple images can still form even if $f(\theta)$ has no local extrema, but this involves discontinuous lens mappings (Sect. 2.5.2). As we will see, yet another possibility is that the lens mapping has local extrema but is nevertheless non-differentiable at the origin (Sect. 2.5.3).

2.5.1 *Differentiable Lens Mappings*

Suppose that the lens is supercritical. Then, there is some θ_0 for which $\kappa(\theta_0) > 1$, assuming continuity. Furthermore, one can show in Problem 2.10 that

$$\lim_{\theta\to 0} \bar{\kappa}(\theta) = \kappa(0)\,. \tag{2.77}$$

Hence, we find

$$f'(0) \;=\; 1 - 2\kappa(0) + \bar{\kappa}(0) \;=\; 1 - \kappa(0) \;<\; 0\,, \tag{2.78}$$

where the last step follows from our assumption of supercriticality.

As $|\theta|$ gets large, the deflection of a light ray approaches a constant value. This means that $f'(\theta) \to 1$ as $|\theta| \to \infty$. Since $f'(-\infty) > 0$ and $f'(0) < 0$, there must be at least one intermediate value θ_0 on the interval $-\infty < \theta_0 < 0$ such that $f'(\theta_0) = 0$ and $f''(\theta_0) < 0$, i.e., θ_0 is a local maximum of $f(\theta)$. Thus, the value $\beta_0 \equiv f(\theta_0)$ marks a transition such that a source with $\beta > \beta_0$ has two fewer images than a source with $\beta < \beta_0$ (see Fig. 2.9). Hence, supercriticality implies multiple images for at least some source positions. In the language of Sect. 2.2.3, $\theta = \theta_0$ defines a radial critical curve with corresponding caustic $\beta = \beta_0$.

We now show that a lens capable of producing multiple images is supercritical. By assumption $f'(\theta_0) = 0$ for some θ_0. Equation (2.75) then reduces to

$$\kappa(\theta_0) - 1 = \bar{\kappa}(\theta_0) - \kappa(\theta_0) > 0\,. \tag{2.79}$$

Therefore, the lens is supercritical at θ_0. Summing up, we have shown that a lens will produce multiple images if and only if it is supercritical, at least in the case of a differentiable convergence field.

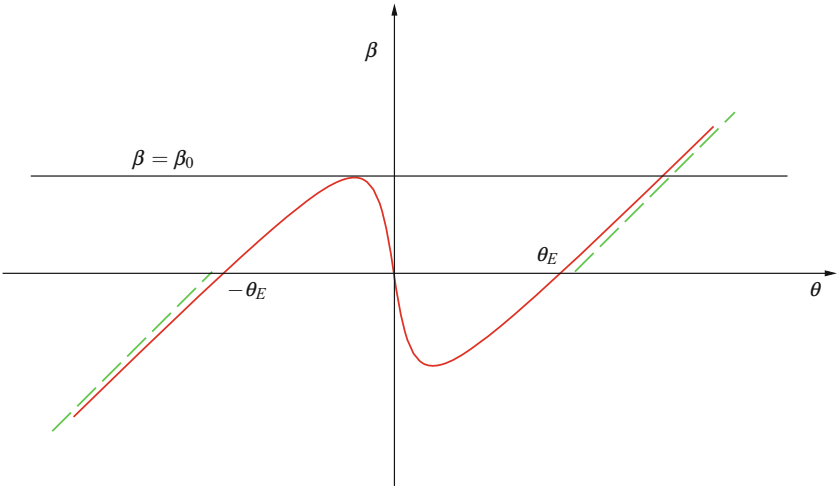


Fig. 2.9 Example of a continuous lens mapping $\beta = f(\theta)$ that produces one image when $\beta > \beta_0$ and three images when $0 < \beta < \beta_0$. An Einstein ring of angular radius θ_E appears when $\beta = 0$. Note that the slope $f'(\theta)$ of the lens mapping asymptotically approaches unity

2.5.2 *Discontinuous Lens Mappings*

Let us now turn to the case that the lens mapping is discontinuous at the origin. We need not consider the possibility of a discontinuity elsewhere since, according to Eq. (2.74), it would imply a nonphysical discontinuity in the mean convergence $\bar{\kappa}$. Note that the lens mapping is continuous at the origin as long as $\bar{\kappa}$ diverges less rapidly than $|\theta|^{-1}$.

Suppose that the deflection angle has the power-law form[7]

$$\alpha(\theta) = A\,|\theta|^{-\eta}\,\frac{|\theta|}{\theta}\,, \tag{2.80}$$

where $\eta \geq 0$ and $A > 0$ are constants, and the final factor ensures that the deflection points away from the lens. In the following, we work in units where $A = 1$; alternatively, we can ignore A, since its value will prove irrelevant anyway. We can now write the mean convergence as

$$\bar{\kappa}(\theta) = |\theta|^{-\eta}\,\frac{|\theta|}{\theta^2}\,. \tag{2.81}$$

[7]This assumption holds for many, but not all, lens models. Fortunately, the conclusions of this and the following subsection remain true in the general case.

Since $\bar{\kappa}(\theta)$ is even, we can take $\theta \geq 0$ without loss of generality. This gives us the simple relation

$$\bar{\kappa}(\theta) = \theta^{-(\eta+1)} . \tag{2.82}$$

In order to obtain the local convergence, we must compute the derivative of the mean convergence:

$$\bar{\kappa}\,'(\theta) = -(\eta+1)\theta^{-(\eta+2)} . \tag{2.83}$$

Equation (2.94) then yields

$$\kappa(\theta) = \frac{1}{2}(1-\eta)\theta^{-(\eta+1)} . \tag{2.84}$$

Since $\kappa > 0$, we must have $\eta \leq 1$. On the interval $0 \leq \eta \leq 1$, the convergence diverges at the origin, and hence a lens is supercritical if the deflection is discontinuous at the origin.

The lens mappings shown in Fig. 2.10 are those of the SIS and point mass, corresponding, respectively, to the boundary cases $\eta = 0$ and $\eta = 1$. If we allowed $\eta < 0$, we would find $\alpha(\theta)$ continuous at $\theta = 0$, which we have already considered. The constraint $\eta \leq 1$ tells us that no lens can be more concentrated than a point mass. The qualitative behavior of the lens mapping differs for $\eta = 0$ and $\eta \neq 0$, so we consider these cases separately. Our subsequent discussion assumes that the lens mapping has no local extrema, since we have dealt with that possibility above.

When η vanishes, $\alpha(\theta)$ is piecewise constant, with a jump discontinuity at the origin. This translates to a lens mapping that is piecewise linear with a jump discontinuity at the origin. In addition to the postulated power-law form for the deflection near $\theta = 0$, we have the asymptotic boundary condition $f\,'(\theta) \to 1$ as $|\theta| \to \infty$. For convenience, we define $f_+(\theta)$ and $f_-(\theta)$ to be identical to $f(\theta)$ for $\theta > 0$ and $\theta < 0$, respectively, and to be undefined elsewhere. Since $f(\theta)$ is continuous on the intervals $-\infty < \theta < 0$ and $0 < \theta < +\infty$, and has no local extrema, the functions $f_\pm(\theta)$ are invertible on their respective domains. The function $f_-(\theta)$ reaches its maximum at 0^- (i.e., just to the left of the β-axis), which we denote by β_0. Note that β_0 is manifestly positive and that $f_-(\theta)$ takes on all values on the closed interval between 0 and β_0. Now, $f_+(0^+) = -f_-(0^-) \equiv -\beta_0$, since $f(\theta)$ is odd. The function $f_+(\theta)$ strictly increases and takes on all values between $-\beta_0$ and ∞. We therefore conclude that for all β satisfying $0 \leq \beta \leq \beta_0$, there exist $\theta_- < 0$ and $\theta_+ > 0$ such that $f_-(\theta_-) = f_+(\theta_+) = \beta$, i.e., there are two images for $0 \leq \beta \leq \beta_0$. Note that the images with $\beta = 0$ will be seen as an Einstein ring, due to symmetry. When $\beta > \beta_0$, there will be a single value of θ for which $f(\theta) = \beta$, resulting from the branch f_+, i.e., there is just one image for $\beta > \beta_0$.

Now we turn to the case $0 < \eta \leq 1$. The jump discontinuity at the origin for $\eta = 0$ is replaced by a vertical asymptote collinear to the β-axis. We again separate

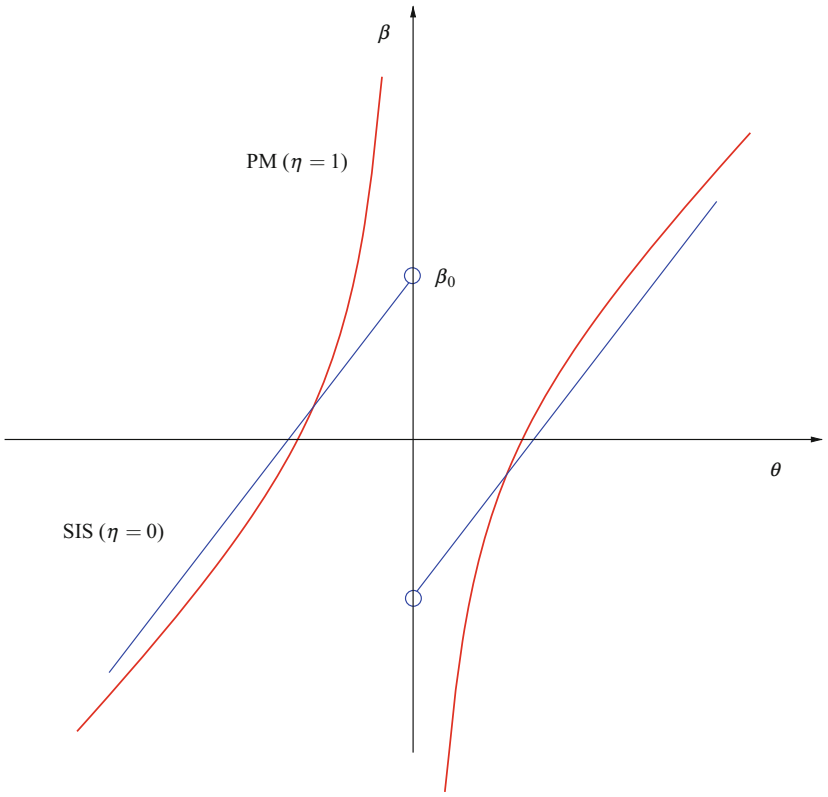


Fig. 2.10 Examples of lens mappings $\beta = f(\theta)$ that are discontinuous at the origin. The lens mapping undergoes a jump discontinuity at $\theta = 0$ for a deflection curve with central (power-law) slope $\eta = 0$. Such lens mappings have the peculiar property that the line $\beta = \beta_0$ marks a change in image number of only one. We plot the SIS lens mapping, for which $\eta = 0$ everywhere. Deflection curves with central slope $0 < \eta \leq 1$ diverge to $\pm\infty$ as $\theta \to 0^{\mp}$, respectively. In the case that such a lens has no local extrema, two images will be produced regardless of source position. We plot the lens mapping of the point mass (PM), for which $\eta = 1$ everywhere

$f(\theta)$ into f_+ and f_- branches. This time, however, $f_{\pm}(0^{\pm}) = \mp\infty$. The analysis then proceeds as before but with $\beta_0 \to \infty$. The upshot is that there are two images for all $\beta \geq 0$.

2.5.3 Divergent Convergence Fields

Before proceeding, let us take stock. We have shown that a lens that can produce multiple images is supercritical, whether the lens mapping is continuous (Sect. 2.5.1) or discontinuous (Sect. 2.5.2). We have also seen in Sect. 2.5.1 that a

supercritical lens with finite central convergence can create multiple images of a source. We have yet to demonstrate, however, that a supercritical lens with divergent central convergence will also produce multiple images; we have only done so in the case that $\kappa(\theta)$ diverges at least as rapidly as $|\theta|^{-1}$.

From Eq. (2.84), we find that the convergence will diverge at the origin whenever $\eta > -1$. We noted on physical grounds that $\eta \leq 1$, as larger values make the convergence negative. In the range $0 \leq \eta \leq 1$, we showed that the deflection is divergent ($\eta \neq 0$) or discontinuous ($\eta = 0$) at the origin. While $\alpha(\theta)$ is continuous at $\theta = 0$, its derivative may not be. We have

$$\alpha'(\theta) = -\eta A \theta^{-(\eta+1)}\,, \tag{2.85}$$

where we set A to unity as before. Strictly speaking, the above expression only holds for $\theta > 0$, but the antisymmetry of $\alpha(\theta)$, and corresponding symmetry of $\alpha'(\theta)$, about $\theta = 0$ allows us to generalize to $\theta < 0$. When $\eta > 0$, $\alpha'(\theta) < 0$, for all θ. Hence, $\alpha(\theta)$ has no local extrema in this case, which prompted our analysis in Sect. 2.5.1. When $\eta < 0$, however, $\alpha'(\theta) > 0$ for all θ. In particular, $\alpha'(0) \to \infty$. Combining this result with the asymptotic boundary condition $\alpha'(\infty) = 0$, we find that $\alpha'(\theta_*) = 1$ for some $\theta_* > 0$, provided that $\alpha'(\theta)$ is continuous for positive θ (see Problem 2.11). In other words, θ_* is a local extremum of the lens mapping, implying multiple images for at least some source positions.

To recap, we have now proved that supercriticality is a necessary and sufficient condition for multiple imaging by axisymmetric lenses. This is a special case of the condition for multiple imaging by arbitrary thin lenses. We will point the way to this general condition in the next section, but a derivation is out of reach at this point. We will take that up in Chap. 4.

2.6 Magnification and Shear

While we have focused so far on the local and mean convergence, the condition for strong lensing by an arbitrary mass distribution can be best approached through the magnification. We begin by writing

$$\begin{aligned}\mu^{-1} &= \left(1 - \frac{\alpha}{\theta}\right)\left(1 - \frac{\mathrm{d}\alpha}{\mathrm{d}\theta}\right) \\ &= (1 - \bar{\kappa})(1 - 2\kappa + \bar{\kappa})\,. \end{aligned} \tag{2.86}$$

If either of these factors is zero, there are multiple images of the background source. In the case that the first factor $1 - \bar{\kappa}$ vanishes, an Einstein ring appears. If the second factor which is just $f'(\theta)$ vanishes, the lens mapping has a local extremum at θ. Conversely, if a lens produces multiple images, it must be supercritical ($\kappa > 1$)

for some θ. This implies that the first factor, and perhaps also the second, is zero somewhere. Thus, strong lensing occurs if and only if the magnification diverges for some θ.

To understand the meaning of this statement more clearly, it is instructive to express each factor in terms of the local convergence and what is known as **shear**, defined by

$$\gamma \equiv \bar{\kappa} - \kappa \,, \tag{2.87}$$

which is nonnegative for all θ (cf. Fig. 2.8). The inverse magnification is then

$$\mu^{-1} = (1 - \kappa - \gamma)(1 - \kappa + \gamma) = \lambda_- \lambda_+ \,, \tag{2.88}$$

where $\lambda_\pm \equiv (1 - \kappa) \pm \gamma$. If $\gamma = 0$, then $\lambda_+ = \lambda_-$, implying that the source is magnified isotropically. We can therefore think of the shear as the cause of anisotropic magnification. In the same way that an extended body is deformed by tidal forces, a nonuniform gravitational field distorts the apparent shape of the light source. (We will explore this topic further in Chap. 4.)

In the case of circular symmetry, λ_- can only vanish at the origin and never becomes negative. Moreover, $\lambda_+ = 0$ for some θ_* if, and only if, $\lambda_- = 0$ at the origin. These conclusions do not extend to noncircular lenses, but the concept of shear does. However, shear cannot be described solely by the nonnegative scalar γ but instead must be thought of as a tensor, with γ being understood as a magnitude. For that matter, magnification itself turns out to be a tensor quantity in the general case, with $\lambda_\pm^{-1}$ being its eigenvalues. Multiple images will form whenever a given source position corresponds to $\lambda_\pm < 0$. For a source far from the lens in projection, $\lambda_\pm \to 1$, i.e., no magnification takes place ($\mu \to 1$). In order for strong lensing to occur, there must be a set of source positions for which the magnification is formally infinite ($\lambda_\pm = 0$). Note that if $\lambda_+ < 0$ at a point, then $\lambda_- < 0$ at the same location. The necessary and sufficient condition for multiple imaging is therefore $\lambda_- = 0$.

Problems

2.1 For a spherically symmetric mass distribution, one can compute the mass $M_{2\mathrm{D}}(R)$ within an infinite cylinder C_R of radius R directly from the mass $M_{3\mathrm{D}}(r)$ within a sphere of radius r. Begin by writing

$$M_{2\mathrm{D}}(R) = \int_{C_R} \rho(\mathbf{x}) \mathrm{d}^3 x \,.$$

Since ρ depends only on $r \equiv |\mathbf{x}|$, it is convenient to perform the integral in spherical coordinates (r, θ, ϕ). Find the limits of integration appropriate for the cylinder

C_R centered on the z-axis, with respect to which θ is measured. Then make the substitution $u \equiv \cos(\theta)$, and show that

$$M_{2D}(R) = \int_0^1 M_{3D}\left(\frac{R}{\sqrt{1-u^2}}\right) \mathrm{d}u \,. \tag{2.89}$$

This result provides a different way to derive Eq. (2.45), and it will be useful in Sect. 7.1.1 when we study galaxy clusters.

2.2 Consider light from a source at angular position β lensed by a point mass with Einstein radius θ_E.

(a) Show that $|\theta_-| < \theta_E$ and $|\theta_+| > \theta_E$ for all $\beta > 0$. Hint: A proof by contradiction may be most natural.
(b) Show that $\mu_+ \equiv \mu(\theta_+) \to 1$ and $\mu_- \equiv \mu(\theta_-) \to 0$ as $\beta \to \infty$. This means that a source sufficiently far from the lens is equivalent to an unlensed source.

2.3 Consider a singular isothermal sphere (SIS) with velocity dispersion σ.

(a) Show that the mass enclosed within a sphere of radius r is given by $M_{3D}(r) = 2\sigma^2 r/G$.
(b) Using Gauss's law, compute the gravitational acceleration at radius r.
(c) Find the corresponding deflection angle using Eq. (2.7). Does this expression agree with Eq. (2.46) for the SIS?
(d) Show that the convergence and shear are identical.
(e) Compute the total magnification as a function of the source position, β. Does the asymptotic behavior of $\mu_{\rm tot}$ as $\beta \to 0$ and $\beta \to \infty$ make sense?

2.4 If $\tilde{\alpha}'(x) < 1$ everywhere, there will be no possibility of multiple images for a nonsingular isothermal lens.

(a) Show that $\tilde{\alpha}'(x)$ reaches its maximum at $x = 0$.
(b) Use this information to show that only a single image can form when $a > 1/2$.

2.5 We saw in Fig. 2.6 that the number of images produced by a nonsingular isothermal sphere changes when the line $y = x - u$ is tangent to the deflection curve, $y = \tilde{\alpha}(x)$. We denote the value of the source position at which this transition occurs by u_r.

(a) Show that the point of tangency occurs at the value of x for which the derivative $\tilde{\alpha}'(x) = 1$. Find an analytic expression for this value, which we denote by x_r (cf. Mollerach and Roulet 2002, p. 47). Recall that the corresponding value of u_r can be obtained directly from the lens equation (2.59).
(b) Show that the radius of the radial critical curve satisfies $|x_r| \approx \sqrt{a}$ for $a \ll 1$.
(c) We derived the result of part (b) using a Taylor series approach in Sect. 2.3.3. To extend the analysis to higher order in a, postulate that x_r can be written as power series in $a^{1/2}$:

$$x_r = c_1 a^{1/2} + c_2 a + \mathcal{O}\left(a^{3/2}\right) , \tag{2.90}$$

which seems natural since we know the lowest-order term scales as $\sqrt{a}$. Set $\tilde{\alpha}'(x_r) = 1$, and find the coefficients c_1 and c_2. Why is it not possible to expand x_r in a Taylor series?

2.6 With the power series approach in Problem 2.5, we essentially treat an NIS with a small core radius as a perturbed SIS. The goal of this problem is to extend the analysis to the images produced by an NIS. There are two salient questions:

(a) How do the two images produced by an SIS change with the addition of a core? Recall that a source inside the pseudocaustic of an SIS lens produces two images that are solutions of Eq. (2.49). Postulate that the image position for an NIS can be written as a Taylor series of the form $x = d_0 + d_1 a + \mathcal{O}\left(a^2\right)$. Substitute this expression into Eq. (2.59), show that d_0 satisfies the SIS lens equation, and then find d_1 in terms of d_0 and the source position.
(b) What is the location of the third image? Since we know this image vanishes when $a = 0$, postulate that its position can be expanded as $x = d_1 a + d_2 a^2 + \mathcal{O}\left(a^3\right)$. Find the coefficients d_1 and d_2 in terms of the source position.
(c) Show that expanding the image positions of an NIS lens in half-integer powers of a yields coefficients that cannot be determined in general.

2.7 In Sect. 2.3.3, we showed that the lens equation for the NIS (2.59) can be rewritten as the cubic equation (2.63) and that the latter equation has real roots that may not correspond to lensed images. A cubic equation of the form $Ax^3 + Bx^2 + Cx + D = 0$ has the discriminant

$$\Delta \equiv B^2C^2 - 4B^3D - 4AC^3 + 18ABCD - 27A^2D^2\,, \tag{2.91}$$

whose sign determines the number of real solutions.

(a) For an NIS lens, show that setting $\Delta = 0$ yields a quadratic equation for the square of the source position u^2 in terms of the core radius a. Which solution for u^2 corresponds to the radial caustic?
(b) For certain source positions, we find three real solutions of Eq. (2.63), although only one of them corresponds to an actual lensed image. For other source positions, all real solutions are physically meaningful. For all source positions $u > 0$, identify the number and position(s) of the lensed image(s).

2.8 While rotation curves of spiral galaxies are typically flat to the largest observable radii, the circular speed must eventually drop off. This means that the total mass of the halo in which the disk resides must be finite. In spite of its many desirable properties, the isothermal sphere does satisfy this condition. The **Plummer model** (Plummer 1911), by contrast, has finite mass and, like the SIS, can be derived from an equation of state (Binney and Tremaine 2008). This model is often used to describe globular clusters.

(a) Consider a polytropic gas, defined by the equation of state $P = K\rho^\gamma$, where K and γ are constants. Using the equation of hydrostatic equilibrium (2.37),

show that the density can be written as $\rho = C_n \Psi^n$, where $n \equiv (\gamma - 1)^{-1}$, $C_n \equiv [K(n+1)]^{-n}$, and $\Psi = -\Phi + \text{const}$. The constant is chosen so that Ψ vanishes at the boundary of the system, which could lie at finite or infinite radius.

(b) Define the dimensionless radius $\xi \equiv r/r_0$ and scaled potential $\theta(\xi) \equiv \Psi(r_0\xi)/\Psi_0$, where $\Psi_0 \equiv \Psi(0)$. For what value of r_0 does Poisson's equation (2.38) take the form

$$\frac{1}{\xi^2}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi^2 \frac{\mathrm{d}\theta}{\mathrm{d}\xi}\right) + \theta^n = 0\,, \tag{2.92}$$

which is known in astrophysics as the **Lane-Emden equation**.

(c) Analytic solutions can be found only for a few values of n; the Plummer model corresponds to $n = 5$. By definition, $\theta(0) = 1$. If we make the physically reasonable assumption that the potential and density are nearly constant for small radii ($0 < r \ll r_0$), we obtain the additional boundary condition $\theta'(0) = 0$. We assume the trial solution $\theta(\xi) = (1 + \xi^2/b^2)^{-1/2}$. Find the value of b for which $\theta(\xi)$ solves the Lane-Emden equation. In terms of the central density, $\rho_c \equiv \rho(0)$, and core radius, $r_c \equiv br_0$, give an expression for the density profile of the Plummer model. Show that the total mass is finite.

(d) If we regard the Plummer model as a gravitational lens, show that its reduced deflection angle is given by

$$\alpha(\theta) = \left(\theta_E^2 + \theta_c^2\right)\left(\frac{\theta}{\theta^2 + \theta_c^2}\right), \tag{2.93}$$

where $\theta_c = r_c/D_l$. What is the Einstein angle θ_E in terms of the total mass, $M_{\rm tot}$? What are the possible numbers of images?

(e) For what values of the source distance D_s would lensing by the globular cluster ω Centauri be observable? As a proxy, we consider lensing to be observable if $\theta_E > 1''$. ω Cen lies at a distance of $D_l = 4.8\,\mathrm{kpc}$ and has mass $M_{\rm tot} = 5 \times 10^6 M_\odot$, and core radius $r_c = 4.5\,\mathrm{pc}$. Assume that $D_{ls} = D_s - D_l$.

2.9 Show that

$$\bar{\kappa}'(\theta) = \frac{2}{\theta}[\kappa(\theta) - \bar{\kappa}(\theta)], \tag{2.94}$$

and hence that if $\bar{\kappa}$ decreases as $|\theta|$ increases, then $\bar{\kappa} > \kappa$ for all $\theta \neq 0$.

2.10 Show that Eq. (2.77), $\lim_{\theta\to 0}\bar{\kappa}(\theta) = \kappa(0)$, holds under either of the following assumptions:

(a) The convergence $\kappa(\theta)$ may be expanded in a Taylor series about the point $\theta = 0$.
(b) The convergence is a continuous function at the origin, so that for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|\theta| < \delta \implies |\kappa(\theta) - \kappa(0)| < \epsilon\,.$$

Hint: Recall that $\left|\int f(x)\mathrm{d}x\right| \leq \int |f(x)|\,\mathrm{d}x$ for any integrable function $f(x)$.

2.11 Show that $\alpha'(\theta)$ is continuous at all θ for which $\kappa(\theta)$ is continuous.

2.12 For a continuous convergence function, show that the shear vanishes if and only if the convergence is constant. Physically, a uniform convergence magnifies all sources by the same amount, regardless of position, and it leads to the **mass-sheet degeneracy** (see Sect. 4.7).

References

Binney, J., & Tremaine, S. (2008). *Galactic dynamics* (2nd ed.). Princeton: Princeton University Press.

Mollerach, S., & Roulet, E. (2002). *Gravitational lensing and microlensing*. Singapore: World Scientific.

Ostriker, J. P., & Peebles, P. J. E. (1973). A numerical study of the stability of flattened galaxies: Or, can cold galaxies survive? *Astrophysical Journal, 186*, 467.

Paczyński, B. (1998). Gravitational microlensing with the space interferometry mission. *Astrophysical Journal Letters, 494*, L23.

Plummer, H. C. (1911). On the problem of distribution in globular star clusters. *Monthly Notices of the Royal Astronomical Society, 71*, 460.

Soldner, J. (1804). Über die Ablenkung eines Lichtstrahls von seiner geradlinigen Bewegung, durch die Attraktion eines Weltkörpers, an welchem er nahe vorbei geht. *Berliner Astronomisches Jahrbuch, 1804*, 161.

Strutt, J. W. (1881). On the electromagnetic theory of light. *Philosophical Magazine, 12*, 81.

Chapter 3
Light Deflection in Curved Spacetime

Before turning to lensing by arbitrary mass distributions, we would do well to remember that the Newtonian deflection angle is off from the actual value by a factor of two. Narrowly speaking, the result of this chapter will be that factor. More broadly, though, understanding how gravitational lensing is an inevitable consequence of general relativity puts the following chapters on solid theoretical ground.

We begin in Sect. 3.1 with a review of special relativity, including the mathematics of four-dimensional spacetime. The geodesic equation, whose solution describes the trajectory of a particle moving through curved spacetime, is presented in Sect. 3.2. The special case of a spherically symmetric mass distribution, which is the basis of weak-field gravitational lensing, is treated in Sects. 3.3 and 3.4. We conclude in Sect. 3.5 with a discussion of a homogeneous, isotropic universe, which is the setting for the cosmological applications of lensing given in later chapters.

3.1 Review of Special Relativity

3.1.1 Galilean Transformations

Central in Newtonian physics is the idea of an *inertial* reference frame, which is simply a system of coordinates in which Newton's second law (Newton II) holds:

$$\mathbf{F}(\mathbf{x}, \dot{\mathbf{x}}; t) = m\ddot{\mathbf{x}}\,,$$

where an overdot indicates the total time derivative. In order for this equation to be meaningful, it should not depend too sensitively on the coordinates being used. To see whether this is the case, suppose that Newton II holds in some reference frame. Under what coordinate transformations will it remain valid?

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_3

Because Newton II involves acceleration, we would not expect it to be preserved under acceleration of the coordinates. Fortunately, Galileo found that a coordinate system moving with constant velocity relative to an inertial one is itself inertial. To see this, consider two coordinate systems, S and S', with points denoted by $\mathbf{x}$ and $\mathbf{x}'$, respectively. If S and S' are related by the transformation

$$t' = t \tag{3.1a}$$

$$\mathbf{x}' = \mathbf{x} - \mathbf{v}t \tag{3.1b}$$

for some constant velocity $\mathbf{v}$, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t'^2}\mathbf{x}' = \frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathbf{x} - \mathbf{v}t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x} = \frac{\mathbf{F}}{m}\,. \tag{3.2}$$

Thus, if Newton II holds in S, it holds in S', as well. In words, we say that Newton's second law is *invariant* under **Galilean transformations** (Eqs. (3.1)). This is known as the principle of **Galilean relativity**. In Newtonian physics, the notion of absolute time renders Eq. (3.1a) superfluous. However, one cannot cleanly separate space and time in special relativity.

3.1.2 Lorentz Transformations

While Newton's laws of mechanics and gravitation have greatly advanced our understanding of the physical world, it was not until the nineteenth century that electromagnetic phenomena were integrated into the classical picture. In the absence of electric charge and current, the four Maxwell equations imply wave equations for the electric and magnetic fields. Since a time-varying field of one type induces a time-varying field of the other, these equations describe a single entity: the electromagnetic wave. As it happens, the speed of such a wave is predicted to be c, the speed of light. What is peculiar is that this speed is independent of inertial reference frame, which patently violates the principle of Galilean relativity. In the case in which the S' frame travels in the x direction with velocity v with the respect to the S frame, the transformation that leaves Maxwell's Equations invariant, known as the **Lorentz transformation**, is

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) \tag{3.3a}$$

$$x' = \gamma\,(x - vt) \tag{3.3b}$$

$$y' = y \tag{3.3c}$$

$$z' = z\,, \tag{3.3d}$$

where

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} . \tag{3.4}$$

For $v/c \ll 1$, the Lorentz transformation (3.3) reduces to the Galilean transformation (3.1).

It is often convenient to express the Lorentz transformation in matrix form. In terms of the quantities $w \equiv ct$, $\beta \equiv v/c$, and $\cosh\chi \equiv \gamma$, the transformation can be written as

$$\begin{bmatrix} w' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cosh\chi & -\sinh\chi & 0 & 0 \\ -\sinh\chi & \cosh\chi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} . \tag{3.5}$$

This can be generalized to a boost in an arbitrary spatial direction (Problem 3.1).

3.1.3 *Four-Vectors in Minkowski Space*

Let us denote a boost by $\mathbf{v}$ with the matrix $\Lambda(\mathbf{v})$. We can then write $\mathbf{X}' = \Lambda(\mathbf{v})\mathbf{X}$, where $\mathbf{X}$ and $\mathbf{X}'$ are four-dimensional vectors. In the context of linear algebra, a vector is simply a matrix consisting of a single row or column. However, we think of a vector here as an object whose length is invariant under boost. One consequence is that if we define $|\mathbf{X}|$ according to the usual Euclidean distance, we will find that $|\Lambda\mathbf{X}| = \left|\mathbf{X}'\right| \neq |\mathbf{X}|$. Instead, for a vector

$$\mathbf{X} \equiv \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} , \tag{3.6}$$

we define

$$|\mathbf{X}|^2 \equiv w^2 - \left(x^2 + y^2 + z^2\right) , \tag{3.7}$$

where $w \equiv ct$ as before. Under this definition of distance, $|\mathbf{X}|$ is invariant under Lorentz transformation and is often called a **four-vector**.[1] In what follows, we use

[1]This is a misnomer because a four-vector describes all objects that transform via the Lorentz transformation regardless of how many dimensions are assumed. Moreover, simply having four dimensions does not make a vector a *four*-vector.

uppercase letters to denote four-vectors and lowercase letters to denote vectors in Euclidean space. For any four-vectors, $\mathbf{X}_1$ and $\mathbf{X}_2$, the dot product takes the form

$$\mathbf{X}_1 \cdot \mathbf{X}_2 = w_1 w_2 - (x_1 x_2 + y_1 y_2 + z_1 z_2) \ . \tag{3.8}$$

Four-vectors are elements of Minkowski space, denoted by $\mathbb{R}^{(3,1)}$. The difference between $\mathbb{R}^{(3,1)}$ and $\mathbb{R}^4$ is the **metric**, which specifies the infinitesimal distance between neighboring points. In "flat" spacetimes, such as $\mathbb{R}^n$ and $\mathbb{R}^{(3,1)}$, differentials can simply be replaced by finite displacements. Anticipating our discussion of general relativity, we will primarily work with the metric in its infinitesimal version. To see most clearly the distinction between Euclidean and Minkowski metrics, we write these two metrics in turn. In four-dimensional Euclidean space, $\mathbb{R}^4$, we have

$$\mathrm{d}s^2 = \mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \equiv \mathrm{d}\mathbf{x}^T \, \mathsf{I} \, \mathrm{d}\mathbf{x} \, , \tag{3.9}$$

where I is the (four-dimensional) identity matrix. We take $\mathrm{d}\mathbf{x}$ to be a column vector, so that its transpose, $\mathrm{d}\mathbf{x}^T$, is a row vector. Four-vectors in $\mathbb{R}^{(3,1)}$ are described by the **Minkowski metric**:

$$\mathrm{d}s^2 = \mathrm{d}w^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \equiv \mathrm{d}\mathbf{X}^T \, \eta \, \mathrm{d}\mathbf{X} \, , \tag{3.10}$$

where

$$\eta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \, . \tag{3.11}$$

In general, the infinitesimal distance between neighboring points can be written as

$$\mathrm{d}s^2 = \mathrm{d}\mathbf{X}^T \, \mathsf{g} \, \mathrm{d}\mathbf{X} \, , \tag{3.12}$$

where g is called the **metric tensor**. Our use of capital letters anticipates the result that g will reduce to η in the absence of nearby massive bodies.

In Newtonian mechanics, $\mathbf{x}$ will be a curve that satisfies Newton's second law, parametrized by the time, t, i.e., $\mathbf{x} = \mathbf{x}(t)$. In special relativity, $\mathbf{X}$ includes time as one of its coordinates, so it cannot be used to parametrize the trajectory of a particle. Instead, we use the **proper time**, τ, which is the time measured in the reference frame of the moving particle; the coordinate time, t, is that measured by an inertial observer. To determine how the proper time is related to the spacetime interval, $\mathrm{d}s$, consider an observer who measures a particle to be at rest. In this case, there is no relative motion between the reference frame of the particle and that of the observer. Thus, $t = \tau$ in this case. Since the particle is at rest, the observer will find that $\mathrm{d}x = \mathrm{d}y = \mathrm{d}z = 0$. Equation (3.10) then gives

$$\mathrm{d}s^2 = \mathrm{d}w^2 \equiv c^2\mathrm{d}t^2 = c^2\mathrm{d}\tau^2 \,. \tag{3.13}$$

Thus, $\mathrm{d}s = c\,\mathrm{d}\tau$. This relation turns out to hold in general.

Let us apply this result to the four-velocity, $\mathbf{U} \equiv \mathrm{d}\mathbf{X}/\mathrm{d}\tau$. We can find its magnitude by "dividing" Eq. (3.10) by $\mathrm{d}\tau^2$ to find

$$\left(\frac{\mathrm{d}s}{\mathrm{d}\tau}\right)^2 = c^2 = \left(\frac{\mathrm{d}w}{\mathrm{d}\tau}\right)^2 - \left(\frac{\mathrm{d}x}{\mathrm{d}\tau}\right)^2 - \left(\frac{\mathrm{d}y}{\mathrm{d}\tau}\right)^2 - \left(\frac{\mathrm{d}z}{\mathrm{d}\tau}\right)^2 \,. \tag{3.14}$$

Therefore, $|\mathbf{U}|^2 = c^2$ in any inertial reference frame. Defining the forward momentum as $\mathbf{P} \equiv m\mathbf{U}$, we find $|\mathbf{P}|^2 = m^2c^2$ where m is the mass of the particle. Our next step is to write $\mathbf{P}$ in terms of the coordinate time, t. By the chain rule, we have

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}\tau} = \frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau} \,. \tag{3.15}$$

We obtain the second factor on the right-hand side by writing Eq. (3.10) in terms of t and τ:

$$c^2\mathrm{d}\tau^2 = c^2\mathrm{d}t^2 - \left(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\right) \,, \tag{3.16}$$

or

$$\begin{aligned}\frac{\mathrm{d}\tau}{\mathrm{d}t} &= \sqrt{1 - \frac{1}{c^2}\left[\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2\right]} \\ &= \sqrt{1 - \frac{|\mathbf{u}|^2}{c^2}} \\ &= \sqrt{1 - |\boldsymbol{\beta}|^2} \\ &= \gamma \,.\end{aligned} \tag{3.17}$$

The vector $\mathbf{u}$ is just the usual (three-dimensional) velocity vector of the particle measured with respect to coordinate time: $\mathbf{u} \equiv \mathrm{d}\mathbf{x}/\mathrm{d}t$. Then it is conventional to write $\boldsymbol{\beta} = \mathbf{u}/c$, $\beta = |\boldsymbol{\beta}|$, and $\gamma = \left(1 - \beta^2\right)^{-1/2}$.

Now we can write the components of the four-momentum, $\mathbf{P} = m\mathbf{U}$, in terms of the ordinary three-momentum, $m\mathbf{u}$:

$$\mathbf{P} = (\gamma mc, \gamma m\mathbf{u}) \,. \tag{3.18}$$

To make sense of the components of $\mathbf{P}$, we first note that particles moving at a significant fraction of the speed of light have their momentum augmented by the

factor γ. The physical significance of the time component, γmc, is less clear. In the limit $|\mathbf{u}| \ll c$, we can hope to make contact with our Newtonian intuition. To this end, we make the Taylor expansion

$$\begin{aligned} \gamma mc &= mc\left(1 - \frac{u^2}{c^2}\right)^{-1/2} \\ &\approx mc + \frac{mu^2}{2c} \\ &\approx \frac{1}{c}\left(mc^2 + \frac{1}{2}mu^2\right), \end{aligned} \tag{3.19}$$

where $u \equiv |\mathbf{u}|$. We immediately identify the second term in parentheses with the kinetic energy, while the first term is independent of velocity and is known as the rest energy. There is no Newtonian analogue to the rest energy, and it has no bearing on the laws of motion in that case.

If we identify the parenthetical quantity on the right hand side of Eq. (3.19) to be the total energy (rest + kinetic) of a particle, we can write the four-momentum as

$$\mathbf{P} = \left(\frac{E}{c}, \mathbf{p}\right), \tag{3.20}$$

where $\mathbf{p} = \gamma m\mathbf{u}$.

We wish to obtain an expression for the energy, E, that is valid for all $u < c$. To do this, we begin by taking the dot product of $\mathbf{P}$ with itself. Just as the magnitude of a vector in Euclidean space is invariant under translations and rotations of coordinates, the magnitude of a four-vector is invariant under Lorentz transformations. This means that we may determine the dot product of interest in a convenient inertial frame; we choose the frame in which $\mathbf{u} = 0$, i.e., the rest frame of the particle. We have $\mathbf{P} = (mc, \mathbf{0})$, and hence,

$$|\mathbf{P}| = mc\,. \tag{3.21}$$

When $\mathbf{u}$ is nonzero, we find

$$c^2\,|\mathbf{P}|^2 = m^2c^4 = E^2 - c^2p^2\,, \tag{3.22}$$

where $p = |\mathbf{p}|$. Alternatively, one can show that

$$E = \gamma mc^2\,. \tag{3.23}$$

We have been using the standard notation for dot products, but we note that $\mathbf{P} \cdot \mathbf{P} \neq \mathbf{P}^T\mathbf{P}$; we must remember to include the Minkowski metric tensor, η, by writing $\mathbf{P} \cdot \mathbf{P} = \mathbf{P}^T\eta\mathbf{P}$. This notation is a bit unwieldy, so we often write $\mathbf{P}$ in terms

of its components, p^μ. We use Greek indices to refer to components of vectors and tensors in spacetime and Roman indices for quantities in Euclidean space, whether in rectangular or curvilinear coordinates. While we continue to denote four-vectors by capital letters, we use lower case letters when writing them in component form. It is conventional to write the dot product as $p^\mu p_\mu$, where repeated indices are summed over if they appear as a subscript-superscript pair. This is known as the Einstein summation convention. We think of p^μ as fundamental and p_μ as auxiliary. These quantities are related by

$$p_\mu = \eta_{\mu\nu} p^\nu \,, \tag{3.24}$$

so that

$$p^\mu p_\mu = \eta_{\mu\nu} p^\mu p^\nu \equiv \mathbf{P}^T \eta \mathbf{P} \,. \tag{3.25}$$

In general the dot product between four-vectors $\mathbf{A}$ and $\mathbf{B}$ in Minkowski space is given by

$$\mathbf{A} \cdot \mathbf{B} \equiv \mathbf{A}^T \eta \mathbf{B} = \eta_{\mu\nu} a^\mu b^\nu \,. \tag{3.26}$$

In addition to offering notational convenience, upper and lower indices convey mathematical content as well (Sect. 3.2.1). We revert to the standard practice of writing vector components with subscripts in future chapters.

3.2 Geodesic Equation

3.2.1 Contravariant and Covariant Vectors

Since relating quantities in one reference frame to those in another occurs frequently in GR, we briefly digress to derive the relevant transformations. Vectors can be written in components that are either **covariant** or **contravariant**. The reason for these names will become clear shortly.

Consider two inertial coordinate systems, S and S', that are related by a spatial rotation. We will work in Euclidean space for now. A translation (including uniform relative motion) between S and S' would not affect our discussion, so we ignore that possibility for convenience. Measuring displacements by $\mathbf{x}$ and $\mathbf{x}'$ in the their respective frames, we can write

$$\mathrm{d}\mathbf{x}' = \left(\frac{\partial \mathbf{x}'}{\partial \mathbf{x}} \right) \mathrm{d}\mathbf{x} \equiv \mathsf{R}\, \mathrm{d}\mathbf{x} \,, \tag{3.27}$$

where R has components (elements)

$$R^i_j = \frac{\partial x'^i}{\partial x^j}\,, \tag{3.28}$$

where x^i and x'^i denote the components of $\mathbf{x}$ and $\mathbf{x}'$, respectively. Because S and S' have a common origin, we can integrate Eq. (3.27) to get

$$\mathbf{x}' = \mathsf{R}\,\mathbf{x}\,, \tag{3.29}$$

or

$$x'^i = R^i_j x^j\,. \tag{3.30}$$

Notice that j appears both as a superscript and a subscript, thus putting the Einstein summation convention into force. *The components of any vector that transform as those of a displacement are said to be contravariant*.

Now consider the gradient of a scalar function f. The chain rule yields

$$\frac{\partial f}{\partial x'^i} = \frac{\partial f}{\partial x^j}\,\frac{\partial x^j}{\partial x'^i}\,, \tag{3.31}$$

where a superscript (subscript) in the denominator is treated as a subscript (superscript) in the numerator by the Einstein convention. We can then write

$$\boldsymbol{\nabla}' f = \mathsf{S}\boldsymbol{\nabla} f\,, \tag{3.32}$$

where the components of S are given this time by

$$S^j_i = \frac{\partial x^j}{\partial x'^i}\,. \tag{3.33}$$

In component form, we can write

$$\partial'_i f = S^j_i\,\partial_j f\,, \tag{3.34}$$

where ∂_i is shorthand for $\partial/\partial x^i$. Defining $u_i \equiv \partial_i f$, we finally obtain

$$u'_i = S^j_i\,u_j\,. \tag{3.35}$$

The components of any vector that transform as those of the gradient are said to be covariant.

It turns out that basis vectors transform covariantly (Problems 3.2 and 3.3). Because of this, any vector whose components transform like basis vectors are

called covariant; those transforming differently are termed contravariant (see Peacock 1999). For the remainder of this chapter, superscripts denote contravariant components, while subscripts denote covariant components. Sums involving the product of a contravariant component with a covariant one appear frequently, inspiring the use of Einstein's convention. The framework developed here can be applied to nonvector tensor components, as well. Indeed, we have already encountered such tensors, namely, the rotation matrices R and S.

3.2.2 Metric Tensor

We mentioned briefly in Sect. 3.1.3 that the separation between points connected by an infinitesimal displacement is given by Eq. (3.12), which can be written in component form as

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu\,, \tag{3.36}$$

where $g_{\mu\nu}$ are the components of the metric tensor g. Since ds is independent of the inertial coordinate system being used, we find

$$\begin{aligned} g_{\mu\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu &= g'_{\alpha\beta}\,\mathrm{d}x'^\alpha \mathrm{d}x'^\beta \\ &= g'_{\alpha\beta}\frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu\,. \end{aligned} \tag{3.37}$$

Since this relation must hold for all coordinate displacements $\mathrm{d}x^\lambda$, we have

$$g_{\mu\nu} = g'_{\alpha\beta}\frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}\,. \tag{3.38}$$

To solve for g′ in terms of g, we recast the above equation as

$$\mathsf{g} = \Lambda^T \mathsf{g}'\Lambda\,, \tag{3.39}$$

whence

$$\mathsf{g}' = \Lambda^{-T}\mathsf{g}\Lambda^{-1}\,, \tag{3.40}$$

where Λ represents either a spatial rotation or Lorentz transformation; it has components $\Lambda^\mu_\nu = \partial_\nu x'^\mu$. Defining Σ by $\Sigma^\mu_\nu \equiv \partial'_\nu x^\mu$, we have

$$\Sigma^\mu_\kappa \Lambda^\kappa_\nu = \frac{\partial x^\mu}{\partial x'^\kappa}\frac{\partial x'^\kappa}{\partial x^\nu} = \frac{\partial x^\mu}{\partial x^\nu} = \delta^\mu_\nu\,, \tag{3.41}$$

where δ^{μ}_{ν} is the Kronecker delta. Thus, $\Lambda^{-1} = \Sigma$ and so

$$\mathrm{g}' = \Sigma^T \mathrm{g} \Sigma \,, \tag{3.42}$$

or equivalently,

$$g'_{\alpha\beta} = g_{\mu\nu} \frac{\partial x^{\mu}}{\partial x'^{\alpha}} \frac{\partial x^{\nu}}{\partial x'^{\beta}} \,. \tag{3.43}$$

Therefore, the metric tensor is *covariant.* Because it has two indices, the metric tensor is said to have rank two; an Nth rank tensor contains the product of N partial derivatives, one factor for each index.

There are a few facts about the metric tensor worth noting before putting it to use. The dot product of four-vectors $\mathbf{A}$ and $\mathbf{B}$ in *curved* spacetime is given by the straightforward generalization of Eq. (3.26):

$$\mathbf{A} \cdot \mathbf{B} \equiv \mathbf{A}^T \mathrm{g} \mathbf{B} = g_{\mu\nu} a^{\mu} b^{\nu} \,. \tag{3.44}$$

In the case $\mathbf{A} = \mathbf{B}$, the dot product simply gives $|\mathbf{A}|^2 = \mathbf{A}^T \mathrm{g} \mathbf{A}$. Since this is a scalar, it is identical to its transpose:

$$|\mathbf{A}|^2 = \left(\mathbf{A}^T \mathrm{g} \mathbf{A} \right)^T = \mathbf{A}^T \mathrm{g}^T \mathbf{A} \,. \tag{3.45}$$

This implies that g is symmetric. Another important property is that g^{-1} is contravariant. By way of justification, note that

$$g^{\mu\lambda} g_{\lambda\nu} = \delta^{\mu}_{\nu} \,, \tag{3.46}$$

where $g^{\mu\lambda}$ are the components of g^{-1}. It would not be possible to write down this identity if g^{-1} were covariant, i.e., if g^{-1} had subscripts rather than superscripts. A genuine proof is left to the interested reader to work out by brute force. A consequence of this result is that we can compute a component a^{μ} in terms of a_{μ}. To see this, we first extend Eq. (3.24) to an arbitrary four-vector in curved spacetime:

$$a_{\mu} = g_{\mu\nu} a^{\nu} \,. \tag{3.47}$$

Multiplying by $g^{\lambda\mu}$ yields

$$g^{\lambda\mu} a_{\mu} = g^{\lambda\mu} g_{\mu\nu} a^{\nu} = \delta^{\lambda}_{\nu} a^{\nu} \,, \tag{3.48}$$

from which we obtain the desired result:

$$a^{\lambda} = g^{\lambda\mu} a_{\mu} \,. \tag{3.49}$$

3.2.3 *Principle of Stationary Interval*

We have seen that the metric tensor is the fundamental quantity for describing spacetime. Two questions that naturally arise are how to determine g and how this relates to the motion of a particle through spacetime. Our approach will be to motivate what the metric tensor should look like under specified assumptions but stop short of obtaining it directly from the Einstein field equations. We refer the interested reader to a monograph or textbook on general relativity for details left out in our discussion. A more relevant topic to gravitational lensing is the connection between g and the trajectory of a photon. The culmination of our effort will be obtaining the **geodesic equation**. We arrive at this result gradually, as there are a few preliminaries to which we must attend.

The first of these is to clarify our terminology. To avoid confusion, we henceforth refer to g as the *metric* and to ds as the (infinitesimal) spacetime **interval**. We would ordinarily refer to ds as a distance, but it is not clear from the outset what its physical meaning would be in spacetime. Indeed, ds need not be positive, nor even real. If $\mathrm{d}s^2 > 0$, we say that the interval is *timelike;* if $\mathrm{d}s^2 < 0$, the interval is *spacelike;* if $\mathrm{d}s^2 = 0$, the interval is *lightlike*, or *null*. To find the total interval between two events A and B, we integrate ds along some appropriate path. To build intuition, we begin with the well-known **Twin Paradox** of special relativity. We then introduce the **Equivalence Principle**, which will allow us to consider non-uniform (accelerated) motion. This framework will enable us to find the equations of motion for both massive and massless particles.

Twin Paradox

Consider identical twins A and B at rest in some inertial reference frame. Knowing that time is not an absolute concept in special relativity, the twins decide to conduct an experiment in which A remains on their home planet, while B travels at high speed through space before returning. To see if time flows at the same rate for both, A and B compare the time elapsed on each of their atomic clocks. If there is a significant time difference, one will be visibly older than the other.

Let t be the time as measured by A, and τ be the time as measured by B. Suppose that B leaves at $t = t_1$ and returns at $t = t_2$. The time of departure as measured by B is $\tau_1 = t_1$, but τ_2 is yet to be determined. Regardless of whose clock we refer to, the initial and final spacetime positions (events) are the same. The total time of B's journey as measured by A will be $\Delta t = t_2 - t_1$. The situation is more complicated for B, since there will be an intermediate time $\tau_\star$ when B turns around to head home. The total time as measured by B is $\Delta\tau = (\tau_\star - \tau_1) + (\tau_2 - \tau_\star) = \tau_2 - \tau_1$. Assuming that B turns around instantaneously and moves at a constant speed u (measured as a fraction of c) relative to A at all other times, we have

$$\Delta\tau = \Delta t\sqrt{1 - u^2}\,. \tag{3.50}$$

We see that $\Delta\tau < \Delta t$, i.e., the time measured by A will always be greater than that measured by B. This is true for any $0 < u < 1$, which implies that the shortest spatial path between the initial and final events is the one that maximizes proper time.

If we just calculate without worrying about conceptual nuance, one may wonder where the eponymous paradox lies. Since any inertial frame is as good as any other, one might argue that both twins appear to be younger than the other. The catch is that B accelerates halfway through the journey, breaking the symmetry between the twins' reference frames. Our purpose in presenting the twin paradox here is to motivate the idea of a geodesic, the spacetime analogue of a straight line in Euclidean space. Before we can offer a formal definition, we must generalize our discussion of relativity to include accelerated (non-inertial) coordinate systems.

Equivalence Principle

In our discussion of the twin paradox, we avoided the thorny issue of accelerated reference frames, which special relativity does not address. We did so by treating the trajectory of the moving twin, B, as consisting of two legs of constant speed. Of course, twin B cannot simply turn around instantaneously. The key to resolving this shortcoming of our analysis is the **weak equivalence principle**, which states that inertial and gravitational mass are identical. Put another way, *the results of experiments carried out in a freely falling laboratory will agree with those obtained in an inertial reference frame*. By confining measurements to a laboratory, we can assume that whatever gravitational field is present can be considered uniform. The fact that the lab is freely falling means that gravity is the only external force exerted on it. Note that unless the gravitational field vanishes, the lab is *not* an inertial frame.

In the language of GR, the equivalence principle states that spacetime is locally flat, i.e., described by the Minkowski metric. Einstein went further by introducing the **strong equivalence principle**, which asserts that *the effects of acceleration are indistinguishable from those of gravitation*. This extension of the equivalence principle is necessary if gravitation is truly a manifestation of spacetime curvature, rather than being a conventional force. The relativistic interpretation of gravity leads to physical predictions at odds with Newton's theory. Deriving the equation of motion of a particle in curved spacetime only requires the weak equivalence principle.

To see this, consider timelike - separated events A and B. We divide the trajectory of a particle that moves from A to B into a sequence of small segments during which the proper time is $\delta\tau$. Because spacetime is locally flat, $\delta\tau$ will be maximized along each segment. Thus the total proper time, $\tau_B - \tau_A$, will also be maximized. Such a path through spacetime is known as a geodesic, which we now endeavor to derive.

Equation of Motion

Considering time dilation in both special and general relativistic contexts, we have argued that a particle moving solely under the influence of gravity follows the path of maximal proper time. Using a variational approach, we now derive the equation of motion for an arbitrary metric $g_{\mu\nu}$. Recall that the change in proper time in an infinitesimal displacement of the particle is given by

$$\mathrm{d}\tau = \frac{\mathrm{d}s}{c} = \frac{1}{c} L(\mathbf{X}, \dot{\mathbf{X}})\, \mathrm{d}p\,, \tag{3.51}$$

where

$$L(\mathbf{X}, \dot{\mathbf{X}}) \equiv \sqrt{g_{\mu\nu}(\mathbf{X}) \dot{x}^\mu \dot{x}^\nu}\,, \tag{3.52}$$

as follows from Eq. (3.36), and where $\dot{x}_\alpha \equiv \mathrm{d}x_\alpha/\mathrm{d}p$. Here p is any convenient parameter specifying the path of the particle. Thus, the elapsed proper time as the particle moves from event A to event B is

$$\Delta\tau = \frac{1}{c} \int_{p_A}^{p_B} L(\mathbf{X}, \dot{\mathbf{X}})\, \mathrm{d}p\,. \tag{3.53}$$

Now consider the family of curves $\{x^\mu(p)\}$, which pass through events A and B and satisfy $x^\mu(p_A) = x_A^\mu$ and $x^\mu(p_B) = x_B^\mu$, where x_A^μ and x_B^μ are fixed. One of the curves is the extremal curve or physical path of the particle, which yields a stationary value of the integral in Eq. (3.53) and, according to the calculus of variations, satisfies the Euler-Lagrange equations (Appendix A):

$$\frac{\mathrm{d}}{\mathrm{d}p} \frac{\partial L}{\partial \dot{x}^\alpha} - \frac{\partial L}{\partial x^\alpha} = 0\,. \tag{3.54}$$

In deriving the equation of motion from Eq. (3.54), it is advantageous to work with the quantity $F = L^2$ instead of L, since squaring eliminates the square root in Eq. (3.52). Substituting $L = F^{1/2}$ in Eq. (3.54), one obtains

$$-\frac{1}{2} F^{-1} \frac{\mathrm{d}F}{\mathrm{d}p} \frac{\partial F}{\partial \dot{x}^\alpha} + \frac{\mathrm{d}}{\mathrm{d}p} \frac{\partial F}{\partial \dot{x}^\alpha} - \frac{\partial F}{\partial x^\alpha} = 0\,. \tag{3.55}$$

This equation holds for an arbitrary parameter p. To obtain the equation of motion in a particularly simple form, we choose $p = \tau = c^{-1}s$, where τ and s are the proper time and interval, respectively, along the extremal path. From Eq. (3.51) with $\mathrm{d}p = \mathrm{d}\tau$, we see that $L = F^{1/2} = c$ all along the extremal curve. Thus, for $p = \tau$, $\mathrm{d}F/\mathrm{d}p = 0$, and the first term in Eq. (3.55) vanishes, implying that F satisfies the Euler-Lagrange equations.

To obtain the equation of motion from the Euler-Lagrange equations for $F = g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu$, we need the derivatives

$$\frac{\partial F}{\partial \dot{x}^\alpha} = 2g_{\alpha\mu}\dot{x}^\mu \tag{3.56}$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\partial F}{\partial \dot{x}^\alpha} = 2g_{\alpha\mu}\ddot{x}^\mu + \partial_\nu\left(g_{\alpha\mu} + g_{\mu\alpha}\right)\dot{x}^\mu \dot{x}^\nu \tag{3.57}$$

$$\frac{\partial F}{\partial x^\alpha} = \partial_\alpha g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu\,, \tag{3.58}$$

where we have used the symmetry of the metric tensor in the first line, and we have used the chain rule to evaluate $\mathrm{d}g_{\alpha\mu}/\mathrm{d}\tau$ to obtain the second term in the second line. Putting the pieces together yields

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\partial F}{\partial \dot{x}^\alpha} - \frac{\partial F}{\partial x^\alpha} \\ &= 2g_{\alpha\mu}\ddot{x}^\mu + \left(\partial_\nu g_{\alpha\mu} + \partial_\nu g_{\mu\alpha} - \partial_\alpha g_{\mu\nu}\right)\dot{x}^\mu \dot{x}^\nu\,. \end{aligned} \tag{3.59}$$

Finally, multiplying this result by the inverse metric $g^{\lambda\alpha}$ and summing over α, we arrive at the equation of motion or geodesic equation

$$\boxed{\frac{\mathrm{d}^2 x^\lambda}{\mathrm{d}\tau^2} = -\Gamma^\lambda_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}}\,, \tag{3.60}$$

where the **Christoffel symbol**,[2] also known as the *affine connection*, is defined by

$$\Gamma^\lambda_{\mu\nu} \equiv \frac{1}{2}g^{\lambda\alpha}\left(\partial_\mu g_{\alpha\nu} + \partial_\nu g_{\mu\alpha} - \partial_\alpha g_{\mu\nu}\right)\,. \tag{3.61}$$

3.3 Schwarzschild Solution

3.3.1 Gravitational Time Dilation

While we are not in a position to derive the Schwarzschild metric from Einstein's equations (but see Appendix A.3), we can at least attempt to offer some motivation for this important solution. Our point of entry is **gravitational time dilation**, which can be most readily understood through the related phenomenon of **gravitational redshift**. Consider a source that emits photons of wavelength λ. Let this source accelerate away from an observer at a rate $a(t)$. During a time interval $\mathrm{d}t$, the

[2] The word "symbol" reminds us that $\Gamma^\lambda_{\mu\nu}$ does not transform as a tensor (see Problem 3.4).

wavelength will be redshifted by an amount $\mathrm{d}\lambda = (a\,\mathrm{d}t/c)\lambda$, where c is the propagation speed of the wave, namely the speed of light.

We know from the strong equivalence principle that the effects of acceleration are indistinguishable from those of gravity, allowing us to write $\mathrm{d}\lambda = (\Phi'(r)\,\mathrm{d}r/c^2)\lambda$, where Φ is the (Newtonian) gravitational potential, which we take to be radial for convenience. The absence of a negative sign follows from our convention that acceleration is defined to be positive for a light source moving away from the observer. For a light source at a distance r from the center of the spherically symmetric matter distribution with total mass M, the wavelength observed asymptotically far from the source is obtained by integrating from the source (radius r and wavelength λ_r) to the observed (where the radius can be taken as infinity and the wavelength is λ_∞) as follows:

$$\int_{\lambda_r}^{\lambda_\infty} \frac{\mathrm{d}\lambda}{\lambda} = \frac{GM}{c^2}\int_r^\infty \frac{\mathrm{d}r}{r^2}$$
$$\Longrightarrow\ \lambda_\infty = \exp\left(\frac{GM}{c^2 r}\right)\lambda_r \approx \left(1+\frac{GM}{c^2 r}\right)\lambda_r\,. \tag{3.62}$$

assuming $GM/(c^2 r) \ll 1$. Since wavelength and period are proportional, it stands to reason that

$$t_\infty \approx \left(1+\frac{GM}{c^2 r}\right) t_r\,. \tag{3.63}$$

Strictly speaking, t_r is the intrinsic period of a source at radius r, while t_∞ is the period measured by a distant observer. More broadly, though, we can think of t_r and t_∞ as the time measured by clocks at r and ∞, respectively.

Just as acceleration was the key to making sense of the twin paradox, so too does it explain gravitational time dilation, by way of the equivalence principle. The observer at infinity plays the role of twin A above (Sect. 3.2.3), while the one at r stands in for twin B: a given time interval measured by someone in a gravitational field will appear to be longer to a distant observer. Equation (3.63) is qualitatively correct, but it is important to realize that it holds quantitatively only for a weak gravitational field, where $r \gg GM/c^2$. This assumption crept in when we took the classical Doppler effect to apply to an inherently relativistic setting. In a proper derivation, we would find gravitational time dilation as a direct consequence of the Schwarzschild metric. However, following this path would take us too far afield.

3.3.2 *Spacetime Interval Outside a Static, Spherically Symmetric Mass Distribution*

Our goal in this subsection is to write down the metric and corresponding spacetime interval that will allow us to compute the bending angle of light by matter.

By assuming the deflector to be static, we know the metric cannot have time dependence. Moreover, spherical symmetry implies that, at fixed time and radius, the metric reduces to that of the surface of a sphere, and that radial motion is independent of the angular coordinates. If the total mass of the lens is finite, we must recover flat spacetime at large radii. Symbolically, this translates as

$$\mathrm{d}s^2 = f(r)c^2\mathrm{d}t^2 - g(r)\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right), \tag{3.64}$$

where $f(\infty) = g(\infty) = 1$.

For a particle at a fixed location, we find

$$\mathrm{d}s^2 \equiv c^2\mathrm{d}\tau^2 = f(r)c^2\mathrm{d}t^2, \tag{3.65}$$

where τ is the proper time of an observer at r, and t is the **coordinate time**, measured as $r \to \infty$. In other words, τ and t are identical to t_r and t_∞, respectively, from our discussion of gravitational time dilation above. If Eq. (3.63) held exactly, we would have $f^{-1/2}(r) = 1 + GM/(c^2 r)$. As it happens, the correct functional form is

$$f^{-1/2}(r) = \left(1 - 2\frac{GM}{c^2 r}\right)^{-1/2}, \tag{3.66}$$

which reduces to our approximation for large r. An intuitive route to $g(r)$ does not exist, so we merely quote the result:

$$g(r) = \left(1 - \frac{2GM}{c^2 r}\right)^{-1}. \tag{3.67}$$

With f and g in hand, we can write down the Schwarzschild interval as

$$\boxed{\mathrm{d}s^2 = \left(1 - \frac{2GM}{c^2 r}\right)c^2\mathrm{d}t^2 - \left(1 - \frac{2GM}{c^2 r}\right)^{-1}\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)}. \tag{3.68}$$

The combination $R_S \equiv 2GM/c^2$ is known as the **Schwarzschild radius**.

On making the coordinate transformation

$$r = \rho\left(1 + \frac{GM}{2c^2\rho}\right)^2, \tag{3.69}$$

the Schwarzschild metric takes the "isotropic" form

$$\mathrm{d}s^2 = \left(\frac{1 - \frac{GM}{2c^2\rho}}{1 + \frac{GM}{2c^2\rho}}\right)^2 c^2\mathrm{d}t^2 - \left(1 - \frac{GM}{2c^2\rho}\right)^{-1}\left[\mathrm{d}\rho^2 + \rho^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right]. \tag{3.70}$$

In the weak-field limit $|\Phi(\rho)|/c^2 \ll 1$, with $\Phi(\rho) = -GM/\rho$, this reduces to

$$\mathrm{d}s^2 \approx \left(1 + 2\,\frac{\Phi}{c^2}\right) c^2 \mathrm{d}t^2 - \left(1 - 2\,\frac{\Phi}{c^2}\right) \left[\mathrm{d}\rho^2 + \rho^2 \left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right] . \tag{3.71}$$

This result turns out to hold for an arbitrary weak gravitational field, $|\Phi(\rho, \theta, \phi)| \ll c^2$, that vanishes for $\rho \to \infty$, but the general-relativistic derivation is beyond the scope of this book. Note that the corrections to the Minkowski metric (flat spacetime) are linear in Φ in the weak-field limit.

3.3.3 Circular Orbits

Having written down the Schwarzschild interval, we wish to determine the path through spacetime a particle will follow. For now, we parametrize the trajectory by the proper time, which formally excludes photons. This shortcoming will turn out to have a straightforward remedy. The geodesic equation consists of four separate equations, one for each spacetime coordinate. Since lensing by a spherically symmetric mass distribution occurs entirely in the plane defined by the observer, source, and lens center, we can set $\theta = \pi/2$ without loss of generality. Thus, the number of equations we must solve is reduced to three, one each for t, r, and ϕ.

Including only the non-vanishing Christoffel symbols, the geodesic equation takes the form

$$\ddot{t} = -2\Gamma^t_{tr}\dot{t}\dot{r} \tag{3.72a}$$

$$\ddot{r} = -\Gamma^r_{tt}\dot{t}^2 - \Gamma^r_{rr}\dot{r}^2 - \Gamma^r_{\phi\phi}\dot{\phi}^2 \tag{3.72b}$$

$$\ddot{\phi} = -2\Gamma^\phi_{r\phi}\dot{r}\dot{\phi} . \tag{3.72c}$$

The spherical symmetry of the problem suggests that the motion should be described by a radial equation. In Newtonian mechanics, we would construct an effective potential, consisting of a gravitational term plus a centrifugal term. In addition, we would have two integrals of motion, namely the energy and angular momentum. In its present form, only the left-hand side of Eq. (3.72b) resembles Newton's second law. We can use Eqs. (3.72a) and (3.72c) to find parameters representing the energy per unit mass and angular momentum per unit mass, respectively. This will allow us to replace $\dot{t}^2$ and $\dot{r}^2$ in Eq. (3.72b) by functions of r only.

We begin by writing Eqs. (3.72a) and (3.72c) in the suggestive form

$$\frac{\mathrm{d}\ln\dot{t}}{\mathrm{d}\tau} = -\frac{\mathrm{d}\ln\left(1 - \frac{R_S}{r}\right)}{\mathrm{d}\tau} \tag{3.73a}$$

$$\frac{\mathrm{d}\ln\dot{\phi}}{\mathrm{d}\tau} = -2\,\frac{\mathrm{d}\ln r}{\mathrm{d}\tau} , \tag{3.73b}$$

where $R_S \equiv 2GM/c^2$ is the Schwarzschild radius, and the relevant Christoffel symbols are given by

$$\Gamma^t_{tr} = \frac{R_s}{2r(r - R_S)} \tag{3.74a}$$

$$\Gamma^\phi_{r\phi} = \frac{1}{r}\,. \tag{3.74b}$$

Upon integration, we arrive at

$$\ln \dot{t} = -\ln\left(1 - \frac{R_S}{r}\right) + \ln\frac{\epsilon}{c^2} \tag{3.75a}$$

$$\ln \dot{\phi} = -2\ln r + \ln \ell\,, \tag{3.75b}$$

or

$$\epsilon = c^2\left(1 - \frac{R_S}{r}\right)\dot{t} \tag{3.76a}$$

$$\ell = r^2\dot{\phi}\,, \tag{3.76b}$$

where ϵ and ℓ are constants, respectively measuring the energy (accounting for gravitational redshift) per unit mass and the angular momentum per unit mass. We can use these constants of motion to rewrite Eq. (3.72b) so that the right hand side depends only on r. We can substitute immediately for $\dot{t}^2$ and $\dot{\phi}^2$, but we must appeal to the interval given in Eq. (3.68) to write $\dot{r}^2$ as a function of radius. It is at this point in our derivation where the distinction between massive and massless particles becomes important.

Massive Case

If we continue to parametrize spacetime coordinates by the proper time, Eq. (3.68) can be recast as

$$c^2 = \left(1 - \frac{R_S}{r}\right)c^2\dot{t}^2 - \left(1 - \frac{R_S}{r}\right)^{-1}\dot{r}^2 - r^2\dot{\phi}^2\,, \tag{3.77}$$

which yields

$$\begin{aligned}\dot{r}^2 &= \left(1 - \frac{R_S}{r}\right)^2 c^2\dot{t}^2 - \left(1 - \frac{R_S}{r}\right)\left(r^2\dot{\phi}^2 + c^2\right) \\ &= \frac{\epsilon^2}{c^2} - \left(1 - \frac{R_S}{r}\right)\left(\frac{\ell^2}{r^2} + c^2\right)\,.\end{aligned} \tag{3.78}$$

Our next step is to compute the Christoffel symbols $\Gamma^r_{\mu\mu}$:

$$\Gamma^r_{tt} = \frac{R_S c^2}{2r^2}\left(1 - \frac{R_S}{r}\right) \tag{3.79a}$$

$$\Gamma^r_{rr} = -\frac{R_S}{2r^2}\left(1 - \frac{R_S}{r}\right)^{-1} \tag{3.79b}$$

$$\Gamma^r_{\phi\phi} = R_S - r\,. \tag{3.79c}$$

The radial equation (3.72b) then becomes $\ddot{r} = -\Phi'_{\text{eff}}(r)$, where the **effective potential** is given by

$$\Phi_{\text{eff}}(r) = -\frac{GM}{r} + \frac{\ell^2}{2r^2} - \frac{GM\ell^2}{c^2 r^3}\,. \tag{3.80}$$

If we were working in Newtonian mechanics, the effective potential would consist of a term arising from an external force plus a second "fictitious" force, the centrifugal term. There would be two constants of motion, which we have found to be the case in the Schwarzschild solution as well. The salient difference between Newtonian gravity and GR, at least in this case, is the third term in Eq. (3.80). It is reassuring that this term diminishes faster than the other two, because it implies that the relativistic effective potential reduces to its Newtonian counterpart at large radius. Moreover the factor of c^2 in the denominator means that this term only becomes important if $r \sim 2GM/c^2$, at which point the second and third terms become comparable in magnitude. Once again, the Schwarzschild radius proves to be the relevant length scale!

In keeping with the title of this subsection, we now compute the radii at which a massive particle can execute circular orbits in the effective potential we have established. This means determining the extrema of Φ_{eff}, just as we would for Newtonian gravity. We have

$$\Phi'_{\text{eff}} = \frac{GM}{r^2} - \frac{\ell^2}{r^3} + \frac{3GM\ell^2}{c^2 r^4} = 0\,. \tag{3.81}$$

Notice that when $\ell = 0$ only the Newtonian term survives, implying that only radial orbits are possible. When $\ell \neq 0$ we have the quadratic equation

$$r^2 - \frac{\ell^2}{GM}r + \frac{3\ell^2}{c^2} = 0\,. \tag{3.82}$$

Circular orbits are therefore possible at radii

$$r_\pm = \frac{\ell^2}{2GM} \pm \frac{\ell}{2}\sqrt{\frac{\ell^2}{G^2M^2} - \frac{12}{c^2}}\,. \tag{3.83}$$

Note that the descriminant is positive if $\ell > \ell_{\rm crit} \equiv 2\sqrt{3}GM/c$. In this case, the asymptotic behavior of the effective potential at small and large radii implies that $\Phi_{\rm eff}$ has a local maximum at r_- and a local minimum at r_+. This means that orbits at r_+ will be stable and those at r_- will be unstable. Of particular interest is the smallest value of r_+ at which a particle can execute a stable circular orbit. This occurs when $\ell = \ell_{\rm crit}$, in which case

$$r_+ = \frac{\ell_{\rm crit}^2}{2GM} = \frac{6GM}{c^2}\,, \tag{3.84}$$

three times the Schwarzschild radius.

Massless Case

It is straightforward to follow the same procedure to determine orbits of a massless particle, but with two important modifications. First, a photon (the massless particle of interest to us) experiences no proper time, so we must identify another quantity to describe its trajectory; this is known as an **affine parameter**. For this and subsequent discussions of light, an overdot will denote differentiation with respect to such a parameter, though an exact specification will not be necessary. The second difference between the present case and that of a massive particle is that we must replace Eq. (3.78) by

$$\dot{r}^2 = \frac{\epsilon^2}{c^2} - \frac{\ell^2}{r^2}\left(1 - \frac{R_S}{r}\right)\,. \tag{3.85}$$

The effective potential for a photon then turns out to be

$$\Phi_{\rm eff}(r) = \frac{\ell^2}{2r^2} - \frac{GM\ell^2}{c^2 r^3}\,. \tag{3.86}$$

Unsurprisingly, at least in retrospect, light is not subject to the Newtonian term, $-GM/r$. This greatly simplifies the task of finding the radii at which a circular orbit is possible. In fact, there is only one such radius: $r = 3GM/c^2$. This is half the value we obtained for the innermost stable circular orbit of a massive particle, and is independent of the photon's angular momentum. However, this orbit is unstable, as it occurs at a maximum of the effective potential. This leaves open the possibility of open orbits, where the trajectory of an incoming photon is altered by a massive object, but does not form a closed curve in space (Sect. 3.4.1).

3.4 Light Propagation in the Schwarzschild Metric

3.4.1 Deflection Angle

We found in Sect. 3.3.3 that null geodesics of the Schwarzschild metric describe curves in space that are not closed, i.e., stable orbits do not exist in that case. Instead, a photon approaches a massive object from a given direction, whereupon it is deflected by some amount before resuming a straight course through space. We wish to ascertain the deflection angle as a function of the distance of closest approach, r_0, to the massive object causing the light ray to bend. We set the origin of spatial coordinates at the deflector. As the mass of the deflector, M, tends to zero, the photon will follow a path ever closer to a straight line, in which case the initial azimuthal angle will approach $\phi = 0$, while the final angle will approach $\phi = \pi$. Thus, the change in azimuthal angle, $\Delta\phi$, will be π. Therefore, we compute the deflection angle $\hat{\alpha} = \Delta\phi - \pi$ (see Fig. 3.1). Note that the coordinate system we are now using differs from that of Sect. 2.1, but our final result will not depend on our choice of coordinates.

Combining Eqs. (3.76b) and (3.85), we obtain

$$\frac{\mathrm{d}\phi}{\mathrm{d}r} \equiv \frac{\dot{\phi}}{\dot{r}} = \pm\ell \left[r^2 \sqrt{\frac{\epsilon^2}{c^2} - \frac{\ell^2}{r^2}\left(1 - \frac{2m}{r}\right)} \right]^{-1} , \tag{3.87}$$

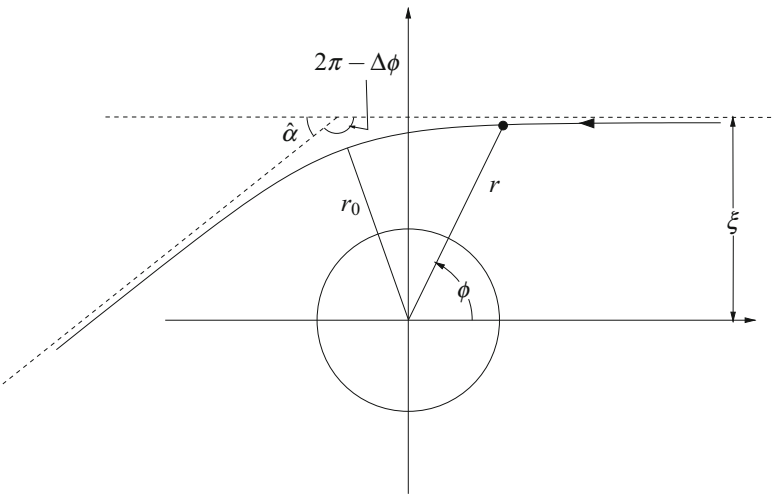


Fig. 3.1 A light ray with impact parameter ξ travels leftward from infinity ($\phi = 0$), and is "bent" by a deflector of mass M. This leads to a change in azimuthal angle of $\Delta\phi$, which is related to the deflection angle, $\hat{\alpha}$, by $\hat{\alpha} = \Delta\phi - \pi$. As expected, $\hat{\alpha} = 0$ when $M = 0$. In the weak-field limit, which we assume throughout this chapter, $\hat{\alpha}$ is small enough so that the distance of closest approach of the photon to the deflector $r_0 \approx \xi$

where we we introduce $m = GM/c^2$ as the mass in geometric units (i.e., length), which is half the Schwarzschild radius. We take the positive (negative) root of Eq. (3.85) for a light ray with r increasing (decreasing). Alternatively, we can take $\dot{r} > 0$ everywhere, and write

$$\begin{aligned} \Delta\phi &= -\int_{\infty}^{r_0} \frac{\dot{\phi}}{\dot{r}} \mathrm{d}r + \int_{r_0}^{\infty} \frac{\dot{\phi}}{\dot{r}} \mathrm{d}r \\ &= 2\int_{r_0}^{\infty} \frac{\mathrm{d}\phi}{\mathrm{d}r} \mathrm{d}r \,. \end{aligned} \tag{3.88}$$

In words, the first line says that a light ray far from the deflector will reach a minimum distance before traveling off to infinity again. Since both parts of the trajectory contribute equally to the overall deflection, we arrive at the second line. To perform the necessary integral, we must determine r_0. Such a turning point occurs at $\dot{r} = 0$, so we can write

$$\frac{\epsilon^2}{c^2} - \frac{\ell^2}{r_0^2}\left(1 - \frac{2m}{r_0}\right) = 0\,. \tag{3.89}$$

In the limit of vanishing mass, the distance of closest approach is the (invariant) impact parameter of the light ray, ξ, so we have

$$\xi = \frac{c\ell}{\epsilon}. \tag{3.90}$$

Introducing ξ allows us to write the condition $\dot{r} = 0$ in the presence of a massive deflector as

$$\frac{r_0^2}{\xi^2} = 1 - \frac{2m}{r_0}\,, \tag{3.91}$$

which can be solved for r_0 if necessary (see Problem 3.5).

We can now write the change in azimuth in terms of r_0 as

$$\begin{aligned} \Delta\phi &= 2\int_{r_0}^{\infty} \frac{\mathrm{d}\phi}{\mathrm{d}r} \mathrm{d}r \\ &= 2\int_{r_0}^{\infty} r^{-2}\left[\frac{1}{\xi^2} - \frac{1}{r^2}\left(1 - \frac{2m}{r}\right)\right]^{-1/2} \mathrm{d}r \\ &= 2\int_{r_0}^{\infty} r^{-2}\left[\frac{1}{r_0^2}\left(1 - \frac{2m}{r_0}\right) - \frac{1}{r^2}\left(1 - \frac{2m}{r}\right)\right]^{-1/2} \mathrm{d}r\,. \end{aligned} \tag{3.92}$$

Changing variables to $u = r_0/r$ leads to

$$\Delta\phi = 2\int_0^1 \left[(1-u^2) - \frac{2m}{r_0}(1-u^3)\right]^{-1/2} \mathrm{d}u\,. \tag{3.93}$$

While this expression is complete, it is not particularly enlightening. We now assume that the deflector represents a small perturbation to flat spacetime, thus allowing us to make a Taylor series expansion. Specifically, we assume $m/r_0 \ll 1$ so we can expand the integrand as

$$\Delta\phi = 2\int_0^1 \left[\frac{1}{(1-u^2)^{1/2}} + \frac{1-u^3}{(1-u^2)^{3/2}}\,\frac{m}{r_0} + \mathscr{O}\left(\frac{m}{r_0}\right)^2\right] \mathrm{d}u\,. \tag{3.94}$$

We can now integrate term by term (using the trigonometric substitution $u = \sin\vartheta$), and obtain

$$\Delta\phi = \pi + \frac{4m}{r_0} + \mathscr{O}\left(\frac{m}{r_0}\right)^2. \tag{3.95}$$

At lowest order, the change in azimuth is π as expected. The deflection is characterized by the first correction term, and at this order of approximation we can set $r_0 = \xi$. Then recalling $m = GM/c^2$ we finally obtain

$$\boxed{\hat{\alpha} = \frac{4GM}{c^2\xi}}\,, \tag{3.96}$$

which is twice the Newtonian value!

3.4.2 Time Delay

An important consequence of light bending is that the spatial path taken by a deflected light ray will be longer than that taken by an undeflected ray, implying a greater travel time for the former. However, the presence of the deflector entails an additional gravitational time delay, the so-called **Shapiro delay** (Shapiro 1964). The most straightforward approach is to compute the total travel time, then subtract off the geometric part. The Shapiro delay was proposed as a test of general relativity independent of the previously measured bending of light. A radar signal would be sent from Earth to Mercury or Venus, and bounced back to the radio telescope that emitted the original signal. It turns out, though, that the geometric and gravitational time delays can be combined into the **Fermat potential**, whose extrema correspond to the positions of the lensed images. We will merely determine the time delay in this section, but wait until Chap. 4 to derive the lens equation.

Whereas our calculation of the deflection angle involved integrating $\mathrm{d}\phi/\mathrm{d}r = \dot{\phi}/\dot{r}$, the time delay is given by the integral of $\mathrm{d}t/\mathrm{d}r = \dot{t}/\dot{r}$. However, the limits of integration must be finite in order for the result to be finite. Intuitively, a light ray will take an infinitely long time to travel an infinite distance. This means that the setup depicted in Fig. 3.1 does not quite apply to the present situation: a photon with $\phi \to 0$ corresponds to $r \to \infty$. The easiest fix is to mark out two points along the solid line that would indicate the initial and final positions, i.e., source and image positions, respectively. We also must ensure that the time delay is positive, as we would expect on physical grounds. This means that we must integrate from smaller to larger (Schwarzschild) radial coordinate.

Consider a photon emitted by a source at radius r_{src} from the center of a deflector of mass M. Suppose that the photon is received by an observer at r_{obs}. Taking the distance of closest approach to the deflector as r_0, the travel time from r_0 to some larger radius R is given by

$$T(R) = \int_{r_0}^{R} \frac{\mathrm{d}t}{\mathrm{d}r}\mathrm{d}r \tag{3.97}$$
$$= \int_{r_0}^{R} \left(1-\frac{2m}{r}\right)^{-1} \left[1-\frac{r_0^2}{r^2}\left(1-\frac{2m}{r}\right)\left(1-\frac{2m}{r_0}\right)^{-1}\right]^{-1/2} \mathrm{d}r\,.$$

We expect $T(R)$ to be of order R/c, so what we actually compute is cT/R, which is of order unity. We change variables to $u = r_0/r$ and simplify the notation by introducing the quantities $h = m/r_0$, which is the combination we saw in the analysis of the bending angle, and $w = r_0/R$. With these substitutions, we can write

$$\frac{cT(R)}{R} = w\int_w^1 u^{-2}(1-2hu)^{-1}\left(1-u^2\frac{1-2hu}{1-2h}\right)^{-1}\mathrm{d}u\,. \tag{3.98}$$

As before, we want to make a Taylor series expansion, but now we must take care because the time delay analysis involves three scales: m, r_0, and R. We know that r_0 will be of order the Einstein radius, and omitting factors of order unity we can write[3] $R_E \sim \sqrt{mR}$. Notice that

$$h = \frac{m}{r_0} \sim \left(\frac{m}{R}\right)^{1/2} \qquad \text{and} \qquad w = \frac{r_0}{R} \sim \left(\frac{m}{R}\right)^{1/2}\,. \tag{3.99}$$

In other words, r_0 is the geometric mean of m and R, such that h and w wind up being of the same order. Thus, when we make a Taylor series expansion, we need

[3] Strictly speaking, the Einstein radius depends on three distances between the observer, lens, and source. We will address the details momentarily, but for this argument we assume that the distances are all of order R.

to treat h and w jointly. First we return to Eq. (3.98) and expand the integrand as a Taylor series in h:

$$\frac{cT(R)}{R} = w\int_0^1 \left[\frac{1}{u^2\sqrt{1-u^2}} + \frac{2+3u}{u(1+u)\sqrt{1-u^2}}h + \mathcal{O}(h)^2\right] \mathrm{d}u \tag{3.100}$$
$$= \sqrt{1-w^2} + \left[\sqrt{\frac{1-w}{1+w}} + 2\ln\left(\frac{1+\sqrt{1-w^2}}{w}\right)\right] wh + w\mathcal{O}(h)^2.$$

Working to second order in w, the first term becomes $1 - w^2/2$. The combination in square brackets becomes $1 + 2\ln(2/w)$. This does not need to be expanded further because it is multiplied by wh which is already second order. Thus, we obtain

$$\frac{cT(R)}{R} = 1 - \frac{w^2}{2} + \left(1 - 2\ln\frac{w}{2}\right) wh + \mathcal{O}(\cdots)^3\,, \tag{3.101}$$

where $\mathcal{O}(\cdots)^3$ is any third order combination of h and w. Lest we lose track of the physics, let us reinsert the scales by writing out h and w:

$$cT(R) = R - \frac{R}{2}\left(\frac{r_0}{R}\right)^2 + m\left(1 - 2\ln\frac{r_0}{2R}\right) + \mathcal{O}^3(\cdots)\,. \tag{3.102}$$

The total travel time of a photon in going from an observer at radius $r_{\rm obs}$ to a radius $r_{\rm src}$ is $T_{\rm tot} \equiv T(r_{\rm obs}) + T(r_{\rm src})$. We saw in Chap. 2 that distances in lensing are measured by angles on the sky, namely the source angle, β, and the image angle, θ. Since the change of direction of a lensed light ray is confined to the proximity of the deflector, we may suppose that the bending takes place in the plane of the lens. We define the distances between the planes of the observer, lens, and source as d_l, d_{ls}, and d_s. Note that these are *coordinate* distances. Unlike the angular diameter distances D_l, D_{ls}, and D_s, shown in Fig. 2.2, $d_{ls} \approx d_s - d_l$.

The radii of the observer and source can be written as

$$r_{\rm obs} = d_l \tag{3.103a}$$
$$r_{\rm src} = \sqrt{d_{ls}^2 + d_s^2\tan^2\beta} \;\approx\; d_{ls}\left(1 + \frac{1}{2}\frac{d_s^2}{d_{ls}^2}\beta^2\right), \tag{3.103b}$$

where we expand $r_{\rm src}$ to second order using the small-angle approximation. In that approximation, $r_0/r_{\rm obs} \approx \theta$, and we can use Eq. (3.102) to write

$$cT(r_{\rm obs}) \approx d_l\left(1 - \frac{\theta^2}{2}\right) + m - 2m\ln\left(\frac{\theta}{2}\right). \tag{3.104}$$

The analysis for $cT(r_{\rm src})$ is similar, except we must use $r_0/r_{\rm src} \approx \theta r_{\rm obs}/r_{\rm src} \approx \theta d_l/d_{ls}$:

$$\begin{aligned} cT(r_{\rm src}) &\approx d_{ls}\left(1+\frac{\beta^2}{2}\frac{d_s^2}{d_{ls}^2}\right)\left(1-\frac{\theta^2}{2}\frac{d_l^2}{d_{ls}^2}\right)+m-2m\ln\left(\frac{\theta}{2}\frac{d_l}{d_{ls}}\right) \\ &\approx d_{ls}\left(1+\frac{\beta^2}{2}\frac{d_s^2}{d_{ls}^2}-\frac{\theta^2}{2}\frac{d_l^2}{d_{ls}^2}\right)+m-2m\ln\left(\frac{\theta}{2}\frac{d_l}{d_{ls}}\right), \end{aligned} \tag{3.105}$$

where we drop a term with $\theta^2\beta^2$ because we are working only to second order. Noting that the unlensed path length is

$$cT_0 \equiv \frac{d_s}{\cos\beta} \approx d_s\left(1+\frac{\beta^2}{2}\right), \tag{3.106}$$

the time delay can be written as

$$c\,\Delta T(\theta|\beta) \approx \frac{\beta^2}{2}\left(\frac{d_s^2}{d_{ls}}-d_s\right)-\frac{\theta^2}{2}\left(\frac{d_l^2}{d_{ls}}+d_l\right)+2m-2m\ln\left(\frac{\theta^2}{4}\frac{d_l}{d_{ls}}\right), \tag{3.107}$$

where the zeroth order term vanishes because $d_l+d_{ls}-d_s=0$. Using that relation, we realize that the distance combinations in the first two terms both simplify to $d_l d_s/d_{ls}$, yielding

$$c\,\Delta T(\theta|\beta) \approx \frac{d_l d_s}{2d_{ls}}\left(\beta^2-\theta^2\right)+2m-2m\ln\left(\frac{\theta^2}{4}\frac{d_l}{d_{ls}}\right). \tag{3.108}$$

Then using the lens equation $\beta=\theta-\theta_E^2/\theta$ we can manipulate the first term and obtain

$$c\,\Delta T(\theta|\beta) \approx \frac{d_l d_s}{2d_{ls}}(\theta-\beta)^2-4m\ln|\theta|+\text{const.} \tag{3.109}$$

This form is conventional because the first term reflects the extra travel time due to the additional path length (the geometric contribution to the time delay), while the second reflects the contribution from gravitational time dilation (the potential or Shapiro delay). All remaining terms are independent of θ and β, and they do not affect any observables (what we measure are differential time delays), so they can be neglected. Thus, our final expression for the time delay is

$$\Delta T(\theta|\beta) = \frac{d_l d_s}{2cd_{ls}}(\theta-\beta)^2-\frac{4GM}{c^3}\ln|\theta|\,. \tag{3.110}$$

Recall that $|\theta| \ll 1$ by assumption, so the gravitational time delay is positive as expected.

3.5 Friedmann-Robertson-Walker Cosmology

The Copernican Principle, which generalizes Copernicus' conclusion that Earth is not at the center of the solar system, led to the **Cosmological Principle**: *the universe is homogeneous and isotropic on large scales.* Its original formulation, which has come to be known as the *Perfect* Cosmological Principle, also assumed that the universe is static. This idea infamously led to Einstein's introduction of the "cosmological constant" into his field equations, and also to the steady-state theory of Hoyle, Bondi, and Gold (Hoyle 1948; Bondi and Gold 1948), which was largely abandoned by cosmologists after the discovery of the cosmic microwave background (see Sect. 9.1).

Our goal in this section is to write down the metric describing a homogeneous, isotropic spacetime. The most general form of this metric is the one described by Robertson (1935) and Walker (1935), although we will not prove this here. Instead, we use geometric and algebraic arguments to find a particular metric that satisfies the cosmological principle, but we will make no attempt to show that it is the only such metric. Furthermore, we argue purely on qualitative grounds that the Robertson-Walker (RW) metric is indeed homogeneous and isotropic. We begin by considering the spatial part of the metric where some degree of intuition can be brought to bear.

Any space that is homogeneous and isotropic will have constant curvature. There are three possibilities we must consider: zero curvature, positive curvature, and negative curvature. Current observations favor a *flat* (zero-curvature) universe, which is to say one in which the Pythagorean Theorem holds at any moment in time. But they do not rule out a universe of constant nonzero curvature, i.e., one that is either spherical (positive curvature) or hyperbolic (negative curvature). The two latter models are respectively termed *closed* and *open*. We describe each in turn, before introducing the RW metric in its full generality.

3.5.1 Homogeneous, Isotropic Universe

Geometry of a Closed Universe

Since no book related to cosmology would be complete without appealing to the analogy of an expanding balloon, we start by considering the geometry of a sphere. Such a universe is closed, since traveling in a fixed direction from any point will eventually lead back to that point. Although not necessary, it is convenient to think of a 2-sphere as a subset of three-dimensional Euclidean space. This is known as embedding, and is at the heart of our subsequent discussion. We refer the interested reader to any text on non-Euclidean geometry for a more thorough treatment.

We can think of a closed universe as a three-dimensional sphere embedded in a (fictitious) four-dimensional space. To get a feel for the terrain, we first consider the two-dimensional unit sphere. In terms of Cartesian coordinates (x, y, z), this sphere is defined by $x^2 + y^2 + z^2 = 1$. It is more convenient to work in spherical

coordinates. We start by defining the azimuthal angle, ϕ, between the x-axis and the projection of the vector $\mathbf{r} = (x, y, z)$ in the xy-plane. We then define the polar angle, θ, between $\mathbf{r}$ and the z-axis. These angles lie in the intervals $0 \leq \phi < 2\pi$ and $0 \leq \theta \leq \pi$. Formally, the Cartesian coordinates (x, y, z) are related to the spherical coordinates (θ, ϕ) through

$$x = \sin\theta \cos\phi \tag{3.111a}$$

$$y = \sin\theta \sin\phi \tag{3.111b}$$

$$z = \cos\theta \,. \tag{3.111c}$$

The distance between two points on the sphere of infinitesimal separation is described by the metric

$$\mathrm{d}S_2^2 = \mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2 \,. \tag{3.112}$$

We now turn to the three-dimensional unit sphere. To find its metric, we embed it in four-dimensional Euclidean space described by Cartesian coordinates (w, x, y, z). The equation to define a sphere is simply $w^2 + x^2 + y^2 + z^2 = 1$. Let ϕ be the angle between the x-axis and the projection of $\mathbf{r} = (w, x, y, z)$ into the xy-plane, and let θ be the angle between $\mathbf{r}$ and the z-axis. Since we are now dealing with a *three*-dimensional sphere, we must introduce a new angular coordinate to specify a given point. Accordingly, we define χ to be the angle between $\mathbf{r}$ and the w-axis. Like θ, the angle χ is restricted to the interval $0 \leq \chi \leq \pi$. The Cartesian coordinates (w, x, y, z) are related to the spherical coordinates (χ, θ, ϕ) by

$$w = \cos\chi \tag{3.113a}$$

$$x = \sin\chi \sin\theta \cos\phi \tag{3.113b}$$

$$y = \sin\chi \sin\theta \sin\phi \tag{3.113c}$$

$$z = \sin\chi \cos\theta \,. \tag{3.113d}$$

When $\chi = \pi/2$, Eq. (3.113a) reduces to $w = 0$, while the remaining three reduce to Eqs. (3.111). In other words, setting $\chi = \pi/2$ allows us to recover the two-dimensional unit sphere from its three-dimensional counterpart. If we make the further substitution $\theta = \pi/2$ in Eqs. (3.111), the unit 2-sphere reduces to the unit circle (1-sphere).

Our next step is to write down the metric describing the 3-sphere in terms of the coordinates (χ, θ, ϕ) without reference to the embedding space defined by (w, x, y, z). We have

$$\mathrm{d}S_3^2 = \mathrm{d}\chi^2 + \sin^2\chi \left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right) \,. \tag{3.114}$$

This metric can be obtained by brute force by substituting Eqs. (3.113) into the metric $\mathrm{d}S_3^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}w^2$. A more elegant method is to construct the $(n+1)$-sphere from the n-sphere, for any integer $n \geq 1$. Denoting the metric of the unit n-sphere by $\mathrm{d}S_n^2$, we have

$$\mathrm{d}S_{n+1}^2 = \mathrm{d}\vartheta_n^2 + \sin^2\vartheta_n \mathrm{d}S_{n-1}^2\,, \tag{3.115}$$

where the polar angles $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ are restricted to $[0, \pi]$. To begin the recursion, we note that the metric of the unit circle is simply $\mathrm{d}S_1^2 = \mathrm{d}\phi^2$, where ϕ is the azimuthal angle as before.

Geometry of an Open Universe

We would like to construct a three-dimensional space of constant *negative* curvature, analogous to our spherical model of a space of constant *positive* curvature. A simple variation of a spherical model would be to replace $\cos\chi$ and $\sin\chi$ with $\cosh\chi$ and $\sinh\chi$, respectively:

$$w = \cosh\chi \tag{3.116a}$$

$$x = \sinh\chi\sin\theta\cos\phi \tag{3.116b}$$

$$y = \sinh\chi\sin\theta\sin\phi \tag{3.116c}$$

$$z = \sinh\chi\cos\theta\,. \tag{3.116d}$$

These quantities satisfy the relation $w^2 - x^2 - y^2 - z^2 = 1$, which is the equation of a three-dimensional hyperboloid of revolution about the w-axis. However, this hypersurface does not have constant negative curvature under the Euclidean metric.

Fortunately, we can get around this problem by introducing the metric

$$\begin{aligned}\mathrm{d}H_3^2 &= -\mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \\ &= \mathrm{d}\chi^2 + \sinh^2\chi\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right)\,,\end{aligned} \tag{3.117}$$

where $0 \leq \chi < \infty$. If we had reversed the signs of the terms on the right-hand side of the first line, we would have obtained a negative definite metric, but we know the distance between neighboring points must be positive. Such would not the case for spacetime, however. We note that one can write the metric of an $(n+1)$-dimensional hyperboloid in terms of that for the unit n-sphere as

$$\mathrm{d}H_{n+1}^2 = \mathrm{d}\vartheta^2 + \sinh^2\vartheta\,\mathrm{d}S_n^2\,, \tag{3.118}$$

where $0 \leq \vartheta < \infty$.

3.5.2 Robertson-Walker Metric

There are two crucial steps we must take to obtain the metric describing an isotropic (and hence also homogeneous) spacetime. The first is to find a metric that can accommodate a closed, open, or flat universe, and the second is to put the "time" into spacetime. We commence by combining $\mathrm{d}S_3^2$, $\mathrm{d}H_3^2$, and the Euclidean metric $\mathrm{d}E_3^2$ of flat, three-dimensional space into a single formula:

$$\mathrm{d}\ell^2 = \mathrm{d}\chi^2 + f^2(\chi)\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right), \tag{3.119}$$

where

$$f(\chi) = \begin{cases} \sin\chi & \text{(spherical)} \\ \chi & \text{(flat)} \\ \sinh\chi & \text{(hyperbolic)}. \end{cases} \tag{3.120}$$

We can eliminate χ in favor of the radial coordinate $r \equiv f(\chi)$ to write the metric as

$$\mathrm{d}\ell^2 = \frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2), \tag{3.121}$$

where we define $k = 1$ for a spherical universe, $k = 0$ for a flat universe, and $k = -1$ for a hyperbolic universe.

Note that the coordinate r is dimensionless. However, it is more common for r to be a length. The curvature k must then have units of inverse area. Under this convention, k is not restricted to the values $\pm1, 0$. Instead, we require only that $k > 0$, $k = 0$, and $k < 0$ in the spherical, flat, and hyperbolic cases, respectively. Since $r = \chi$ in the flat universe, the two quantities must have the same dimension. Thus we replace Eq. (3.120) by the curvature-dependent relation

$$r \equiv f_k(\chi) = \begin{cases} \frac{1}{\sqrt{k}}\sin(\chi\sqrt{k}) & (k>0) \\ \chi & (k=0) \\ \frac{1}{\sqrt{-k}}\sinh(\chi\sqrt{-k}) & (k<0). \end{cases} \tag{3.122}$$

The most straightforward way to incorporate time is to write $\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a^2(t)\mathrm{d}\ell^2$, where the **scale factor** $a(t)$ allows for cosmic expansion or contraction. Thus,

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\right] \tag{3.123a}$$

$$= c^2\mathrm{d}t^2 - a^2(t)\left[\mathrm{d}\chi^2 + f_k^2(\chi)(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\right], \tag{3.123b}$$

which is known as the Robertson-Walker (RW) metric. The coordinates (r, θ, ϕ) and distance χ are said to be *comoving* because they "move" with the expansion or contraction of the universe. To measure physical distance, we define $r' = a(t)r$. Correspondingly, we define $k' = k/a(t)$. As expected, the curvature defined by k' decreases as the universe expands. In the absence of curvature ($k' = 0$), the RW metric reduces to the Minkowski metric in spherical coordinates. Moreover, the RW metric is locally Minkowskian, i.e., assumes the Minkowski form when $r' \ll 1/\sqrt{k'}$. Finally, the metric describes an isotropic spacetime. To see this, we observe that it is invariant under Lorentz transformation (i.e., boosts and spatial rotations). For example, reversing the sign of any of the four coordinates leaves the metric unchanged. Our discussion here is not intended to be mathematically rigorous, but is physically motivated.

3.5.3 *Friedmann's Equation*

Our next task is to understand how the cosmological scale factor $a(t)$ evolves. Rather than a full general-relativistic treatment, we give a simple Newtonian argument which leads to the same time dependence of $a(t)$. Consider a sphere of radius $R(t)$ and spatially uniform density $\rho(t)$. The energy per unit mass of a particle on the surface of the sphere is conserved and given by

$$\epsilon = \frac{1}{2}\dot{R}^2 - \frac{4\pi G}{3}\rho R^2 \equiv -c^2 K \,, \tag{3.124}$$

where the two terms on the right correspond to kinetic and potential energy, and K is a dimensionless constant. Depending on the value of K, the particle either escapes to infinity ($K \leq 0$) or attains a maximum displacement before falling toward the center of the sphere ($K > 0$).

In terms of the dimensionless scale factor $a(t)$, where $R(t) = R_0\, a(t)$, Eq. (3.124) takes the form

$$\boxed{\frac{1}{2}\dot{a}^2 - \frac{4\pi G}{3}\rho a^2 = -\frac{1}{2}c^2 k} \,, \tag{3.125}$$

where $k = 2K/R_0^2$. The boxed evolution equation, known as Friedmann's equation,[4] also follows from the full general-relativistic treatment combining the Robertson-Walker metric and the Einstein field equations, with k equal to the curvature discussed in Sect. 3.5.1.

[4]This relation is sometimes called Friedmann's first equation. Friedmann's second equation, which is equivalent to Newton's law of motion, will not be needed here.

In the Big Bang picture, there are three contributions to the density: matter, radiation, and dark energy. For each of these three components, the density has the form $\rho = \rho_0 a^{-n}$. For a matter-dominated universe, $n = 3$, since the density varies inversely with volume for a fixed total mass. For a radiation-dominated universe, there is an extra factor of a^{-1} from scaling the wavelength, and $n = 4$. For a universe dominated by dark energy or the cosmological constant, $n = 0$. According to these values of n, the density is dominated by radiation for $a \ll 1$ and by dark energy for $a \gg 1$, with an important matter contribution in the intermediate regime.

Solutions for a Flat Universe

It is straightforward to solve Friedmann's equation for $\rho = \rho_0 a^{-n}$. In the case $k = 0$ of a flat universe, the equation becomes

$$\frac{1}{2}\dot{a}^2 - \frac{4\pi G}{3}\rho_0 a^{2-n} = 0\,, \tag{3.126}$$

where ρ_0 is the current cosmic density. For the expansion following a Big Bang ($a = 0$ at $t = 0$), Eq. (3.126) predicts

$$a(t) = \left(\frac{2\pi G}{3}\rho_0\, n^2 t^2\right)^{1/n}\,. \tag{3.127}$$

Thus, $a(t)$ grows as $t^{2/3}$ and as $t^{1/2}$ for matter-dominated ($n = 3$) and radiation-dominated ($n = 4$) universes, respectively. In the case $n = 0$, corresponding to expansion driven by the cosmological constant or dark energy, Eq. (3.126) has the solution

$$a(t) = \exp\left[\sqrt{\frac{8\pi G}{3}\rho_0}\; t\right] \tag{3.128}$$

for $-\infty < t < \infty$, where $t = 0$ corresponds to the present.

Solution for the Einstein-de Sitter Universe

Friedmann's equation can also be solved analytically for a universe with nonzero curvature k and with $\rho = \rho_0 a^{-n}$. In the Einstein-de Sitter or matter-dominated case $n = 3$, we rescale the variables a and t according to

$$a = \frac{4\pi G\rho_0}{3c^2|k|}\,\hat{a}\,, \qquad t = \frac{4\pi G\rho_0}{3c^3|k|^{3/2}}\,\hat{t}\,. \tag{3.129}$$

Friedmann's equation then becomes

$$\left(\frac{\mathrm{d}\hat{a}}{\mathrm{d}\hat{t}}\right)^2 = \frac{2}{\hat{a}} - \mathrm{sgn}(k)\,, \tag{3.130}$$

where $\mathrm{sgn}(k) = k/|k|$ if $k \neq 0$ and $\mathrm{sgn}(0) = 0$. Equation (3.130) is readily solved by separation of variables

$$\int \sqrt{\frac{\hat{a}}{2 - \hat{a}\,\mathrm{sgn}(k)}}\,\mathrm{d}\hat{a} = \int \mathrm{d}\hat{t}\,, \tag{3.131}$$

and we choose the integration constant so that $\hat{a} = 0$ when $\hat{t} = 0$.

In the case $k > 0$ of positive curvature, corresponding to a closed universe, integrating Eq. (3.131) with the help of the substitution $\hat{a} = 1 - \cos\phi = 2\sin^2(\phi/2)$, yields

$$\hat{a} = 1 - \cos\phi \tag{3.132a}$$

$$\hat{t} = \phi - \sin\phi\,. \tag{3.132b}$$

The graph of $\hat{a}$ versus $\hat{t}$ has the form of a cycloid. The universe begins expansion from a point at $\hat{t} = 0$, reaches maximum size $\hat{a} = 2$ at $\hat{t} = \pi$, begins collapsing, and at $\hat{t} = 2\pi$ is again pointlike.

In the case $k < 0$ of negative curvature, corresponding to an open universe, integrating Eq. (3.131) with the substitution $\hat{a} = \cosh\phi - 1 = 2\sinh^2(\phi/2)$, yields

$$\hat{a} = \cosh\phi - 1 \tag{3.133a}$$

$$\hat{t} = \sinh\phi - \phi\,. \tag{3.133b}$$

As expected, $\hat{a}$ increases monotonically with time, with $\hat{a} \approx \frac{1}{2}(6\hat{t})^{2/3}$ for $0 < \hat{t} \ll 1$ and $\hat{a} \approx \hat{t}$ for $\hat{t} \gg 1$.

Friedmann's equation can also be solved analytically for radiation-dominated and dark-energy-dominated universes with nonzero curvature (see Problem 3.9).

The Hubble Parameter

In terms of the **Hubble parameter** $H \equiv \dot{a}/a$, Friedmann's equation (3.125) can be written as

$$H^2 = \frac{8\pi G\rho}{3} - \frac{c^2 k}{a^2}\,. \tag{3.134}$$

According to this relation, the critical density ρ_{crit}, which marks the threshold between an open universe ($k < 0$) and a closed universe ($k > 0$) is

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G} \,. \tag{3.135}$$

Combining this definition and Eq. (3.134) leads to

$$\rho - \rho_{\text{crit}} = \frac{3}{8\pi G} \frac{c^2 k}{a^2} \,. \tag{3.136}$$

Thus, the sign of $\rho - \rho_{\text{crit}}$ is the same as the sign of k, and this sign is fixed for all time.

To determine the present critical density, one needs the present value H_0 of the Hubble parameter. From observational data, Hubble (1929) concluded that the recession speed of galaxies beyond the Local Group varies with distance d according to the linear relation

$$v = H_0 \, d \,, \tag{3.137}$$

known as **Hubble's law**. The Hubble constant is often expressed as

$$H_0 \equiv 100 \, \text{h km s}^{-1} \, \text{Mpc}^{-1} \,, \tag{3.138}$$

where h is dimensionless. Measurements of h point to a value around 0.7 (Freedman and Madore 2010). According to Eqs. (3.136) and (3.138), the current value of the critical density is

$$\rho_{\text{crit}} = 1.8788 \times 10^{-26} \, \text{h}^2 \, \text{kg m}^{-3} \,, \tag{3.139}$$

of the order of a few hydrogen atoms per cubic meter. For recessional motion driven by the global expansion of the universe, $v/d = \dot{d}/d = \dot{a}/a$, and Hubble's law implies $\dot{a} = H_0 \, a$.

With the value of H_0 given above, Hubble's law holds approximately for galaxies which are outside the Local Group but not too distant. The Hubble parameter depends on distance and time through the scale factor, as we now consider. Although radiation was dominant in the early universe, it can now be neglected in comparison with the contributions of matter and dark energy. We define the constant quantities

$$\Omega_m = \frac{\rho_m}{\rho^0_{\text{crit}}} \,, \qquad \Omega_\Lambda = \frac{\rho_\Lambda}{\rho^0_{\text{crit}}} \,, \qquad \Omega_k = -\frac{c^2 k}{H_0^2} \,, \tag{3.140}$$

where ρ^0_{crit} is the current value of the critical density and Ω_k is the (dimensionless) **curvature density**. Then Friedmann's equation (3.134) can be written as

$$H^2 = H_0^2(\Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda)\,. \tag{3.141}$$

At present, $H = H_0$ and $a = 1$, which implies $\Omega_k = 1 - \Omega_m - \Omega_\Lambda$.

It is convenient to express H in terms of redshift z, which it is a directly measurable quantity. Since the wavelength of light transforms with the same scale factor as the universe, light emitted with wavelength λ from a distant galaxy at time t reaches us at time t_0 with a shift in wavelength $\Delta\lambda \equiv z\lambda$, where

$$\frac{\lambda + \Delta\lambda}{\lambda} = 1 + z = \frac{a(t_0)}{a(t)}\,. \tag{3.142}$$

On setting $a(t_0) = 1$ and substituting Eq. (3.142) in Eq. (3.141), Friedmann's equation becomes

$$H^2(z) = H_0^2[\Omega_m(1+z)^3 + (1 - \Omega_m - \Omega_\Lambda)(1+z)^2 + \Omega_\Lambda]\,. \tag{3.143}$$

Galaxies in the neighborhood of a galaxy with redshift z obey Hubble's law (3.137), except that H_0 is replaced by $H(z)$.

Deviations from Hubble's law may also be quantified by expanding the scale factor about t_0 in the form

$$\begin{aligned} a(t) &= a(t_0) + \dot{a}(t_0)(t - t_0) + \frac{1}{2}\ddot{a}(t_0)(t - t_0)^2 + \cdots \\ &= a(t_0)\left(1 + H_0(t - t_0) - \frac{q_0}{2}H_0^2(t - t_0)^2 + \cdots\right)\,, \end{aligned} \tag{3.144}$$

where $H_0 \equiv H(t_0)$, and

$$q_0 = -\left(\frac{\ddot{a}a}{\dot{a}^2}\right)_{t=t_0} \tag{3.145}$$

is the *dimensionless* **deceleration parameter**. This parameter was expected to be positive, due to gravitational attraction, until the discovery of the acceleration attributed to dark energy (Riess et al. 1998; Perlmutter et al. 1999). To first order, $H \equiv \dot{a}(t)/a(t) = H_0\,[1 - (q_0 + 1)H_0(t - t_0)]$, which reduces to Hubble's law for $|t - t_0| \ll [(q_0 + 1)H_0]^{-1}$. In practice, one usually works with redshift rather than time.

3.5.4 Cosmological Distances

If we liken the increase in wavelength due to cosmic expansion to the Doppler effect, we can assign a recession speed v to a given redshift z. To first order in v/c we have

$$\Delta\lambda \equiv z\lambda = \frac{v}{c}\,\lambda\,, \tag{3.146}$$

where $\Delta\lambda > 0$ for recession. Hubble's law can then be written as $v = cz = H_0 d$. This only holds when $z \ll v/c$. The dependence of distance on redshift is the focus of this subsection.

The most frequently used cosmological distances are the *angular-diameter* distance and the *luminosity* distance. These can be related to the *comoving* distance χ in the RW metric (3.123), which, in the next paragraph, is expressed as an integral involving the Hubble parameter.

Comoving Distance

The **comoving distance** χ traversed by a photon emitted at time t from a distant source can be determined by setting $\mathrm{d}s = 0$ in the RW metric. For fixed θ and ϕ, we find

$$\chi = \int_t^0 \frac{c\,\mathrm{d}t'}{a(t')} = \int_a^1 \frac{c\,\mathrm{d}a}{a\dot{a}} = \int_a^1 \frac{c\,\mathrm{d}a}{a^2 H}\,. \tag{3.147}$$

The upper limits of integration are the current time and scale factor. Since $a = (1+z)^{-1}$, the comoving distance to the emitter is given by

$$\boxed{\chi(z) = c\int_0^z \frac{\mathrm{d}z'}{H(z')}} \tag{3.148}$$

in terms of the redshift of the source, where $H(z)$ is given by Eq. (3.143). Note that the comoving distance between objects with redshifts z_1 and $z_2 > z_1$ along the same line of sight is the difference $\chi(z_2) - \chi(z_1)$. Other cosmological distances discussed below do not necessarily have this property.

Angular-Diameter Distance

The **angular-diameter distance** of an object with width or transverse size Δw that subtends an angle $\Delta\phi$ on the sky is defined by

$$d^{\mathrm{ang}} \equiv \frac{\Delta w}{\Delta\phi}\,. \tag{3.149}$$

According to the RW metric (3.123),

$$\Delta w = a(t) f_k(\chi)\Delta\phi = (1+z)^{-1} r(z)\Delta\phi\,, \tag{3.150}$$

where $f_k(\chi)$ is given by Eq. (3.122) and $r(z) = f_k(\chi(z))$. Thus,

$$d^{\mathrm{ang}}(z) = (1+z)^{-1} r(z)\,. \tag{3.151}$$

Now consider an observer in a plane at redshift $z_1 \neq 0$ and a source in a plane at redshift $z_2 > z_1$. The angular-diameter distance from plane 1 to plane 2 is

$$\boxed{d_{12}^{\rm ang}(z_1, z_2) = (1+z_2)^{-1} r_{12}(z_1, z_2) \equiv (1+z_2)^{-1} f_k(\chi_2 - \chi_1)}\,, \tag{3.152}$$

where $\chi_i = \xi(z_i)$. In a *closed* universe,

$$\begin{aligned} r_{12} &= \frac{1}{\sqrt{k}} \sin\Big(\sqrt{k}(\chi_2 - \chi_1)\Big) \\ &= \frac{1}{\sqrt{k}} \Big[\sin(\sqrt{k}\chi_2)\cos(\sqrt{k}\chi_1) - \sin(\sqrt{k}\chi_1)\cos(\sqrt{k}\chi_2)\Big]\,. \end{aligned} \tag{3.153}$$

With the help of Eq. (3.122), this can be rewritten as

$$r_{12} = r_2\sqrt{1 - kr_1^2} - r_1\sqrt{1 - kr_2^2}\,. \tag{3.154}$$

It is straightforward to verify that this holds for all k. In particular, in a *flat* universe ($k = 0$), $r_{12} = r_2 - r_1$.

Substituting Eq. (3.151) in Eq. (3.154) and replacing the subscripts 1 and 2 by l and s, we obtain

$$\begin{aligned} (1+z_s) D_{ls} = (1+z_s) D_s \sqrt{1 - k(1+z_l)^2 D_l^2} \\ -(1+z_l) D_l \sqrt{1 - k(1+z_s)^2 D_s^2}\,, \end{aligned} \tag{3.155}$$

where D_l, D_s, and D_{ls} are the usual distances in discussions of lensing. Thus, even in the case $k = 0$ of a flat universe, $D_{ls} \neq D_s - D_l$.

Luminosity Distance

The distance to a source can also be inferred from the observed flux F and the luminosity (power output) L of the source. For a static, flat universe, this distance is

$$d = \sqrt{\frac{L}{4\pi F}}\,. \tag{3.156}$$

The expansion of the universe does not affect L but does alter F. This is because the energy of an emitted photon is diminished by a factor of $1 + z$. Meanwhile, the rate at which photons are received decreases by the same factor. At the end of the day, this means making the substitution $F \to (1+z)^{-2} F$. The luminosity distance is then

$$d^{\rm lum} = (1+z) d\,. \tag{3.157}$$

From the RW metric, the physical radius $d(t)$ of a sphere is given in terms of its coordinate radius r as $d(t) = a(t)r$. At the present time, this reduces to $d = r$. Thus

$$d^{\text{lum}}(z) = (1+z)r(z) = (1+z)^2 d^{\text{ang}}(z)\,. \tag{3.158}$$

The luminosity distance of a source at redshift z_2 as seen by an observer at z_1 is

$$\boxed{d_{12}^{\text{lum}}(z_1, z_2) = (1+z_2)r_{12}(z_1, z_2) = (1+z_2)^2 d_{12}^{\text{ang}}(z_1, z_2)}\,. \tag{3.159}$$

Problems

3.1 The Lorentz transformation for a boost with velocity v parallel to the x axis is given in Eq. (3.3).

(a) Show that for a boost with velocity $\mathbf{v}$ in an arbitrary direction

$$t' = \gamma\left(t - \frac{\mathbf{v}\cdot\mathbf{x}}{c^2}\right) \tag{3.160a}$$

$$\mathbf{x}' = \mathbf{x} + (\gamma - 1)|\mathbf{v}|^{-2}(\mathbf{v}\cdot\mathbf{x})\mathbf{v} - \gamma\,\mathbf{v}\,t\,. \tag{3.160b}$$

(b) The Lorentz transformation for a boost with velocity v parallel to the x axis is shown in matrix form in Eq. (3.5). From Eq. (3.160), find the corresponding 4×4 matrix for a boost with velocity $\mathbf{v}$ in an arbitrary direction.

3.2 Consider a set of curvilinear coordinates x^i in an n-dimensional Euclidian space. A corresponding set of basis vectors $\mathbf{e}_i$ is defined by expanding the infinitesimal displacement $\mathrm{d}\mathbf{r}$ as $\mathrm{d}\mathbf{r} = \mathrm{d}x^i\,\mathbf{e}_i$. In general these basis vectors are neither of unit length nor orthogonal, and their lengths and directions depend on position.

(a) Show that $\mathbf{e}_i \cdot \mathbf{e}_j = g_{ij}$, where g is the metric tensor.

(b) Now consider a different set of generalized coordinates and basis vectors, which we denote by $\tilde{x}^{\,i}$ and $\tilde{\mathbf{e}}_i$, respectively. Since the infinitesimal displacement vector is independent of the particular basis, $\mathrm{d}\mathbf{r} = \mathrm{d}x^i\,\mathbf{e}_i = \mathrm{d}\tilde{x}^{\,i}\,\tilde{\mathbf{e}}_i$. Show that the basis vectors satisfy $\tilde{\mathbf{e}}_i = \mathbf{e}_j\,a^j{}_i$, and evaluate $a^j{}_i$ in terms of derivatives of the coordinates x^i and $\tilde{x}^i$.

(c) An arbitrary vector $\mathbf{v}$ may be expanded as $\mathbf{v} = v^i\,\mathbf{e}_i = \tilde{v}^{\,i}\,\tilde{\mathbf{e}}_i$, in terms of the basis vectors introduced in part (b). Show that the coordinates satisfy $\tilde{v}^{\,i} = b^i{}_j\,v^j$, and evaluate $b^i{}_j$ in terms of derivatives of the coordinates x^i and $\tilde{x}^i$.

(d) Show that $b^j{}_i\,a^i{}_k = \delta^j{}_k$, i.e., matrix b is the inverse of matrix a. Thus, the tranformation of the coordinates "opposes" the transformation of the basis vectors, whence the term contravariant.

3.3 The basis vectors $\mathbf{e}_i$ defined in the preceding problem are neither of unit length nor orthogonal. However, a set of reciprocal basis vectors $\mathbf{e}^i$ can be constructed from them which satisfy the orthonormality condition $\mathbf{e}^i \cdot \mathbf{e}_j = \delta^i_{\ j}$. Considering a three dimensional space and making use of cross products such as $\mathbf{e}_1 \times \mathbf{e}_2$, construct $\mathbf{e}^1$, $\mathbf{e}^2$, and $\mathbf{e}^3$ from $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$.

Remark The basis vector $\mathbf{e}_i$ may be expanded in terms of the reciprocal basis in the form $\mathbf{e}_i = c_{ij}\mathbf{e}^j$. From the property $\mathbf{e}_i \cdot \mathbf{e}_j = g_{ij}$ (see preceding problem), it follows that $c_{ij} = g_{ij}$. Thus, $\mathbf{e}_i = g_{ij}\mathbf{e}^j$, i.e., the metric tensor serves to raise and lower the indices of the basis vectors.

3.4

(a) For a Euclidian (flat) three-dimensional space parametrized by Cartesian coordinates (x, y, z), find all elements of the metric tensor and of the Christoffel symbol or affine connection defined in Eq. (3.61).
(b) For a Euclidian (flat) three-dimensional space parametrized by spherical coordinates (r, θ, ϕ), with $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, find all elements of the metric tensor and of the corresponding Christoffel symbol.
(c) Does the Christoffel symbol $\Gamma^{\lambda}_{\mu\nu}$, which has three indices, transform, in general, like a tensor of rank three under a change of coordinates? Do your answers to parts (a) and (b) shed any light on this question?

3.5 Trigonometric substitutions are very useful in evaluating integrals. A much earlier use historically was in solving cubic equations (Viète 1591).

(a) Show that the general cubic equation $y^3 + ay^2 + by + c = 0$ may be rewritten as $x^3 - Ax = B$, without a quadratic term, by making a substitution of the form $y = x + x_0$. Express x_0, A, and B in terms of a, b, and c.
(b) Using the relation $\cos\theta = \frac{1}{2}\left(e^{i\theta} + e^{-i\theta}\right)$, prove the identity $4\cos^3\theta - 3\cos\theta = \cos 3\theta$.
(c) Substituting $x = k\cos\theta$ in $x^3 - Ax = B$ and comparing with the above trigonometric identity, solve for k, θ, and x in terms of A and B.
(d) Using the result of part (c), show that the solution to Eq. (3.91) is

$$r_0 = \frac{2|\xi|}{\sqrt{3}}\cos\left[\frac{1}{3}\cos^{-1}\left(-\frac{3^{3/2}m}{|\xi|}\right)\right]. \tag{3.161}$$

3.6 In Sect. 3.4.1 we derived the lowest order term in a Taylor expansion for the deflection angle in the Schwarzschild metric. Now derive the second-order term in the m/ξ expansion. Hint: It is natural to work first in terms of m/r_0 and then convert to m/ξ; if you do that, recall that $r_0 \approx \xi$ at lowest order, but there are correction terms that become important when you expand $\Delta\phi$ to second order in m/ξ.

3.7 In Sect. 2.1 we derived the Newtonian deflection angle to lowest order using the Born approximation. Now derive it using an analysis similar to what we did for the relativistic bending angle. Specifically, consider a test particle passing a mass M such that the impact parameter is ξ, the distance of closest approach is r_0, and

the initial speed is c. Use Newtonian mechanics to derive an equation of motion involving $\mathrm{d}\phi/\mathrm{d}r$, and then obtain an expression for $\Delta\phi$ that is analogous to the GR expression in Eq. (3.92). Recall from our earlier analysis that the lowest order terms should be

$$\Delta\phi \approx \pi + \frac{2m}{\xi}\,.$$

Verify this result, and also derive the second-order term in a series expansion in m/ξ.

3.8 Friedmann's equation is a nonlinear first-order differential equation for the scale factor $a(t)$. For an Einstein-de Sitter or matter-dominated universe ($n = 3$), that first-order differential equation is solved in Sect. 3.5.3. The goal of this problem is to solve Friedmann's equation for $n = 3$ in a different way, first converting it to a linear second-order differential equation with constant coefficients.

(a) Show that the curvature k is given in terms of H_0 and $\Omega_0 \equiv \rho_0/\rho^0_{\mathrm{crit}}$ by

$$k = \frac{H_0^2}{c^2}(\Omega_0 - 1)\,.$$

(b) Introducing the parameter ϕ through the relation $\mathrm{d}\phi/\mathrm{d}t = f(a)$, find the function $f(a)$ for which Friedmann's equation takes the form

$$\left(\frac{\mathrm{d}a}{\mathrm{d}\phi}\right)^2 \pm a^2 = \frac{\Omega_0}{|\Omega_0 - 1|}\,a\,,$$

where the upper and lower signs correspond to positive and negative curvature, respectively.

(c) Show that differentiating the above equation with respect to ϕ leads to a linear second-order differential equation, reminiscent of Newton's law of motion for a driven harmonic oscillator. Solve the differential equation for a as a function of ϕ, using the initial conditions $a = 0$ and $\phi = 0$ at $t = 0$. Show that the solution is equivalent to the one in Sect. 3.5.3.

3.9 Using the method of Sect. 3.5.3 for a matter-dominated universe, solve Friedmann's equation with nonzero curvature in the cases $n = 4$ (radiation domination) and $n = 0$ (domination by the cosmological constant or dark energy).

References

Bondi, H., & Gold, T. (1948). The steady-state theory of the expanding universe. *Monthly Notices of the Royal Astronomical Society, 108*, 252.

Freedman, W. L., & Madore, B. F. (2010). The Hubble constant. *Annual Review of Astronomy and Astrophysics, 48*, 673.

Hoyle, F. (1948). A new model for the expanding universe. *Monthly Notices of the Royal Astronomical Society, 108*, 372.

Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences, 15*, 168.

Peacock, J. A. (1999). *Cosmological physics*. Cambridge: Cambridge University Press.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., et al. (1999). Measurements of Ω and Λ from 42 high-redshift supernovae. *Astrophysical Journal, 517*, 565.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., et al. (1998). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astronomical Journal, 116*, 1009.

Robertson, H. P. (1935). Kinematics and world-structure. *Astrophysical Journal, 82*, 284.

Shapiro, I. I. (1964). Fourth test of general relativity. *Physical Review Letters, 13*, 789.

Viète, F. (1591). De aequationum recognitione et emendatione, in his *Opera*, 82–162.

Walker, A. G. (1935). On the formal comparison of Milne's kinematical system with the systems of general relativity. *Monthly Notices of the Royal Astronomical Society, 95*, 263.

Chapter 4
Multiple Imaging in the Weak-Field Limit

Now that we have had some practice in carrying out lensing calculations and have derived the all-important factor of two, we turn to non-axisymmetric lens models in the context of the thin lens approximation. We show in Sect. 4.1 that the vector counterpart of the deflection angle is given as the gradient of a scalar potential and uses Fermat's principle to obtain the lens equation. We then solve this equation to find the image positions for a simple but important special case. What follows thereafter is a general discussion of image magnification (Sect. 4.2) and time delay (Sect. 4.3), which can be observed and compared against specified lens models. Some general properties of the lens mapping are described in Sects. 4.4 and 4.5. We derive the conservation of surface brightness in Sect. 4.6, which allows us to consider lensing of spatially extended sources. Mathematical degeneracies in the lens equation, which constrain our ability to compare theory and observation, are discussed in Sect. 4.7. The case of a light ray deflected by multiple lenses (Sect. 4.8) rounds out our presentation of the theory of strong lensing in the weak-field limit.

4.1 Lens Equation

4.1.1 Lens Potential

Since we are now dealing with general lenses, the vectorial nature of the deflection angle must be explicitly taken into account. As in electrostatics and gravitation, we can define a scalar potential whose gradient gives the vector field of interest. Recall from Eq. (2.6) that the deflection vector is given by

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{2}{c^2}\int_{-\infty}^{\infty} \boldsymbol{\nabla}_{\boldsymbol{\xi}}\Phi(\boldsymbol{\xi}, z)\mathrm{d}z \equiv \boldsymbol{\nabla}_{\boldsymbol{\xi}}\hat{\psi}(\boldsymbol{\xi})\,, \tag{4.1}$$

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_4

where $\nabla_{\boldsymbol{\xi}} \equiv \partial/\partial\boldsymbol{\xi}$, which is identical to $\nabla_\perp$ in the notation of Chap. 2. The function $\hat{\psi}$ is given by

$$\hat{\psi}(\boldsymbol{\xi}) \equiv \frac{2}{c^2}\int_{-\infty}^{\infty} \Phi(\boldsymbol{\xi}, z)\mathrm{d}z\,. \tag{4.2}$$

Although the focus of this chapter is on general lenses described by $\hat{\psi}(\boldsymbol{\xi})$, it is instructive to begin with the case of a point mass. In terms of $R \equiv |\boldsymbol{\xi}|$, the corresponding gravitational potential in three dimensions is

$$\Phi(R, z) = -\frac{GM}{\sqrt{R^2 + z^2}}\,. \tag{4.3}$$

The projected potential is then

$$\begin{aligned}\hat{\psi}_{\mathrm{PM}}(R) &= -\frac{2GM}{c^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}z}{\sqrt{R^2+z^2}} \\ &= -\frac{2GM}{c^2}\ln\left|\sqrt{1+\frac{z^2}{R^2}}+\frac{z}{R}\right|_{-\infty}^{\infty}\,,\end{aligned} \tag{4.4}$$

which diverges. However, recall that the actual source, lens, and observer coordinates are $z = -D_{ls}$, $z = 0$, and $z = D_l$, respectively. Thus,

$$\begin{aligned}\hat{\psi}_{\mathrm{PM}}(R) &= -\frac{2GM}{c^2}\ln\left(\sqrt{1+\frac{z^2}{R^2}}+\frac{z}{R}\right)\Bigg|_{-D_{ls}}^{D_l} \\ &\approx -\frac{2GM}{c^2}\left[\ln\left(\frac{2D_l}{R}\right)+\ln\left(\frac{2D_{ls}}{R}\right)\right]\,,\end{aligned} \tag{4.5}$$

where $R \ll D_l$ and $R \ll D_{ls}$ if the deflection is small.

For an arbitrary mass distribution described by the density $\rho(\mathbf{r}) \equiv \rho(\boldsymbol{\xi}, z)$, the gravitational potential can be written as

$$\begin{aligned}\Phi &= -G\int_{\mathbb{R}^3}\frac{\rho(\mathbf{r}')\mathrm{d}^3r'}{|\mathbf{r}-\mathbf{r}'|} \\ &= -G\int_{\mathbb{R}^3}\frac{\rho(\boldsymbol{\xi}', z')\mathrm{d}^2\xi'\mathrm{d}z'}{\sqrt{|\boldsymbol{\xi}-\boldsymbol{\xi}'|^2+(z-z')^2}} \\ &= -G\int_{\mathbb{R}^3}\frac{\Sigma(\boldsymbol{\xi}')\delta(z')\mathrm{d}^2\xi'\mathrm{d}z'}{\sqrt{|\boldsymbol{\xi}-\boldsymbol{\xi}'|^2+(z-z')^2}}\end{aligned}$$

$$= -G \int_{\mathbb{R}^2} \frac{\Sigma(\boldsymbol{\xi}') \mathrm{d}^2 \xi'}{\sqrt{|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2 + z^2}}, \tag{4.6}$$

where we have used the thin lens approximation

$$\rho(\boldsymbol{\xi}', z') = \Sigma(\boldsymbol{\xi}') \delta(z') . \tag{4.7}$$

Taking the gradient perpendicular to the line of sight, we have

$$\nabla_{\boldsymbol{\xi}} \Phi(\boldsymbol{\xi}, z) = G \int_{\mathbb{R}^2} \frac{\Sigma(\boldsymbol{\xi}') \left(\boldsymbol{\xi} - \boldsymbol{\xi}'\right) \mathrm{d}^2 \xi'}{(|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2 + z^2)^{\frac{3}{2}}} . \tag{4.8}$$

The deflection vector is then

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) &= \frac{2G}{c^2} \int_{\mathbb{R}^2} \Sigma(\boldsymbol{\xi}') \left(\boldsymbol{\xi} - \boldsymbol{\xi}'\right) \mathrm{d}^2 \xi' \int_{-\infty}^{\infty} \left(\left|\boldsymbol{\xi} - \boldsymbol{\xi}'\right|^2 + z^2\right)^{-\frac{3}{2}} \mathrm{d}z \\ &= \frac{4G}{c^2} \int_{\mathbb{R}^2} \frac{\Sigma(\boldsymbol{\xi}') \left(\boldsymbol{\xi} - \boldsymbol{\xi}'\right) \mathrm{d}^2 \xi'}{|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2} \\ &= \frac{4G}{c^2} \nabla_{\boldsymbol{\xi}} \int_{\mathbb{R}^2} \Sigma(\boldsymbol{\xi}') \ln\left(\left|\boldsymbol{\xi} - \boldsymbol{\xi}'\right|\right) \mathrm{d}^2 \xi' , \end{aligned} \tag{4.9}$$

where we use the relation

$$\nabla_{\boldsymbol{\xi}} \ln\left(\left|\boldsymbol{\xi} - \boldsymbol{\xi}'\right|\right) = \frac{\boldsymbol{\xi} - \boldsymbol{\xi}'}{|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2} . \tag{4.10}$$

We can then write

$$\hat{\psi}(\boldsymbol{\xi}) = \frac{4G}{c^2} \int_{\mathbb{R}^2} \Sigma(\boldsymbol{\xi}') \ln\left(\left|\boldsymbol{\xi} - \boldsymbol{\xi}'\right|\right) \mathrm{d}^2 \xi' . \tag{4.11}$$

Instead of starting from Eq. (4.1), we could have derived Eq. (4.9) by noting that it is the straightforward generalization of the deflection due to a point mass (2.10). That approach seems much simpler, but it does not make explicit the assumption of a thin lens (4.7).

In terms of the angular vector $\boldsymbol{\theta} \equiv \boldsymbol{\xi}/D_l$, the **reduced deflection vector** is given by

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) \equiv \frac{D_{ls}}{D_s} \hat{\boldsymbol{\alpha}}(D_l \boldsymbol{\theta}) . \tag{4.12}$$

Defining

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) \equiv \nabla \psi(\boldsymbol{\theta}), \tag{4.13}$$

the **reduced lens potential** takes the form

$$\begin{aligned}\psi(\boldsymbol{\theta}) &\equiv \frac{D_{ls}}{D_l D_s}\hat{\psi}(D_l\boldsymbol{\theta}) \\ &= \frac{4G}{c^2}\frac{D_l D_{ls}}{D_s}\int_{\mathbb{R}^2}\Sigma(D_l\boldsymbol{\theta}')\ln\left(\left|\boldsymbol{\theta}-\boldsymbol{\theta}'\right|\right)\mathrm{d}^2\theta' \\ &= \frac{1}{\pi}\int_{\mathbb{R}^2}\kappa(\boldsymbol{\theta}')\ln\left(\left|\boldsymbol{\theta}-\boldsymbol{\theta}'\right|\right)\mathrm{d}^2\theta' , \end{aligned} \tag{4.14}$$

where the convergence of a two-dimensional mass distribution is

$$\kappa(\boldsymbol{\theta}) \equiv \frac{\Sigma(D_l\boldsymbol{\theta})}{\Sigma_{\mathrm{crit}}} , \tag{4.15}$$

where Σ_{crit} is defined by Eq. (2.68).

Using Eq. (4.5), we obtain

$$\begin{aligned}\psi_{\mathrm{PM}}(\boldsymbol{\theta}) &= \frac{D_{ls}}{D_l D_s}\hat{\psi}_{\mathrm{PM}}(D_l|\boldsymbol{\theta}|) \\ &= -\frac{2GM}{c^2}\frac{D_{ls}}{D_l D_s}\left[\ln\left(\frac{2D_l}{D_l|\boldsymbol{\theta}|}\right)+\ln\left(\frac{2D_{ls}}{D_l|\boldsymbol{\theta}|}\right)\right] \\ &= \frac{2GM}{c^2}\frac{D_{ls}}{D_l D_s}\left[2\ln|\boldsymbol{\theta}|-\ln\left(\frac{4D_{ls}}{D_l}\right)\right] \end{aligned} \tag{4.16}$$

for the reduced lens potential of a point mass at the origin. Since the gradient of the potential is the physical quantity of interest, we can ignore terms that are independent of $\boldsymbol{\theta}$, as we did in going from Eq. (4.11) to Eq. (4.14). This leaves us with

$$\psi_{\mathrm{PM}}(\boldsymbol{\theta}) = \frac{4GM}{c^2}\frac{D_{ls}}{D_l D_s}\ln|\boldsymbol{\theta}| \tag{4.17a}$$

$$= \frac{M}{\pi D_l^2 \Sigma_{\mathrm{crit}}}\ln|\boldsymbol{\theta}| . \tag{4.17b}$$

The potential at $\boldsymbol{\theta}$ due to a point mass with $M = D_l^2\Sigma_{\mathrm{crit}}$ at $\boldsymbol{\theta}'$ is

$$g(\boldsymbol{\theta}-\boldsymbol{\theta}') = \frac{1}{\pi}\ln|\boldsymbol{\theta}-\boldsymbol{\theta}'| . \tag{4.18}$$

If we compare this with Eq. (4.14), we see that the potential of an arbitrary mass distribution is the convolution of the point-mass potential and the convergence, i.e.,

$$\psi(\boldsymbol{\theta}) = \int_{\mathbb{R}^2}\kappa(\boldsymbol{\theta}')g(\boldsymbol{\theta}-\boldsymbol{\theta}')\mathrm{d}^2\theta' . \tag{4.19}$$

We can invert Eq. (4.14) to find the convergence $\kappa(\boldsymbol{\theta})$ for a given potential $\psi(\boldsymbol{\theta})$. We start by writing

$$\begin{aligned}\nabla^2\psi(\boldsymbol{\theta}) &= \nabla^2\left[\int_{\mathbb{R}^2}\kappa(\boldsymbol{\theta}')g(\boldsymbol{\theta}-\boldsymbol{\theta}')\mathrm{d}^2\theta'\right] \\ &= \int_{\mathbb{R}^2}\kappa(\boldsymbol{\theta}')\nabla^2 g(\boldsymbol{\theta}-\boldsymbol{\theta}')\mathrm{d}^2\theta' \,. \end{aligned} \tag{4.20}$$

Since

$$\nabla^2 g(\boldsymbol{\theta}-\boldsymbol{\theta}') = 2\delta^{(2)}(\boldsymbol{\theta}-\boldsymbol{\theta}') \tag{4.21}$$

(see Problem 4.1), we can write

$$\nabla^2\psi(\boldsymbol{\theta}) \;=\; 2\int_{\mathbb{R}^2}\kappa(\boldsymbol{\theta}')\delta^{(2)}(\boldsymbol{\theta}-\boldsymbol{\theta}')\mathrm{d}^2\theta' \;=\; 2\kappa(\boldsymbol{\theta})\,. \tag{4.22}$$

This is *Poisson's equation* in two dimensions.

To complete our discussion of the lens potential, we describe how to solve Poisson's equation (4.22) for a general convergence $\kappa(\boldsymbol{\theta})$. The convergence at $\boldsymbol{\theta}$ due to a point mass at $\boldsymbol{\theta}'$ is simply $\delta^{(2)}(\boldsymbol{\theta}-\boldsymbol{\theta}')$. Poisson's equation in this case is then identical to Eq. (4.21). The solution $g(\boldsymbol{\theta}-\boldsymbol{\theta}')$ is called **Green's function** and is given by Eq. (4.18). Since Poisson's equation is linear, the potential due to an arbitrary mass distribution $\kappa(\boldsymbol{\theta})$ is the appropriate superposition of Green's functions for point masses at different locations (see Appendix D.3 for more details on the Green's function method). The result is Eq. (4.14), as desired.

4.1.2 Fermat's Principle

If we compare the lens potential of a point mass (Eq. (4.17a)) with the gravitational contribution to the time delay in the Schwarzschild metric (Eq. (3.110)), we notice something curious: apart from constant factors, one is the negative of the other! To understand the connection between the lens potential and the gravitational time delay, we must adapt Eq. (3.110) to a cosmological context. This involves two steps. First, we must interpret the Schwarzschild coordinate distances d as angular diameter distances D. This is justified, since the observer and source are far from the lens, where spacetime is approximately described by the RW metric. The second step is a bit more subtle: Eq. (3.110) gives the time delay relative to the lens, but we seek the time delay as measured by an observer, whose distance from the lens increases with the cosmic expansion. This introduces a time dilation factor of $a^{-1}(t)_l = 1 + z_l$.

The total (geometric plus gravitational) time delay due to a Schwarzschild lens lying in an expanding universe described by the RW metric is then

$$\begin{aligned}\Delta T_{\rm PM}(\theta|\beta) &= \frac{1+z_l}{c}\left[\frac{D_l D_s}{2D_{ls}}(\theta-\beta)^2 - \frac{4GM}{c^2}\ln|\theta|\right] \\ &= T_*\left[\frac{1}{2}(\theta-\beta)^2 - \psi_{\rm PM}(\theta)\right], \end{aligned} \tag{4.23}$$

where

$$T_* = \frac{1+z_l}{c}\,\frac{D_l D_s}{D_{ls}}\,, \tag{4.24}$$

$\psi_{\rm PM}$ is given by Eq. (4.17a), and the source position β is fixed. It is straightforward to extend Eq. (4.23) to a general potential by noting that $\Delta T_{\rm PM}$ is linear in $\psi_{\rm PM}$; the potential due to a collection of point masses (or infinitesimal mass elements) at redshift z_l is just the sum (integral) of the individual potentials. We must also replace the scalar source and image positions by their vector counterparts. If we denote the potential of an arbitrary mass distribution at redshift z_l by $\psi(\boldsymbol{\theta})$, we can write the time delay in units of T_* as

$$\tau(\boldsymbol{\theta}|\boldsymbol{\beta}) \equiv \frac{\Delta T(\boldsymbol{\theta}|\boldsymbol{\beta})}{T_*} = \frac{1}{2}\,|\boldsymbol{\theta}-\boldsymbol{\beta}|^2 - \psi(\boldsymbol{\theta})\,, \tag{4.25}$$

where τ is known as the **Fermat potential**.[1]

Once $\tau(\boldsymbol{\theta}|\boldsymbol{\beta})$ is known, the image position(s) $\boldsymbol{\theta}$ for a given source position $\boldsymbol{\beta}$ may be found from Fermat's principle, also known as the principle of extremal time. We look for stationary points of τ, where $\boldsymbol{\nabla}\tau = \mathbf{0}$. Using Eq. (4.25) we obtain

$$\begin{aligned}\boldsymbol{\beta} &= \boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta}) \\ &\equiv \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})\,, \end{aligned} \tag{4.26}$$

which is the vector generalization of the lens equation (2.18). But there is a problem with this derivation: the geometric contribution proportional to $(\theta-\beta)^2$ in Eqs. (3.110), (4.23), and (4.24) was obtained assuming a point mass. We now argue that this geometric term is generally valid.

From Problem 4.2, we know that the time delay in Euclidean geometry has the form

$$c\Delta T_{\rm geom} = \frac{D_l D_s}{2D_{ls}}(\theta-\beta)^2\,. \tag{4.27}$$

[1] This τ is not related to the proper time we encountered in Chap. 3.

Replacing θ and β by their vectorial counterparts $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ yields

$$c\Delta T_{\rm geom} = \frac{D_l D_s}{2D_{ls}}|\boldsymbol{\theta} - \boldsymbol{\beta}|^2 \,. \tag{4.28}$$

Schneider et al. (1992) show that this expression also holds if D_l, D_s, and D_{ls} are angular diameter distances, provided that we include a redshift factor $1 + z_l$ to account for the cosmic expansion. The final form for the geometric time delay is therefore

$$\Delta T_{\rm geom} = \frac{1+z_l}{c}\frac{D_l D_s}{2D_{ls}}|\boldsymbol{\theta} - \boldsymbol{\beta}|^2 \,, \tag{4.29}$$

which is completely consistent with the geometric time delay in Eqs. (4.23) and (4.24) and not limited to a point mass.

Having established the general form of the lens equation (4.26), we can now look for solutions following Chap. 2 but allowing more general geometry. For simplicity, we first consider cases where the lens equation can be recast in the form of a matrix equation, which can be inverted to determine the positions of lensed images.

4.1.3 Convergence and Shear

Suppose that a light ray undergoes sufficiently small deflection that the lens potential may be approximated by a Taylor series in the x and y components of $\boldsymbol{\theta}$. A deflection angle of linear order requires a quadratic lens potential:

$$\psi(x, y) \approx \psi^0 + \psi^0_x x + \psi^0_y y + \frac{1}{2}\left(\psi^0_{xx}x^2 + \psi^0_{yy}y^2 + 2\psi^0_{xy}xy\right) , \tag{4.30}$$

where the superscript "0" denotes evaluation at the origin and the subscripts indicate partial differentiation with respect to the specified coordinates. The constant can be neglected since the deflection angle (4.13) is independent of ψ^0. The linear terms can also be ignored, because they produce a constant deflection, which amounts to a translation $\boldsymbol{\beta} \rightarrow \boldsymbol{\beta} + \boldsymbol{\alpha}^0$, of the origin of the source plane, where $\boldsymbol{\alpha}^0 \equiv \psi^0_x \mathbf{e}_x + \psi^0_y \mathbf{e}_y$.

It is often convenient to write the lens potential as a multipole expansion, as in electrodynamics (Jackson 1962). In general, a narrow "bundle" of light rays undergoes both isotropic and anisotropic deflection, corresponding to monopole and higher-order terms, respectively. Because gravitation is purely attractive, the dipole term vanishes, and the lowest-order anisotropic deflection comes from the quadrupole term. In the language of lensing, the monopole and quadrupole moments are known as convergence (cf. Sects. 2.4 and 4.1.1) and shear (cf. Sect. 2.6), respectively. Since the light rays in the bundle are close together, we may assume

that they pass through a region of uniform density in the lens plane. This means the convergence (4.22) has the constant value

$$\kappa = \frac{1}{2}\left(\psi^0_{xx} + \psi^0_{yy}\right). \tag{4.31}$$

In terms of the shear components

$$\gamma_+ \equiv \frac{1}{2}\left(\psi^0_{xx} - \psi^0_{yy}\right) \tag{4.32a}$$

$$\gamma_\times \equiv \psi^0_{xy} \tag{4.32b}$$

and the convergence, the lens potential (4.30) can be written as

$$\begin{aligned}\psi(x, y) &\approx \frac{1}{2}(\kappa + \gamma_+)\, x^2 + \frac{1}{2}(\kappa - \gamma_+)\, y^2 + \gamma_\times xy \\ &= \frac{\kappa}{2}(x^2 + y^2) + \frac{\gamma_+}{2}(x^2 - y^2) + \gamma_\times xy\,. \end{aligned} \tag{4.33}$$

To see that the κ-term is isotropic while the γ-terms are not, we introduce polar coordinates (R, ϕ):

$$\begin{aligned}\psi(R, \phi) &= \frac{R^2}{2}[\kappa + \gamma_+(\cos^2\phi - \sin^2\phi) + 2\gamma_\times \sin\phi\cos\phi] \\ &= \frac{R^2}{2}[\kappa + \gamma_+\cos(2\phi) + \gamma_\times \sin(2\phi)]\,. \end{aligned} \tag{4.34}$$

Setting

$$\gamma_+ \equiv \gamma\cos(2\phi_\gamma) \tag{4.35a}$$

$$\gamma_\times \equiv \gamma\sin(2\phi_\gamma)\,, \tag{4.35b}$$

with $\gamma > 0$ and $0 \le \phi_\gamma < \pi$, we rewrite Eq. (4.34) as

$$\begin{aligned}\psi(R, \phi) &= \frac{R^2}{2}[\kappa + \gamma\cos(2\phi_\gamma)\cos(2\phi) + \gamma\sin(2\phi_\gamma)\sin(2\phi)] \\ &= \frac{R^2}{2}\{\kappa + \gamma\cos[2(\phi - \phi_\gamma)]\}\,. \end{aligned} \tag{4.36}$$

Notice that the first term is isotropic, i.e., independent of ϕ, whereas the second term depends on both R and ϕ. We also find that the second term is invariant under the rotation $\phi \to \phi + \pi$. If the shear were a vector, it would be invariant under

$\phi \to \phi + 2\pi$ but not under $\phi \to \phi + \pi$. This implies that γ and ϕ_γ (or equivalently γ_+ and $\gamma_\times$) define a second rank *tensor*—a vector being a first rank tensor.

Reverting to Cartesian coordinates, we compute the deflection angle as

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \begin{bmatrix} \kappa + \gamma_+ & \gamma_\times \\ \gamma_\times & \kappa - \gamma_+ \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \equiv \Gamma \boldsymbol{\theta} \,. \tag{4.37}$$

Letting I be the two-dimensional identity matrix, we can write the lens equation (4.26) in the form

$$\boldsymbol{\beta} = (\mathsf{I} - \Gamma)\, \boldsymbol{\theta} \,. \tag{4.38}$$

Thus, there is a single image at

$$\boldsymbol{\theta} = (\mathsf{I} - \Gamma)^{-1} \boldsymbol{\beta} \equiv \mathsf{A}_0 \boldsymbol{\beta} \,, \tag{4.39}$$

where the subscript "0" reminds us that Γ involves derivatives of the lens potential (4.33) evaluated at the origin.

4.2 Amplification Tensor

While the lens equation (4.26) involves first derivatives of the Fermat potential (4.25), the magnification arises from its second derivatives. According to Eq. (2.23), when the lens is axisymmetric the inverse magnification of a source with angular position β and with an image at θ is

$$\mu^{-1}(\theta) = \frac{\beta}{\theta} \frac{\mathrm{d}\beta}{\mathrm{d}\theta} \,, \tag{4.40}$$

namely, the ratio of differential area in the source plane to that in the image plane. For general lenses, we need the **amplification tensor** A, whose inverse is the Jacobian of the lens mapping:

$$\mathsf{A}^{-1}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} \,, \tag{4.41}$$

where $\boldsymbol{\beta} \equiv [\, u \;\; v \,]^T$ and $\boldsymbol{\theta} \equiv [\, x \;\; y \,]^T$. We use the convention that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are column vectors. The magnification of an image is given by the relation

$$\mu(\boldsymbol{\theta}) = \det \mathsf{A}(\boldsymbol{\theta}) \,. \tag{4.42}$$

4.2.1 Magnification for Constant Convergence and Shear

To begin exploring the properties of the amplification tensor, let us return to the scenario from Sect. 4.1.3 in which the lens potential is characterized by a constant convergence and shear. Equation (4.39) implies that the amplification tensor is $\mathsf{A}_0 = (\mathsf{I} - \Gamma)^{-1}$. The magnification of a lensed image can therefore be written in terms of the convergence κ and shear γ_+ and $\gamma_\times$ as

$$\mu_0 = \det \mathsf{A}_0 = \left[(1-\kappa)^2 - \gamma^2\right]^{-1}, \tag{4.43}$$

where $\gamma \equiv \sqrt{\gamma_+^2 + \gamma_\times^2}$.

Consider a small circular source with radius ϵ. In terms of the azimuthal angle ϕ, we can parametrize the boundary of the source as

$$\boldsymbol{\beta} = \epsilon \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix}. \tag{4.44}$$

In the absence of shear ($\gamma_+ = \gamma_\times = 0$), $\mathsf{A}_0 = (1-\kappa)^{-1}\mathsf{I}$. This implies that

$$\boldsymbol{\theta} = \frac{\epsilon}{1-\kappa} \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix} = (1-\kappa)^{-1}\boldsymbol{\beta}. \tag{4.45}$$

Thus, the (angular) radius of the image will be $\epsilon' = |1-\kappa|^{-1}\epsilon$. The absolute value ensures that $\epsilon' > 0$ even when $\kappa > 1$. We see in the upper left panel of Fig. 4.1 that the image will be larger than the source ($\epsilon' > \epsilon$) as long as $0 < \kappa < 2$. If $\kappa > 2$, however, the image will be smaller than the source ($\epsilon' < \epsilon$). The magnification will be given by

$$\mu = (\epsilon'/\epsilon)^2 = (1-\kappa)^{-2} > 0 \tag{4.46}$$

(see the upper right panel of Fig. 4.1).

When $\mu > 1$, we say the source is magnified; when $\mu < 1$ we say it is demagnified. As expected, $\mu = 1$ when $\kappa = 0$. When the lens is critical ($\kappa = 1$), the source will be infinitely magnified. This situation does not arise in practice because real sources are not infinitesimal.

In the presence of shear, a circular source appears as an elliptical image. For convenience let us set $\kappa = 0$. Suppose for now that $\gamma_\times = 0$. We can then write

$$\mathsf{A}_0 = \frac{1}{1-\gamma_+^2} \begin{bmatrix} 1+\gamma_+ & 0 \\ 0 & 1-\gamma_+ \end{bmatrix}, \tag{4.47}$$

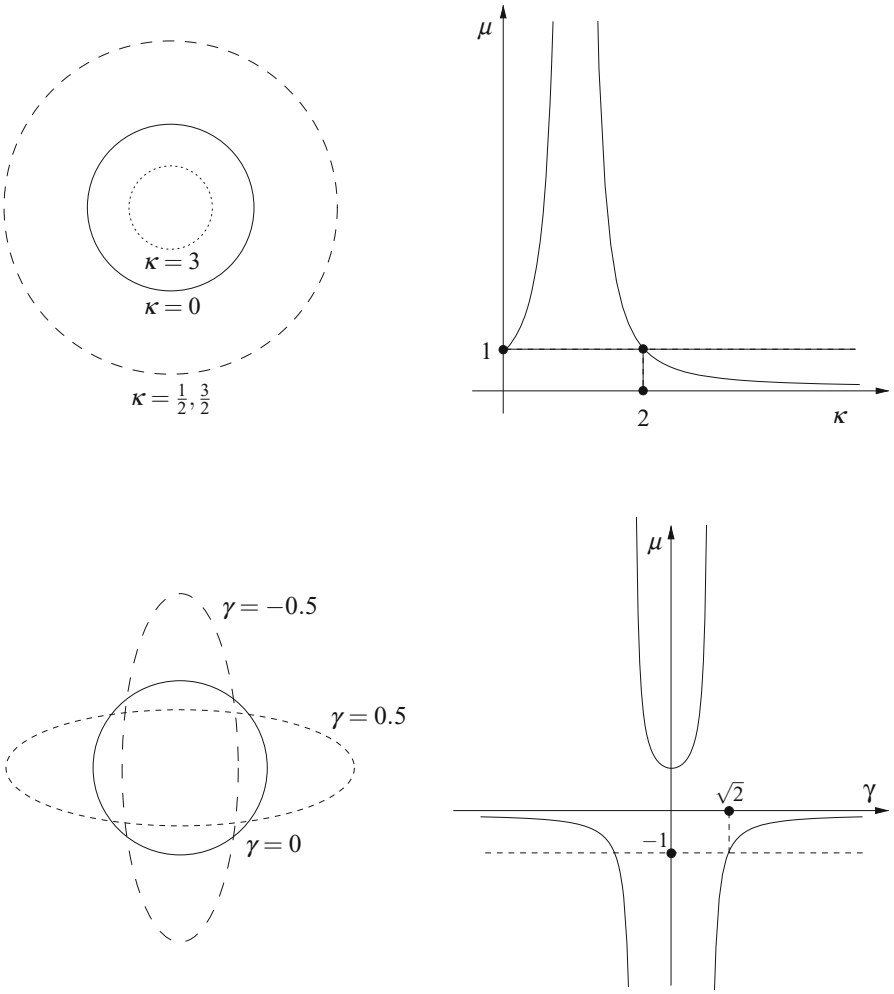


Fig. 4.1 Image (left-hand panels) and magnification (right-hand panels) of a circular source for different values of convergence ($\kappa \neq 0$; $\gamma_+ = \gamma_\times = 0$) and shear ($\gamma_+ \equiv \gamma \neq 0$; $\kappa = \gamma_\times = 0$). Note that if $0 \leq \kappa < 1$, $1 - \kappa$ produces the same magnification, μ, as κ. When $\kappa = 0$, either $\mu \geq 1$ or $\mu < 0$; each value of $\mu \neq 1$ corresponds to two distinct values of γ. If neither the γ_+ component nor the $\gamma_\times$ component of shear vanishes, the horizontal and vertical axes are aligned with the eigenvectors of A, and $\gamma = \sqrt{\gamma_+^2 + \gamma_\times^2}$

so that

$$\boldsymbol{\theta} = \frac{\epsilon}{1 - \gamma_+^2} \begin{bmatrix} (1 + \gamma_+) \cos\phi \\ (1 - \gamma_+) \sin\phi \end{bmatrix} . \tag{4.48}$$

Defining

$$a \equiv \frac{\epsilon}{1-\gamma_+} \tag{4.49a}$$

$$b \equiv \frac{\epsilon}{1+\gamma_+}\,, \tag{4.49b}$$

and using the identity $\cos^2\phi + \sin^2\phi = 1$, we see that $x^2/a^2 + y^2/b^2 = 1$. This is just the equation of an ellipse with semimajor axis a and semiminor axis b, assuming $\gamma_+ > 0$. The roles of a and b are reversed if $\gamma_+ < 0$ (see the lower left panel of Fig. 4.1). Notice that $a < 0$ ($b < 0$) when $\gamma_+ > 1$ ($\gamma_+ < -1$). We offer a physical interpretation in Sect. 4.2.3.

Even without appealing to Eq. (4.42), we can compute the magnification as the ratio of image area to source area:

$$\mu = \frac{\pi ab}{\pi\epsilon^2} = \frac{1}{1-\gamma_+^2}\,. \tag{4.50}$$

Unlike the case with convergence and no shear, it is possible in the present situation for μ to be negative (see the lower right panel of Fig. 4.1): this happens when $|\gamma_+| > 1$. This curious fact has the same explanation (see Sect. 4.2.3) as the negative semimajor axis we encountered above. We also note that demagnification ($|\mu| < 1$) occurs when $|\gamma_+| > \sqrt{2}$. The case with $\gamma_\times \neq 0$ is the subject of Problem 4.4.

4.2.2 General Case

In terms of the lens potential, the inverse amplification tensor can be written as

$$\mathsf{A}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} 1-\psi_{xx}(\boldsymbol{\theta}) & -\psi_{xy}(\boldsymbol{\theta}) \\ -\psi_{xy}(\boldsymbol{\theta}) & 1-\psi_{yy}(\boldsymbol{\theta}) \end{bmatrix}, \tag{4.51}$$

where the subscripts denote partial derivatives as in Sect. 4.1.3. While we have been thinking of the components of shear as constants, it is convenient to generalize these quantities to be functions of $\boldsymbol{\theta}$. In particular, we define

$$\gamma_+(\boldsymbol{\theta}) \equiv \frac{1}{2}[\psi_{xx}(\boldsymbol{\theta}) - \psi_{yy}(\boldsymbol{\theta})] \tag{4.52a}$$

$$\gamma_\times(\boldsymbol{\theta}) \equiv \psi_{xy}(\boldsymbol{\theta})\,. \tag{4.52b}$$

The amplification tensor at $\boldsymbol{\theta}$ can then be expressed in terms of the convergence and shear:

$$\mathsf{A} = \frac{1}{(1-\kappa)^2 - \gamma_+^2 - \gamma_\times^2} \begin{bmatrix} 1-\kappa+\gamma_+ & \gamma_\times \\ \gamma_\times & 1-\kappa-\gamma_+ \end{bmatrix} . \tag{4.53}$$

Note that the lens equation may have more than one solution and the magnification must be evaluated separately at each of the image positions.

4.2.3 Eigenvalues and Image Parity

We can characterize each of the several images formed by a strong lens by the convergence and shear in its neighborhood. As we have seen, both the size and the shape of a source can be distorted by lensing. An image can also be inverted along either or both of two principal axes. To see this, we need go no further than the case of a lens with nonzero shear and vanishing convergence. Specifically, if $\gamma_+ \neq 0$ and $\gamma_\times = 0$, the principal axes are just the major and minor axes of the elliptical image. The analogous result for $\gamma_+ = 0$ and $\gamma_\times \neq 0$ is the subject of Problem 4.4. We now turn to the general situation, where the convergence and both components of shear may be nonzero.

While it is straightforward enough to compute the magnification as $\mu = \det \mathsf{A}$, a geometric interpretation leads to a more intuitive understanding. Recall from linear algebra that the determinant of a matrix can be written as the product of eigenvalues. For a real, symmetric matrix, the eigenvalues will be real, and if they are distinct, then the associated eigenvectors will be orthogonal; these vectors define the principal axes we are seeking. Each lensed image will have a different set of principal axes, determined by the local convergence and shear. In the axisymmetric case, the eigenvalues of A give the radial and tangential magnifications. For a general lens, the eigenvalues give the magnifications along the principal axes.

It turns out that the simplest approach is to compute the eigenvalues, λ, of A^{-1}; the eigenvalues of A will then be λ^{-1}. The characteristic equation of A^{-1} is

$$0 = \det(\mathsf{A}^{-1} - \lambda \mathsf{I}) = \lambda^2 - 2(1-\kappa)\lambda + (1-\kappa)^2 - \gamma^2 = 0 , \tag{4.54}$$

where $\gamma \equiv \sqrt{\gamma_+^2 + \gamma_\times^2}$. The solutions are $\lambda_\pm = (1-\kappa) \pm \gamma$, so that the eigenvalues of A are simply $[(1-\kappa) \pm \gamma]^{-1}$.

Recall that an eigenvector of A corresponding to the eigenvalue λ^{-1} is an eigenvector of A^{-1} corresponding to the eigenvalue λ. Thus, it suffices to solve the equation

$$\mathsf{A}^{-1}\mathbf{x} = \lambda \mathbf{x} . \tag{4.55}$$

Using the notation defined in Eqs. (4.35), the normalized eigenvectors of A are

$$\mathbf{x}_+ = \begin{bmatrix} -\sin\phi_\gamma \\ \cos\phi_\gamma \end{bmatrix} \tag{4.56a}$$

$$\mathbf{x}_- = \begin{bmatrix} \cos\phi_\gamma \\ \sin\phi_\gamma \end{bmatrix}, \tag{4.56b}$$

where $\mathbf{x}_\pm$ corresponds to $\lambda_\pm^{-1}$.

The magnification of an image is just the product

$$\mu = (\lambda_+\lambda_-)^{-1} = [(1-\kappa)^2 - \gamma^2]^{-1}, \tag{4.57}$$

which depends on the components of shear through the parameter γ. Geometrically, the source will be magnified by an amount λ_+^{-1} (λ_-^{-1}) parallel to $\mathbf{x}_+$ ($\mathbf{x}_-$). The upshot is that an infinitesimal circular source will appear as an infinitesimal elliptical image whose axis ratio is given by λ_-/λ_+. This will be smaller (larger) than unity for $\kappa < 1$ ($\kappa > 1$), since $\gamma > 0$.

The eigenvectors do not depend on the convergence, because that is an isotropic effect. Moreover, they are independent of the magnitude of shear and involve only the angle $\phi_\gamma = \tan^{-1}(\gamma_\times/\gamma_+)/2$.

A negative eigenvalue means that the image will be inverted along the direction of the corresponding eigenvector. The parity of an image will be positive if the orientation of the source is maintained by lensing and negative if an inversion occurs. This effect will be unobservable for a purely elliptical source but can be measured for a source whose structure is more general.

4.3 Time Delay and Parity

Along with the positions and fluxes of lensed images, the time delay between any pair of images is sometimes observable. Such a measurement requires that the unlensed light source vary in brightness over time. Since each lensed image corresponds to a distinct light path, the travel time of light from the source to the observer differs from one image to another. To compare observed time delays to those predicted by a particular model, we must appeal to the Fermat potential. We have seen that its stationary points give the locations of the lensed images, but so far we have not attempted to determine whether a given image occurs at a minimum, maximum, or saddle point. In this section we establish a connection between the type of stationary point and the parity of the corresponding image.

Suppose that $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ solves the lens equation, so that $\boldsymbol{\nabla}_{\boldsymbol{\theta}}\,\tau|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}} = \mathbf{0}$. In the one-dimensional case, this condition takes the form $\tau'(\vartheta) = 0$, and the sign of $\tau''(\vartheta)$ determines the type of extremum. In two dimensions one deduces the nature of a stationary point from the Hessian matrix

$$\mathsf{H} = \begin{bmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{yx} & \tau_{yy} \end{bmatrix} \tag{4.58}$$

of second derivatives, evaluated at the point (see any calculus textbook). If the determinant $\det \mathsf{H} < 0$, then the stationary point is a saddle point. This is because $\det \mathsf{H} < 0$ requires the eigenvalues of the Hessian to have opposite signs. If $\det \mathsf{H} > 0$, either both eigenvalues are positive, or they are both negative. The former case corresponds to a minimum of τ and, the latter, to a maximum.

According to Eq. (4.25), $\tau_{ij} = \delta_{ij} - \psi_{ij}$, where ψ is the lens potential. Together with Eq. (4.51), this implies $\mathsf{H} = \mathsf{A}^{-1}$. Thus, a saddle point of τ has negative parity, i.e., $\mu = \det \mathsf{A} < 0$, while a maximum or minimum has positive parity ($\mu > 0$). Note that the partial parities of a maximum of τ are negative, i.e., the image is inverted along both of its principal axes.

4.4 Burke's Theorem

In Sect. 2.5.1 we showed that a differentiable lens mapping produces at least one image. When the lens is supercritical ($\kappa > 1$), it produces more than one image of any source that is sufficiently close to the lens. Our argument closely parallels that given by Dyer and Roeder (1980), who showed that a spherical lens with finite mass and continuous deflection function has an odd number of images. This result is of limited utility, since it excludes the point mass, as well as the singular and nonsingular isothermal spheres; the deflection of the point mass is infinite at the origin, and isothermal spheres have infinite mass. Burke (1981) showed that the conclusion of Dyer and Roeder remains true even for nonspherical lenses with infinite mass, provided that the deflection is bounded and continuous. In particular, Burke's theorem applies to the nonsigular isothermal sphere (and its elliptical counterpart), though not to the singular isothermal sphere or to the point mass. A rigorous proof of Burke's theorem is beyond the scope of this book, so we offer a conceptual argument.

We define two vector fields on the lens plane: the reduced deflection angle $\boldsymbol{\alpha}(\boldsymbol{\theta})$ due to the lens and the reduced deflection angle $\tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \equiv \boldsymbol{\theta} - \boldsymbol{\beta}$ that the light ray would need to undergo in order to reach the observer. Images form at points where the difference $\tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta}) - \boldsymbol{\alpha}(\boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}} \tau$ vanishes. Since $\boldsymbol{\alpha}(\boldsymbol{\theta})$ is bounded by assumption and $\boldsymbol{\beta}$ is constant, $\nabla \tau \approx \boldsymbol{\theta}$ for large $|\boldsymbol{\theta}|$.

Given a zero $\mathbf{x}_0$ of a vector field $\mathbf{A}$ defined in the plane, we define the *index* of $\mathbf{x}_0$ to be the net rotation (in units of 2π) of the vectors on a small simple closed curve $\mathscr{C}$ around $\mathbf{x}_0$ as it is traversed once counterclockwise. The index corresponding to a minimum or maximum point of τ is $+1$, while that corresponding to a saddle point is -1. Because $\nabla \tau(\boldsymbol{\theta}) \approx \boldsymbol{\theta}$ for large $|\boldsymbol{\theta}|$, the *Poincaré-Hopf index theorem* (see, e.g., Guillemin and Pollack 1974, p. 134) implies that the sum of the indices of all the zeros of $\nabla \tau$ is $+1$. Thus, $\nabla \tau$ has an odd number of zeros, which implies that the lens produces an odd number of images. We may also conclude that $n_{\min} + n_{\max} = n_{\text{sad}} + 1$, where n is the number of images of the specified type. Proving the Poincaré-Hopf index theorem is far beyond the scope of this book, but the *residue theorem*, presented in Appendix B.3, offers a pleasing analogy from complex analysis.

4.5 Critical Curves and Caustics

We showed in Sect. 2.2.3 that an image of a source with infinitesimal area has divergent magnification if the source lies on a caustic. Point caustics and point critical curves, which arise for axisymmetric lenses (see Sect. 2.3), are replaced by closed curves when the lens has some degree of asymmetry. We will find the caustics and corresponding critical curves for representative lens models in Chap. 6. Our goal in this section is to show that the number of images changes by two when a source crosses a caustic.

The critical points form a finite set of closed curves (Schneider et al. 1992, §6.2), which generically consist of smooth sections called **folds** that meet at zero or more points known as **cusps** (Whitney 1955). Fold points and cusps are collectively referred to as **stable critical points**.[2] To study their lensing properties, we begin by expanding the lens potential ψ in a neighborhood of a critical point in a Taylor series. We have seen that the second-order expansion (4.33) leads to a single image with finite magnification, except in the special case that the total shear γ and the convergence κ satisfy $\gamma = |1-\kappa|$. Since the magnification diverges at a critical point, higher-order terms must be included in the Taylor expansion. For a stable critical point at the origin,

$$\psi(x,y) = \frac{1}{2}(1-K)\,x^2 + \frac{1}{2}\,y^2 + e\,x^3 + f\,x^2y + g\,xy^2 + h\,y^3 \\ +k\,x^4 + m\,x^3y + n\,x^2y^2 + p\,xy^3 + r\,y^4 \tag{4.59}$$

to fourth order (Petters et al. 2001, p. 346), where $K \neq 0$. Note that this expansion gives information only about images near the critical point, and not about any that lie far away from it.

4.5.1 Folds

At a fold critical point the additional condition $h \neq 0$ holds[3] (Petters et al. 2001, p. 347). In this case we may set the other third-order coefficients e, f, and g and all the fourth-order coefficients k, m, n, p, and r equal to zero without affecting the results qualitatively (see Keeton et al. (2005) for the general treatment).

The corresponding lens equation is

$$u = Kx\,, \quad v = -3hy^2\,, \tag{4.60}$$

[2] Other types of critical points are more difficult to analyze and are encountered rarely in practice. We refer the interested reader to the book by Petters et al. (2001) for details.

[3] Equivalently, one can define a fold point by the condition $e \neq 0$.

or, equivalently,

$$x = \frac{u}{K}, \quad y = \pm\sqrt{\frac{v}{-3h}}. \tag{4.61}$$

The magnification is

$$\mu = -\frac{1}{6hKy}. \tag{4.62}$$

Thus, the critical curve is the line $y = 0$, and the corresponding caustic line is $v = 0$. If $h > 0$, the lens equation has two real solutions for $v < 0$ and no real solutions for $v > 0$. Sources in these regimes are *inside* or *outside* the caustic, respectively. The magnifications of the two images for $v < 0$ are

$$\mu_\pm = \pm\frac{1}{2K\sqrt{-3hv}}. \tag{4.63}$$

The key results are as follows:

1. A source outside a fold caustic does not produce any images near the fold critical curve.
2. A source inside a fold caustic produces two images that are equidistant from the fold critical curve.
3. For a source inside the caustic, the magnification diverges as the inverse square root of the distance from the caustic.
4. The magnification is discontinuous across the caustic.
5. The magnifications for a source inside the caustic satisfy $\mu_+ + \mu_- = 0$, implying that the images are equally bright and have opposite parity.

4.5.2 Cusps

For a cusp critical point at the origin, the expansion (4.59) again applies but with the additional restrictions[4] $h = 0$, $g \neq 0$, and $r \neq 0$ (Petters et al. 2001, p. 347). In this case we may set the remaining third-order and fourth-order coefficients equal to zero without affecting the results qualitatively (see Keeton et al. (2003) for the general treatment). The corresponding lens equation is

$$u = Kx - gy^2 \tag{4.64a}$$

$$v = -2gxy - 4ry^3, \tag{4.64b}$$

[4] Or equivalently, $e = 0$, $f \neq 0$, and $k \neq 0$.

and the magnification is

$$\mu = -\left[2Kgx + 4\left(3Kr + g^2\right)y^2\right]^{-1} . \tag{4.65}$$

The critical curve is the parabola

$$x = -\frac{2\left(3Kr + g^2\right)}{Kg}y^2 . \tag{4.66}$$

If $\left(3Kr + g^2\right)/(Kg) > 0$, the parabola opens to the left and otherwise to the right.

The caustic can be written parametrically as

$$u = -\frac{3(2Kr + g^2)}{g}\, y^2 \tag{4.67a}$$

$$v = \frac{4(2Kr + g^2)}{K}\, y^3 . \tag{4.67b}$$

It opens to the right if $u > 0$ in Eq. (4.67a), i.e., if $(2Kr + g^2)/g < 0$, and otherwise to the left. An example is shown in Fig. 4.2.

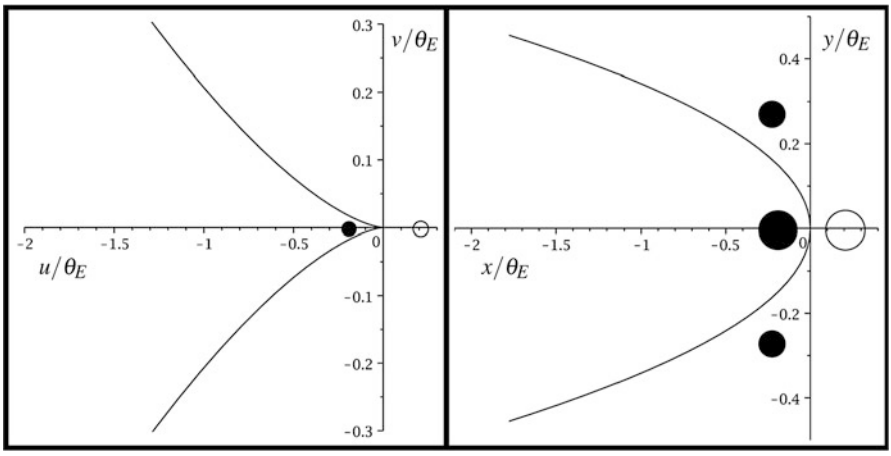


Fig. 4.2 Caustic (left panel) and critical curve (right panel) near a cusp produced by a singular isothermal ellipsoid (see Sect. 6.1.2) with axis ratio $q = 0.8$. The angle θ_E is the Einstein radius for the singular isothermal sphere obtained by setting $q = 1$. A source outside the caustic produces one image, while a source inside the caustic produces three images. The ratio of image area to source area gives the (unsigned) magnification. Note that a circular source appears as elliptical images, but the distortion is not depicted here

To find the image positions, we first eliminate x from Eqs. (4.64), which leads to

$$y^3 + \frac{gu}{2Kr + g^2}\, y + \frac{Kv}{2(2Kr + g^2)} = 0\,. \tag{4.68}$$

Since this equation is cubic in y, there are either one or three images.

First, let us consider a source on the u-axis. One solution of the cubic equation is $y = 0$, corresponding to an image at $x = u/K$. Its magnification is $\mu = -1/(2gu)$, which diverges for a source approaching the cusp from either side of the caustic. There are two other solutions of the lens equation:

$$x = \frac{2r}{2Kr + g^2}u \tag{4.69a}$$

$$y = \pm\sqrt{\frac{-gu}{2Kr + g^2}}\,, \tag{4.69b}$$

which are real when the argument of the square root is positive. Both of these images have magnification $\mu = 1/(4gu)$, which diverges as the source approaches the cusp from inside the caustic. From the discussion below Eqs. (4.67), we conclude that there are three images of a source inside the caustic and one image of a source outside.

For a general source position (u, v), we can obtain some useful relations without explicitly computing the image positions. Let us denote the three solutions of the cubic equation (4.68) by y_1, y_2, y_3. We can then write the cubic polynomial as

$$\begin{aligned} 0 &= (y - y_1)(y - y_2)(y - y_3) \\ &= y^3 - (y_1 + y_2 + y_3)y^2 + (y_1y_2 + y_1y_3 + y_2y_3)y - (y_1y_2y_3)\,. \end{aligned} \tag{4.70}$$

Matching coefficients with Eq. (4.68) yields the relations

$$y_1 + y_2 + y_3 = 0 \tag{4.71a}$$

$$y_1y_2 + y_1y_3 + y_2y_3 = \frac{gu}{2Kr + g^2} \tag{4.71b}$$

$$y_1y_2y_3 = -\frac{Kv}{2(2Kr + g^2)}\,. \tag{4.71c}$$

For the image at y_i, we can use Eq. (4.64a) to find x_i and, inserting these coordinates in Eq. (4.65), obtain the magnification. Combining the results with Eqs. (4.71) leads to the magnification sum rule

$$\mu_1 + \mu_2 + \mu_3 = 0\,. \tag{4.72}$$

The key results are as follows:

1. A source outside a cusp produces one image.
2. A source inside a cusp produces three images.
3. For sources both inside and outside a cusp, the magnification diverges as the inverse of the distance from the caustic.
4. In the case of three images, the magnifications satisfy $\mu_1 + \mu_2 + \mu_3 = 0$.

4.6 Surface Brightness and Extended Sources

Our discussion so far has focused on point-like or infinitesimal sources, for mathematical convenience. From the definition of magnification, it would seem that our task is to determine the ratio of image area (or more precisely, solid angle) to source area in an observed lens. This approach does not work in practice, however, since the area of the unlensed source is unknown. For a strong lens, where there is more than one image, we can compute the ratio of image areas, which provides $n - 1$ observables for a system of n images.

There is yet another difficulty we must overcome: it is typically not possible to measure the area of an observed image with sufficient precision. Fortunately, it turns out that the magnification is also the ratio of image flux to source flux. We can obviate knowledge of the source flux by computing flux ratios among the various images. We can replace area by flux due to the principle of conservation of surface brightness, which we will come to in a moment. We can then compute the magnification of an extended source as a flux ratio.

4.6.1 Conservation of Surface Brightness

Consider a source subtending an infinitesimal solid angle $\mathrm{d}\Omega$ that gives rise to an image subtending an angle $\mathrm{d}\Omega'$. The flux per unit frequency of the source is given in terms of the **surface brightness** by

$$\mathrm{d}F_\nu = I_\nu \mathrm{d}\Omega \tag{4.73}$$

and that of the image by

$$\mathrm{d}F'_{\nu'} = I'_{\nu'} \mathrm{d}\Omega' \,. \tag{4.74}$$

The quantity I_ν is the surface brightness (flux per unit frequency per unit solid angle) of the source and $I'_{\nu'}$ is that of the image. For the time being, we allow for the possibility that the lensed frequency ν' may differ from the unlensed frequency ν.

In order to show that the ratio of fluxes is identical to that of solid angles, we must demonstrate that $I'_{\nu'} = I_\nu$ (see Petters et al. 2001, §3.2.5). To do this, we

invoke Liouville's theorem, familiar from classical mechanics, which says that the number of particles per unit volume of phase space is conserved, provided that the particles do not interact with one another. During a time interval Δt, photons from a distant source will move a distance $\Delta z = c\Delta t$ toward an observer on earth. If these photons traverse a cross sectional area ΔA, the volume element in position space will be $\Delta V_{\rm pos} = c\,\Delta A\,\Delta t$. Meanwhile, photons with momenta between p and $p + \Delta p$ emanating from a solid angle $\Delta\Omega$ about the z-axis will occupy a volume in momentum space $\Delta V_{\rm mom} = p^2\,\Delta p\,\Delta\Omega$. For a system of N noninteracting photons, the phase space density is given by

$$\begin{aligned} n &= \frac{N}{\Delta V_{\rm pos}\Delta V_{\rm mom}} \\ &= \frac{N}{c\,\Delta A\,\Delta t\,p^2\,\Delta p\,\Delta\Omega} \\ &= \frac{c^2 N}{h^3\nu^2\Delta\nu\,\Delta t\,\Delta A\,\Delta\Omega}\,. \end{aligned} \tag{4.75}$$

The total energy of the system is $h\nu N$, so that the surface brightness takes the form

$$I_\nu = \frac{h\nu N}{\Delta\nu\,\Delta t\,\Delta A\,\Delta\Omega}\,. \tag{4.76}$$

Comparing this expression to the phase space density yields the relation

$$\frac{I_\nu}{\nu^3} = \frac{nh^4}{c^2} = \text{const.} \tag{4.77}$$

(see Misner et al. 1973, Fig. 22.2). The lensed and unlensed surface brightnesses satisfy

$$\frac{I'_{\nu'}}{\nu'^3} = \frac{I_\nu}{\nu^3}\,. \tag{4.78}$$

If the gravitational potential of the lens remains static while the N photons pass by, any blueshift of the photons as they approach the lens will be offset by a redshift as they move away. This means that $\nu' = \nu$ and hence that $I'_{\nu'} = I'_\nu = I_\nu$; this is the conservation of surface brightness.

4.6.2 Magnification of an Extended Source

Since the surface brightness of an image is the same as that of the source, the monochromatic flux of an image (4.74) becomes

$$\mathrm{d}F'_\nu = I_\nu \mathrm{d}\Omega' = I_\nu\,|\mu|\,\mathrm{d}\Omega = |\mu|\mathrm{d}F_\nu\,, \tag{4.79}$$

where we take the absolute value of the magnification to ensure that the flux is positive. Now consider an extended source whose surface brightness is a function of position: $I_\nu = I_\nu(\mathbf{x})$. The surface brightness at a position $\boldsymbol{\beta}'$ for a source with centroid $\boldsymbol{\beta}$ is then $I_\nu(\boldsymbol{\beta}' - \boldsymbol{\beta})$. The total flux of the image is given by

$$F'_\nu(\boldsymbol{\beta}) = \int_{\mathbb{R}^2} I_\nu(\boldsymbol{\beta}' - \boldsymbol{\beta})|\mu(\boldsymbol{\beta}')|\mathrm{d}^2\beta' \,. \tag{4.80}$$

If a specified source position $\boldsymbol{\beta}$ gives rise to n distinct images, we replace $\left|\mu(\boldsymbol{\beta}')\right|$ by the total magnification,

$$\mu_{\rm tot}(\boldsymbol{\beta}') \equiv \sum_{i=1}^{n} \left|\mu_i(\boldsymbol{\beta}')\right| \,. \tag{4.81}$$

Since a source of flux F_ν is magnified by the factor $\mu_{\rm ext} = F'_\nu/F_\nu$,

$$\mu_{\rm ext}(\boldsymbol{\beta}) = \frac{\displaystyle\int_{\mathbb{R}^2} I_\nu(\boldsymbol{\beta}' - \boldsymbol{\beta})\mu_{\rm tot}(\boldsymbol{\beta}')\mathrm{d}^2\beta'}{\displaystyle\int_{\mathbb{R}^2} I_\nu(\boldsymbol{\beta}')\mathrm{d}^2\beta'} \,. \tag{4.82}$$

Thus, the magnification of an extended source is the convolution of the surface brightness and the magnification due to a point source, normalized by the flux of the unlensed source.

We saw in Sect. 4.5 that the magnification diverges as a power law as a point source approaches a caustic. However, the divergences of $\mu_{\rm tot}$ are weak enough so that the integral in the numerator of Eq. (4.82) converges. Thus, the magnification $\mu_{\rm ext}(\boldsymbol{\beta})$ of an extended source remains finite for all positions $\boldsymbol{\beta}$ of the source centroid.

4.7 Degeneracies in the Lens Equation

There are certain transformations that leave the form of the lens equation unchanged. It is important to understand such degeneracies because they can lead to ambiguity in the lens parameters that give rise to a given set of observables. Here we identify several types of transformations that play a role in applications discussed later; more detailed presentations can be found in the literature (e.g., Gorenstein et al. 1988; Saha 2000; Schneider and Sluse 2014).

4.7.1 Similarity Transformations

One simple transformation is to rescale by a constant multiplicative factor. If a is a constant, the transformation

$$\frac{D_l D_s}{D_{ls}} \rightarrow a \frac{D_l D_s}{D_{ls}} \tag{4.83}$$

leaves the source position, image positions, and magnifications unchanged but rescales the time delays by the factor a.

Also consider the transformation

$$\boldsymbol{\beta} \rightarrow a^{1/2}\boldsymbol{\beta} \tag{4.84a}$$

$$\boldsymbol{\theta} \rightarrow a^{1/2}\boldsymbol{\theta} \tag{4.84b}$$

$$\kappa \rightarrow a\,\kappa\,. \tag{4.84c}$$

This transformation is valid only if the images are not resolved, such that $\boldsymbol{\theta}$ is not measured directly. The transformed quantities still solve the lens equation. The magnifications are unchanged, while the time delays are again rescaled by a.

Saha (2000) labels these the **distance degeneracy** and **angular degeneracy**, respectively. He also notes that they can be combined in different ways to yield what he terms "parallax" and "perspective" degeneracies. These degeneracies are important in microlensing (see Chap. 5).

4.7.2 Mass-Sheet Transformation

A more subtle transformation leads to the **mass-sheet degeneracy**. Consider the lens equation (4.26). In addition to the lens potential $\psi(\boldsymbol{\theta})$, we add an infinite sheet of constant mass density. The vector version of Eq. (2.73), assuming circular symmetry, is

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \bar{\kappa}(|\boldsymbol{\theta}|)\boldsymbol{\theta}\,, \tag{4.85}$$

where $\bar{\kappa}(|\boldsymbol{\theta}|)$ is the mean convergence within angular radius $|\boldsymbol{\theta}|$. If the convergence has some constant value κ_0, we have $\bar{\kappa} = \kappa_0$. The potential corresponding to Eq. (4.85) is then

$$\psi_0(\boldsymbol{\theta}) = \frac{1}{2}\kappa_0|\boldsymbol{\theta}|^2\,. \tag{4.86}$$

The lens equation for the combined potential $\psi + \psi_0$ is

$$\begin{aligned}\boldsymbol{\beta} &= \boldsymbol{\theta} - \boldsymbol{\nabla}\left(\psi(\boldsymbol{\theta}) + \frac{1}{2}\kappa_0|\boldsymbol{\theta}|^2\right) \\ &= (1-\kappa_0)\boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta})\,. \end{aligned} \tag{4.87}$$

If we now scale $\boldsymbol{\beta}$ and ψ by the factor $1-\kappa_0$, we recover Eq. (4.26). In other words, the form of the lens equation is preserved under the transformations:

$$\boldsymbol{\beta} \to (1-\kappa_0)\boldsymbol{\beta} \equiv \tilde{\boldsymbol{\beta}} \tag{4.88a}$$

$$\psi(\boldsymbol{\theta}) \to (1-\kappa_0)\psi(\boldsymbol{\theta}) + \frac{1}{2}\kappa_0|\boldsymbol{\theta}|^2 \equiv \tilde{\psi}(\boldsymbol{\theta})\,. \tag{4.88b}$$

Thus, the solutions $\boldsymbol{\theta}$ of $\tilde{\boldsymbol{\beta}} = \boldsymbol{\theta} - \boldsymbol{\nabla}\tilde{\psi}(\boldsymbol{\theta})$ are identical to those of Eq. (4.26).

It is not possible to distinguish between $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ or ψ and $\tilde{\psi}$ from the observed image positions alone. This degeneracy is unimportant for the source position, which is physically irrelevant, but poses a significant difficulty in determining the lens potential. Of course, no truly infinite mass sheet exists, but there may be an unknown quantity of matter along the line of sight to the lens that can effectively be modeled by a constant convergence field. The mass-sheet degeneracy seems not to apply to the magnification, since

$$\tilde{\mu}^{-1} \equiv \det\left(\frac{\partial\tilde{\boldsymbol{\beta}}}{\partial\boldsymbol{\theta}}\right) = (1-\kappa_0)^2\mu^{-1}\,, \tag{4.89}$$

where $\tilde{\mu}$ is the magnification in the presence of the mass sheet. If the actual source flux is known, then the observed magnifications enable us to distinguish between ψ and $\tilde{\psi}$. If not, we must work with magnification ratios. In that case, the factor $(1-\kappa_0)^2$ cancels, leaving the degeneracy intact. The time delay between a pair of images i and j seems to offer another way to break the degeneracy, because $\Delta\tilde{T}_{ij} = (1-\kappa_0)\Delta T_{ij}$, where the differential time delay is defined by

$$\Delta T_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \Delta T(\boldsymbol{\theta}_j) - \Delta T(\boldsymbol{\theta}_i)\,. \tag{4.90}$$

Unfortunately, there remains a degeneracy between κ_0 and the time scale T_* of Eq. (4.24). The mass sheet degeneracy is particularly important for galaxy-scale lensing (see Chap. 6).

4.7.3 Source Position Transformation

Schneider and Sluse (2014) point out that the mass sheet transformation is an example of a more general class that they call **source plane transformations**. Consider the transformation

$$\boldsymbol{\beta} \rightarrow \mathbf{f}(\boldsymbol{\beta}) \tag{4.91a}$$

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) \rightarrow \boldsymbol{\theta} - \mathbf{f}\left(\boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})\right) . \tag{4.91b}$$

The function $\mathbf{f}(\boldsymbol{\beta})$ must be one-to-one but is otherwise arbitrary. This transformation leaves the image positions unchanged, but it does modify the magnifications (by the Jacobian determinant of $\mathbf{f}$) and the time delays.

It is not guaranteed that the transformed deflection field can be written as the gradient of a potential or equivalently that it can be associated with a physically plausible surface mass distribution. A transformation that preserves axial symmetry yields a model that corresponds to a surface mass distribution. In the general case, the only transformation that maintains an exact connection to a surface mass distribution is $\mathbf{f}(\boldsymbol{\beta}) = a\boldsymbol{\beta}$ for constant a, which is equivalent to the mass sheet transformation. However, a broader range of transformations yield models that can be closely related to surface mass distributions. (See Schneider and Sluse (2014) for detailed discussion of these points.)

4.7.4 Connection to Electrodynamics

Equations (4.88) resemble gauge transformations in electrodynamics. Since the magnetic field satisfies $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$, we can write $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$, for some vector potential $\mathbf{A}$. From Faraday's law (in Gaussian units), we have

$$\begin{aligned} 0 &= \boldsymbol{\nabla} \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \\ &= \boldsymbol{\nabla} \times \left(\mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}\right) . \end{aligned} \tag{4.92}$$

Just as $\boldsymbol{\nabla} \times \boldsymbol{\alpha} = \mathbf{0}$ implies $\boldsymbol{\alpha} = \boldsymbol{\nabla}\psi$ in the context of lensing, Eq. (4.92) has the solution

$$\mathbf{E} = -\boldsymbol{\nabla}\Phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} ,$$

for some scalar potential Φ. The electric and magnetic fields are invariant under the transformation

$$\mathbf{A} \rightarrow \mathbf{A} + \boldsymbol{\nabla} f \equiv \tilde{\mathbf{A}} \tag{4.93a}$$

$$\Phi \rightarrow \Phi - \frac{1}{c}\frac{\partial f}{\partial t} \equiv \tilde{\Phi} \tag{4.93b}$$

for any function f. We say that the potentials $\mathbf{A}$ and Φ are degenerate with their counterparts $\tilde{\mathbf{A}}$ and $\tilde{\Phi}$.

Unlike lensing degeneracies, the gauge invariance of electrodynamics has no physically observable consequences, with the choice of gauge being a matter of convenience. Two popular gauges are those due to Coulomb and Lorenz, which require $\nabla \cdot \mathbf{A} = 0$ and $\nabla \cdot \mathbf{A} = -(1/c)\partial\Phi/\partial t$, respectively. The scalar potential satisfies Poisson's equation in the Coulomb gauge, and the scalar and vector potentials satisfy the inhomogeneous wave equation in the Lorenz gauge. Which (if either) of the gauges is preferred depends on the particular application. Since both the original and transformed fields in Eqs. (4.93) must be consistent with the choice of gauge, we conclude that the function f satisfies either Laplace's equation (Coulomb gauge) or the homogeneous wave equation (Lorenz gauge).

4.8 Multiplane Lensing

So far we have considered the deflection of a light ray by a single lensing object, but it is possible for the ray to encounter several lenses. If these lenses are approximately equidistant from the observer, they may be treated in a single lens plane according to Eq. (4.26). We now discuss the case of lenses that are at different distances along the line of sight. If the interlens distances are large compared to the size of each lens, the total deflection is the sum of individual deflections that can be described by the thin lens approximation. (See Schneider et al. (1992), Petters et al. (2001), and McCully et al. (2014) for more discussion.)

We begin with the case of two lenses at distances D_1 and $D_2 > D_1$ from the observer. We assume that the lens planes are parallel, i.e., the angular separation between the lenses on the sky is small. A light ray emitted by a source at distance D_s and angular position $\boldsymbol{\beta}$ relative to some line of sight undergoes a deflection of $\hat{\boldsymbol{\alpha}}_2(D_2\boldsymbol{\theta}_2)$, followed by a deflection of $\hat{\boldsymbol{\alpha}}_1(D_1\boldsymbol{\theta}_1)$. Generalizing the geometric derivation of the lens equation in Sect. 2.2.2 yields the relation

$$D_s\boldsymbol{\beta} = D_s\boldsymbol{\theta}_1 - D_{1s}\hat{\boldsymbol{\alpha}}_1(D_1\boldsymbol{\theta}_1) - D_{2s}\hat{\boldsymbol{\alpha}}_2(D_2\boldsymbol{\theta}_2)\,, \tag{4.94}$$

where $\boldsymbol{\theta}_1$ is the image position and D_{is} is the distance from lens i to the source. Dividing through by D_s, and recalling the definition of the reduced deflection angle (2.17), we have

$$\boldsymbol{\beta} = \boldsymbol{\theta}_1 - \boldsymbol{\alpha}_1(\boldsymbol{\theta}_1) - \boldsymbol{\alpha}_2(\boldsymbol{\theta}_2)\,. \tag{4.95}$$

The vector $\boldsymbol{\theta}_2$ would be the image position if the lens at D_1 were not present. In other words, we can think of $\boldsymbol{\theta}_2$ as the source position and write down the lens equation due to the single deflection $\hat{\boldsymbol{\alpha}}_1(D_1\boldsymbol{\theta}_1)$:

$$D_2\boldsymbol{\theta}_2 = D_2\boldsymbol{\theta}_1 - D_{12}\hat{\boldsymbol{\alpha}}_1(D_1\boldsymbol{\theta}_1)\,. \tag{4.96}$$

In terms of the reduced deflection angle, we obtain

$$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1 - \beta_{12}\boldsymbol{\alpha}_1(\boldsymbol{\theta}_1)\,, \tag{4.97}$$

where

$$\beta_{12} = \frac{D_{12}D_s}{D_2 D_{1s}}\,. \tag{4.98}$$

Thus, we have two equations for the two unknowns $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Note that if the planes are at the same distance, then $D_{12} = 0$ so $\beta_{12} = 0$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1$. Equation (4.95) then reduces to the single-plane lens equation for two lenses.

Now consider N lenses at distances D_1, D_2, ..., D_N from the observer, with $D_i \le D_{i+1}$. Applying the lens equation recursively, Eqs. (4.95), (4.97), and (4.98) can be replaced by

$$\boldsymbol{\beta} = \boldsymbol{\theta}_1 - \sum_{i=1}^{N} \boldsymbol{\alpha}_i(\boldsymbol{\theta}_i) \tag{4.99a}$$

$$\boldsymbol{\theta}_j = \boldsymbol{\theta}_1 - \sum_{i=1}^{j-1} \beta_{ij}\boldsymbol{\alpha}_i(\boldsymbol{\theta}_i) \qquad \text{for } j = 2, 3, \ldots, N \tag{4.99b}$$

$$\beta_{ij} = \frac{D_{ij}D_s}{D_j D_{is}} \qquad \text{for } i \le j\,. \tag{4.99c}$$

Note that Eq. (4.99a) becomes a special case of Eq. (4.99b) if we also let j take the value $N + 1$ and define $\boldsymbol{\theta}_{N+1} = \boldsymbol{\beta}$.

The time delay again has both geometric and potential terms, now summed over planes:

$$T(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N|\boldsymbol{\beta}) = \sum_{i=1}^{N} \tau_{i,i+1}\left[\frac{1}{2}\,|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i|^2 - \beta_{i,i+1}\psi_i(\boldsymbol{\theta}_i)\right], \tag{4.100}$$

where index $N + 1$ again represents the source plane, ψ_i is the lens potential for lens i, and

$$\tau_{ij} = \frac{1 + z_i}{c}\frac{D_i D_j}{D_{ij}} \qquad \text{for } i \le j\,. \tag{4.101}$$

(Note a useful identity: $\beta_{ij}\tau_{ij} = \tau_{is}$ for all i and j.) Petters et al. (2001) show that the multiplane lens equation and time delay are related through a Fermat's principle: the lens equation can be obtained by finding stationary points of T with respect to all of the angular positions $\{\boldsymbol{\theta}_i\}$.

Problems

4.1 The goal of this problem is to prove that $\nabla^2 g(\boldsymbol{\theta}) = 2\delta^{(2)}(\boldsymbol{\theta})$, where $g(\boldsymbol{\theta}) = \pi^{-1} \ln|\boldsymbol{\theta}|$ and ∇^2 is the two-dimensional Laplacian.

(a) Writing ∇^2 in polar coordinates, show that $\nabla^2 g(\boldsymbol{\theta}) = 0$ at all points away from the origin.
(b) The origin requires special attention. To begin, show that

$$\int_A \nabla^2 \ln|\boldsymbol{\theta}| \, \mathrm{d}^2\theta = \int_C \boldsymbol{\nabla} \ln|\boldsymbol{\theta}| \cdot \mathbf{n} \, \mathrm{d}s \,, \tag{4.102}$$

where $\mathrm{d}s$ is an element of arc length, the integration contour C bounds the area A, the direction of integration is counterclockwise, and $\mathbf{n}$ is an outward-pointing unit vector normal to the contour.
(c) Taking the area A to be a disk of radius b centered at the origin, evaluate the integral on the right-hand side of Eq. (4.102).
(d) Combine your results to prove that $\nabla^2 \ln|\boldsymbol{\theta}| = 2\pi\delta^{(2)}(\boldsymbol{\theta})$, as desired.

4.2 Using the lensing diagram shown in Fig. 2.2, compute the difference between light travel time for the actual light ray and a hypothetical ray that travels straight from the source to the observer. Assuming Euclidean geometry and the small angle approximation, show that the excess light travel time has the form given in Eq. (4.27) for the geometric time delay.

4.3 This problem explores how the general theory presented in this chapter applies to axisymmetric lenses. Express your answers in terms of the reduced deflection angle $\alpha(R)$ and its derivatives.

(a) Write Poisson's equation (4.22) for the lens potential in polar coordinates.
(b) Show that the convergence κ and the reduced deflection angle α are related by

$$\kappa = \frac{1}{2}\left(\frac{\alpha}{R} + \frac{\mathrm{d}\alpha}{\mathrm{d}R}\right) . \tag{4.103}$$

(c) Find general expressions for the eigenvalues and eigenvectors of the inverse amplification tensor. Assuming that the model has two critical curves, use the eigensystem to explain why we label one of the critical curves "tangential" and the other one "radial."

4.4 Following Sect. 4.2.1, suppose that $\gamma_\times$ is constant and nonzero, while γ_+ and κ both vanish.

(a) Once again parametrize the boundary of the source as a circle, and find the boundary of the image.
(b) Show that the image is an ellipse whose semimajor and semiminor axes are given by

$$a = \epsilon(1 - \gamma_\times)^{-1} \tag{4.104a}$$

$$b = \epsilon(1 + \gamma_\times)^{-1}\,. \tag{4.104b}$$

(c) What is the orientation of the ellipse?

4.5 The goal of this problem is to use the methods from Sect. 4.2 to analyze how images are distorted in the vicinity of fold and cusp caustics discussed in Sect. 4.5.

(a) Consider the two images produced by a fold caustic. What is the parity of each image? In what direction is each image stretched?
(b) Repeat the analysis from part (a) for the three images produced by a cusp caustic.

4.6 Consider an axisymmetric power-law lens model with lens potential $\psi(\theta) = A\theta^\eta$, where the power-law index η and normalization A are constants.

(a) Find the range of η such that the model is physically and astrophysically plausible, meaning that the density and deflection are nonnegative and the density does not increase with radius. Is the model able to produce multiple images for the full physical range of η?
(b) Rewrite the normalization in terms of the Einstein radius, θ_E (instead of A).
(c) For what range of η does the model have a radial critical curve? What is the radius of that critical curve?
(d) Plot magnification as a function of θ/θ_E for different values of η that show the range of behaviors for this model.
(e) Sketch the image configuration (indicating parity) for a multiply-imaged source when $\eta = 0.5$ and 1.5.

References

Burke, W. L. (1981). Multiple gravitational imaging by distributed masses. *Astrophysical Journal Letters, 244*, L1.

Dyer, C. C., & Roeder, R. C. (1980). Possible multiple imaging by spherical galaxies. *Astrophysical Journal Letters, 238*, L67.

Gorenstein, M. V., Shapiro, I. I., & Falco, E. E. (1988). Degeneracies in parameter estimates for models of gravitational lens systems. *Astrophysical Journal, 327*, 693.

Guillemin, V., & Pollack, A. (1974). *Differential topology*. Englewood Cliffs: Prentice-Hall.

Jackson, J. D. (1962). *Classical electrodynamics*. New York: Wiley.

Keeton, C. R., Gaudi, B. S., & Petters, A. O. (2003). Identifying lenses with small-scale structure. I. Cusp lenses. *Astrophysical Journal, 598*, 138.

Keeton, C. R., Gaudi, B. S., & Petters, A. O. (2005). Identifying lenses with small-scale structure. II. Fold lenses. *Astrophysical Journal, 635*, 35.

McCully, C., Keeton, C. R., Wong, K. C., & Zabludoff, A. I. (2014). A new hybrid framework to efficiently model lines of sight to gravitational lenses. *Monthly Notices of the Royal Astronomical Society, 443*, 3631.

Misner, C. W., Thorne, K. S., & Wheeler, J. A. (1973). *Gravitation*. San Francisco: W.H. Freeman and Co.

Petters, A. O., Levine, H., & Wambsganss, J. (2001). *Singularity theory and gravitational lensing*. Boston: Birkhäuser.

Saha, P. (2000). Lensing degeneracies revisited. *Astronomical Journal, 120*, 1654.

Schneider, P., Ehlers, J., & Falco, E. E. (1992). *Gravitational lenses*. Berlin: Springer.

Schneider, P., & Sluse, D. (2014). Source-position transformation: An approximate invariance in strong gravitational lensing. *Astronomy & Astrophysics, 564*, A103.

Whitney, H. (1955). On singularities of mappings of Euclidean spaces. I. Mappings of the plane into the plane. *Annals of Mathematics (2), 62*, 374.

Chapter 5
Microlensing Within the Local Group

Broadly speaking, there are two types of strong lensing: macrolensing and microlensing. In the former case, the multiple images of the source can be spatially resolved, while in the latter case, they cannot. The separations of the images are of the order of the Einstein radius θ_E, which is determined by the distances between the source, lens, and observer, as well as the mass of the lens. Thus, for a given resolution limit θ_0, there is macrolensing for $\theta_E \gtrsim \theta_0$ and microlensing for $\theta_E \lesssim \theta_0$.

In this chapter, we investigate microlensing within the Local Group of galaxies. The discovery of vast, unseen quantities of "dark matter" in spiral galaxies prompted researchers to ascertain its composition. The two leading proposals of the day were that dark matter consists of as yet undetected particles distributed continuously in space or that faint compact objects orbit within the extended spherical "halo" of the galaxy along with visible stars. The first type became known as "weakly interacting massive particles" (WIMPs); the second became known as "massive astrophysical compact halo objects" (MACHOs). The growing consensus is that the WIMPs will win in the end, but both merit discussion, as each has played an important role in the ever-increasing prominence of gravitational lensing as a probe of the universe. Since it is the aggregate distribution of WIMPs in the form of massive dark matter halos that is observable in galaxies and clusters, we consider them in subsequent chapters on macrolensing. MACHOs, on the other hand, are massive enough to magnify stars passing behind them, though the images they produce are too close together to be resolved. Thus, these objects fall into the microlensing regime. A more recent application of microlensing is the detection of planets outside the solar system, so-called exoplanets.

Except for the point mass, the analysis in Chap. 2 assumed the lensing object was a galaxy or something larger. In both Chaps. 2 and 4, we solved the lens equation to find the image positions, with the understanding that they could be observed individually. Here we assume only that the centroid (i.e., center of light) and total magnification of the images can be measured. Another difference between earlier

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_5

chapters and this one is that before, the lenses were at cosmological distances so we could not simply add the observer-lens and lens-source distances to compute the observer-source distance. However, the distances encountered in the galaxy and to its neighbors are small enough that we can add them in the conventional way. Another advantage of microlensing is that the flux of the unlensed source is generally known (see Sect. 5.1.1). Although observer, lens, and source always move relative to one another, the corresponding angular motion on the sky is smaller for objects that are farther away. Thus, a microlensing event can be observed (more or less) from start to finish, while the timescale for macrolensing far exceeds the lifetime of an astronomer. The lens potential encountered in microlensing is a discrete sum over the contributions of point masses, rather than an integral over a mass distribution.

We begin by considering microlensing of a point source. The simplest such case is for a lens consisting of a single point mass (Sect. 5.1). We then turn to the case of lensing by several point masses (Sect. 5.2). After formulating the problem in terms of complex variables (Sect. 5.2.1), we specialize to the binary lensing (Sect. 5.2.2) of a source by two stars close together or by a single star orbited by a planet. Extended sources are considered in Sect. 5.3. Putting microlensing theory into practice requires the statistical approach presented in Sect. 5.4. Applications of microlensing are discussed in Sect. 5.5.

5.1 Microlensing by a Point Mass

In Chap. 2 we analyzed lensing by a point mass and obtained in Eq. (2.31) the formula

$$\mu_{\rm tot}(u) = \frac{u^2+2}{u\sqrt{u^2+4}} \tag{5.1}$$

for the total magnification of the two images. Here, u denotes the distance of the source from the lens in units of the Einstein angle θ_E. Note that the value of u, and hence that of $\mu_{\rm tot}$, changes due to the motion of the source relative to the lens. It is conventional to say that a *microlensing event* occurs if $u < 1$ or equivalently, $\mu_{\rm tot} > 3/\sqrt{5} \approx 1.34$, at some time. This restriction ensures that the magnification changes appreciably as u varies with time.

5.1.1 Light Curves

In microlensing observations, one measures the flux of a source as a function of time, i.e., a **light curve**. Expressed in units of the unlensed flux F_0, the light curve gives the variation of the magnification over time. Note that F_0 cannot be determined

exactly, because that would require observing the source at $u \to \infty$. However, it can be determined to within 1% of its actual value (i.e., $\mu_{\rm tot} = 1.01$) when the source is only $\sim 3.5\theta_E$ from the lens. Therefore, we may assume that F_0 is known, so that our light curves show the function $f(t) \equiv \mu_{\rm tot}[u(t)]$, where $u(t)$ is the projected lens-source separation at time t.

To construct a theoretical light curve, we must specify a trajectory for the source. The simplest trajectory, resulting in the so-called Paczyński light curve, is uniform motion across the sky. Working in units of length, we denote the minimum projected lens-source separation by $\xi_{\rm min}$. For a source with linear velocity $v_\perp$ relative to the observer-lens line of sight, the transverse distance between the source and lens at an arbitrary time t takes the form

$$\xi = \sqrt{\xi_{\rm min}^2 + v_\perp^2 t^2}\,. \tag{5.2}$$

For a source at a distance D_s from the observer, a given lens-source separation ξ corresponds to an angular separation on the sky of $\beta = \xi/D_s$. Making the further substitution $u = \beta/\theta_E$, we arrive at the parameterization

$$u(t) = \sqrt{u_{\rm min}^2 + \left(\frac{t}{t_E}\right)^2}\,, \tag{5.3}$$

where

$$t_E = \frac{D_s \theta_E}{v_\perp} \tag{5.4}$$

is the time required for the source to traverse a distance one Einstein radius and therefore known as the **Einstein time**. In the microlensing literature, the Einstein time is often given by the equivalent expression $t_E = D_l \theta_E / v_\perp$, where $v_\perp$ denotes the transverse velocity of the lens relative to the observer-source line of sight.

The quantities we can determine directly from the light curve are the peak magnification $\mu_{\rm max} = \mu_{\rm tot}(0)$ and the time Δt during which $\mu_{\rm tot} \geq 3/\sqrt{5} \approx 1.34$. With the help of Eq. (5.1), we can calculate $u_{\rm min}$ from $\mu_{\rm max}$. The time in which the source moves from $u = u_{\rm min}$ to $u = 1$ is $\Delta t/2$. Thus, according to Eq. (5.3),

$$t_E = \frac{\Delta t}{2\sqrt{1 - u_{\rm min}^2}}\,. \tag{5.5}$$

If the distance to the source is known, then the value of $v_\perp$ follows from the angular velocity of the source, and the value of θ_E follows from Eq. (5.4). If the distance to the lens is also known—no mean feat!—we can then calculate the mass of the lens from Eq. (2.25) for θ_E, with $D_{ls} = D_s - D_l$.

We show light curves for three values of $u_{\rm min}$ in Fig. 5.1. These impact parameters are attained at $t = 0$. Since the magnification depends only on the

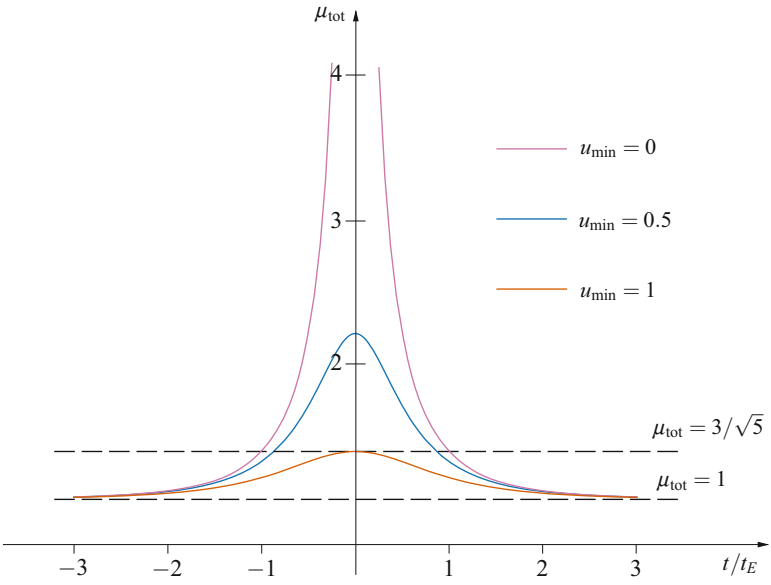


Fig. 5.1 Light curves of a point source passing near a point mass with three different values of the impact parameter $u_{\rm min}$. Notice that the curves become taller and broader as $u_{\rm min}$ decreases

distance $u(t)$ of the source from the lens, the light curves are symmetric about $t = 0$. When $u_{\rm min} = 0$, the source passes directly behind the lens, causing the magnification to diverge there. When $u_{\rm min} = 1$, the source is microlensed only at $t = 0$, since its magnification is below our chosen threshold at all other points along its trajectory. At intermediate values of $u_{\rm min}$, the microlensing event lasts for an interval satisfying $0 < \Delta t < 2t_E$. The peak magnification is finite, but exceeds $3/\sqrt{5}$.

5.1.2 Parallax

If the only information we have is the light curve of a given microlensing event, the parameters of the model are degenerate, i.e., there are continuously many combinations of parameter values that reproduce the same observational data. For example, doubling θ_E in Eq. (5.4) yields no change in the Einstein time if D_s is also halved. These are consequences of the similarity transformations discussed in Sect. 4.7.1.

Parallax observations offer one way around this difficulty. For microlensing events that are longer than about a month, the motion of the Earth around the Sun must be taken into account. In that case, we can no longer approximate the

trajectory of the source as a straight line across the sky. This complication is more than offset by the advantage of having multiple spatial perspectives of the event. Alternatively, one can observe the same event simultaneously from different locations, e.g., from a telescope on Earth and from another far away in space. Of the space-based observatories currently online, the Spitzer Space Telescope is best suited. This observatory is in heliocentric orbit, trailing the Earth at a distance of the order of 1 AU. In addition, it is sensitive to infrared radiation, which most stars emit copiously.

5.1.3 Astrometric Microlensing

The change in magnification with time is one manifestation of microlensing. Another, known as **astrometric microlensing**, is a changing position of the centroid of the images. If the angular separation of the source and the lens is sufficiently small, both of these effects can be observed. Otherwise the magnification is always close to one, and the centroid position is determined, to a good approximation, by the brighter image. Until recently this second case has been of purely theoretical interest. If we know the actual trajectory of the source, we can compare the true position of the source with its apparent position at several epochs to infer the Einstein radius of the lens. If we also know the distances to the lens and source, we can deduce its mass as well.

5.2 Microlensing by Multiple Point Masses

Microlenses often consist of more than one object. Here we set out the formalism for multiple point lenses and then analyze the binary lens in detail. Stars with multiple planetary companions are rather common, so the case of more than two lenses is also important. However, solving the lens equation and finding critical curves and caustics become exceedingly difficult in such cases. Once we have made the conceptual leap from one to two lensing objects, it is straightforward, in principle, to generalize to an arbitrary number of lenses.

What distinguishes the binary lens from a single point mass is twofold: circular symmetry is broken, and caustics are formed. In the axisymmetric case, we found the light curve by taking the composition of $(\mu_{\text{tot}} \circ u)(t)$, where $u(t)$ depended only on two parameters: the minimum lens-source separation, $u_{\min}$, and the Einstein time, t_E. In the present section, the source trajectory is given by the two-component vector $\mathbf{u}(t)$, which requires additional parameters containing directional information. Furthermore, caustic crossings alter light curves by introducing sharp peaks. This means that the maximum magnification need not occur at $u_{\min}$.

5.2.1 Complex Formulation

Consider a collection of n point lenses. We denote the mass of lens j by M_j, its Einstein radius by b_j, and its position on the sky by $\boldsymbol{\theta}_j$. Under the assumption that all of the lenses are at the same distance D_l from the observer,[1] then it follows that $b_j^2 \propto M_j$. If $b^2 = \sum_j b_j^2$, then b is the Einstein radius of a point lens with mass $M = \sum_j M_j$.

The deflection vector of such a system is

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \sum_{j=1}^{n} b_j^2 \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_j}{\left|\boldsymbol{\theta} - \boldsymbol{\theta}_j\right|^2} = b^2 \sum_{j=1}^{n} m_j \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_j}{\left|\boldsymbol{\theta} - \boldsymbol{\theta}_j\right|^2}\,, \tag{5.6}$$

where $m_j = M_j/M$ is the fraction of the total mass contained in lens j. (Note that $\sum_j m_j = 1$.) The lens equation is then

$$\boldsymbol{\beta} = \boldsymbol{\theta} - b^2 \sum_{j=1}^{n} m_j \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_j}{\left|\boldsymbol{\theta} - \boldsymbol{\theta}_j\right|^2}\,. \tag{5.7}$$

Setting $\mathbf{U} \equiv \boldsymbol{\beta}/b$ and $\mathbf{X} \equiv \boldsymbol{\theta}/b$, the lens equation takes the form

$$\mathbf{U} = \mathbf{X} - \sum_{j=1}^{n} m_j \frac{\mathbf{X} - \mathbf{X}_j}{\left|\mathbf{X} - \mathbf{X}_j\right|^2}\,, \tag{5.8}$$

where the source position $\mathbf{U}$ has components (u, v) and the image positions $\mathbf{X}$ have components (x, y).

The two-dimensional nature of this equation suggests switching to complex variables (see Appendix B), where vectors are replaced by numbers. Introducing the variables $\zeta = u + iv$ and $z = x + iy$, we arrive at the **complex lens equation**

$$\begin{aligned} \zeta &= z - \sum_{j=1}^{n} m_j \frac{z - z_j}{\left|z - z_j\right|^2} \\ &= z - \sum_{j=1}^{n} \frac{m_j}{z^* - z^*{}_j}\,, \end{aligned} \tag{5.9}$$

[1] The distances between lens stars are ∼AU, while those between observer, lenses, and source are ∼kpc.

where an asterisk denotes complex conjugation. We define the complex deflection function

$$\boxed{\alpha(z) \equiv \sum_{j=1}^{n} \frac{m_j}{z - z_j}} \tag{5.10}$$

so that the lens equation can be written as

$$\zeta = z - [\alpha(z)]^* \,. \tag{5.11}$$

Note that $\alpha(z)$ is *meromorphic* (Appendix B.3), i.e., *analytic* (Appendix B.1) for all z except the z_j, where it has poles.

We caution the reader that the complex deflection function of an *arbitrary* mass distribution need not be rational nor even meromorphic. For example, it may depend on both z and z^*. While we use complex variables and functions extensively in this chapter, their drawbacks outweigh their advantages when continuous matter distributions are involved.

The lens equation (5.9) has at most $5n - 5$ solutions for $n > 1$ (Khavinson and Neumann 2006, Theorem 1; see also Rhie 2003). The physical interpretation is that a system of n point lenses may produce a maximum of $5n - 5$ images. The actual number of images depends on the value of ζ, but is at least $n+1$ (Petters et al. 2001, Theorem 11.7). Recall that a source far from a single point mass appears as two images: the primary image asymptotically approaches the true source position with a magnification of unity, while the secondary image approaches the lens position with vanishing magnification. In the case of $n > 1$ point lenses, a source far from the lenses ($|\zeta| \gg \max_j |z_j|$) will appear as a primary image with a magnification of unity and n secondary images with vanishing magnification at the lens positions. (The cases of $n = 2$ and $n > 2$ are explored in Problems 5.3 and 5.4, respectively.)

As usual, the inverse magnification is given by

$$\mu^{-1} = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}.$$

Our goal is to express μ^{-1} as a function of z and z^*. To this end, we introduce the **Wirtinger operators**

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right) \tag{5.12a}$$

$$\frac{\partial}{\partial z^*} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right) \tag{5.12b}$$

(see Appendix B.1). For any complex function $f(z, z^*)$, one can show that

$$\left(\frac{\partial f}{\partial z}\right)^* = \frac{\partial f^*}{\partial z^*} \tag{5.13a}$$

$$\left(\frac{\partial f}{\partial z^*}\right)^* = \frac{\partial f^*}{\partial z}\,. \tag{5.13b}$$

The partial derivatives with respect to x and y can be expressed as

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial z} + \frac{\partial}{\partial z^*} \tag{5.14a}$$

$$\frac{\partial}{\partial y} = i\left(\frac{\partial}{\partial z} - \frac{\partial}{\partial z^*}\right)\,. \tag{5.14b}$$

Thus, we see that

$$\begin{aligned}\mu^{-1}(z, z^*) &= \frac{\partial\zeta}{\partial z}\frac{\partial\zeta^*}{\partial z^*} - \frac{\partial\zeta}{\partial z^*}\frac{\partial\zeta^*}{\partial z} \\ &= \left|\frac{\partial\zeta}{\partial z}\right|^2 - \left|\frac{\partial\zeta}{\partial z^*}\right|^2 \\ &= 1 - \left|\frac{\partial\zeta}{\partial z^*}\right|^2\,,\end{aligned} \tag{5.15}$$

since $\partial\zeta/\partial z = 1$ according to Eq. (5.9). From Eqs. (5.11) and (5.13a), we have

$$\frac{\partial\zeta}{\partial z^*} = -\frac{\partial\alpha^*}{\partial z^*} = -\left(\frac{\partial\alpha}{\partial z}\right)^* = -\left(\frac{\mathrm{d}\alpha}{\mathrm{d}z}\right)^*\,, \tag{5.16}$$

where the last equality follows from the fact that α is analytic at each point where it is defined. The inverse magnification then takes the simple form

$$\boxed{\mu^{-1}(z, z^*) = 1 - \left|\frac{\mathrm{d}\alpha}{\mathrm{d}z}\right|^2}\,. \tag{5.17}$$

On a critical curve, we have $\mu^{-1} = 0$ so that $|\mathrm{d}\alpha/\mathrm{d}z| = 1$. This, in turn, implies that

$$\frac{\mathrm{d}\alpha}{\mathrm{d}z} = \pm e^{\pm i\phi}\,, \tag{5.18}$$

where $0 \leq \phi < 2\pi$. The choice of signs on the right-hand side is arbitrary. For the sake of consistency with other authors (e.g., Petters et al. 2001; Mollerach and Roulet 2002), we write

$$\frac{\mathrm{d}\alpha}{\mathrm{d}z} = -e^{-i\phi}\,. \tag{5.19}$$

Comparing this to the derivative of Eq. (5.10), we find

$$\sum_{j=1}^{n} \frac{m_j}{(z-z_j)^2} = e^{-i\phi}\,. \tag{5.20}$$

This is equivalent to a set of polynomial equations $Q(z;\phi) = 0$ of degree $2n$ in z, parametrized by ϕ. We denote the solutions by $z^{(1)}(\phi),\ z^{(2)}(\phi), \ldots, z^{(2n)}(\phi)$. Note that if $z^{(i)}(\phi_1) = z^{(j)}(\phi_2)$, then $\phi_1 = \phi_2$, since the left-hand side of Eq. (5.20) is independent of ϕ and the right-hand side is independent of z. Thus, the solutions $z^{(i)}$ are pairwise disjoint curves, which are not necessarily closed. However, their union does consist of a number ($\leq 2n$) of closed curves. In the lensing literature, it is these closed curves that are the critical curves.

Critical curves merge or separate when $z^{(i)}(\phi_0) = z^{(j)}(\phi_0)$ for some ϕ_0 and $i \neq j$. Suppose $z^{(1)}(\phi_0) = z^{(2)}(\phi_0) \equiv \xi$. Then

$$\begin{aligned} Q(z;\phi_0) &= \prod_{j=1}^{2n} \left(z - z^{(j)}(\phi_0)\right) \\ &= (z-\xi)^2 \prod_{j=3}^{2n} \left(z - z^{(j)}(\phi_0)\right). \end{aligned} \tag{5.21}$$

Hence, $Q'(\xi;\phi_0) = 0$. Similarly, if $k > 1$ of the $z^{(i)}(\phi_0)$ are all equal to ξ, then the first $k-1$ derivatives of $Q(z;\phi_0)$ are zero at $z = \xi$.

5.2.2 *Binary Lens*

We now specialize to the binary case ($n = 2$). For convenience we define the real axis to be the line containing the two lenses, with the origin midway between them. Letting d be the distance from the origin to each lens, we have $z_1 = d$ and $z_2 = -d$ and a lens separation of $\Delta \equiv 2d$. We also define the mass ratio to be $q \equiv M_1/M_2 = m_1/m_2$. The lens equation (5.9) reduces to

$$\zeta = z - \frac{m_1}{z^* - d} - \frac{m_2}{z^* + d}\,. \tag{5.22}$$

The complex conjugate of this equation allows us to solve for z^* in terms of z:

$$z^* = \zeta^* + \frac{m_1}{z-d} + \frac{m_2}{z+d}\,. \tag{5.23}$$

Substituting this result into Eq. (5.22) leads to the fifth order polynomial equation

$$\begin{aligned} P(z) \equiv (z^2 - d^2) \left\{ z + d(m_1 - m_2) + (z^2 - d^2)[\zeta^* + d(m_1 - m_2)] \right\} \\ -(z - \zeta) \left\{ [\zeta^*(z^2 - d^2) + z + d(m_1 - m_2)]^2 - d^2(z^2 - d^2)^2 \right\} \\ = 0 \end{aligned} \tag{5.24}$$

(cf. Mollerach and Roulet 2002, p. 130). Thus, the binary lens admits a maximum of five images (see Problem 5.3). A given solution of Eq. (5.24) is physical if and only if its conjugate agrees with the value obtained from Eq. (5.23).

Recall that we can determine the number of images for a given source position ζ, even without solving the lens equation. We start by generating the critical curves in the image plane, from which the caustics in the source plane can be constructed. For a binary lens, the critical curves satisfy Eq. (5.20) with $n = 2$:

$$\frac{m_1}{(z - d)^2} + \frac{m_2}{(z + d)^2} = e^{-i\phi} . \tag{5.25}$$

Multiplying both sides by $(z - d)^2(z + d)^2$ yields the quartic equation

$$Q(z) \equiv m_1(z + d)^2 + m_2(z - d)^2 - e^{-i\phi}(z^2 - d^2)^2 = 0 \tag{5.26}$$

(cf. Petters et al. 2001; Mollerach and Roulet 2002), which, with the help of Eqs. (5.22) and (5.25) can be rewritten as

$$Q(z) = (z - d)^2(z + d)^2 \left(\frac{\partial \zeta^*}{\partial z} - e^{-i\phi} \right) . \tag{5.27}$$

Recall that critical curves touch when $Q'(z) = 0$. It follows from Eq. (5.27) that if $Q(z) = 0$, then $Q'(z) = 0$ if and only if $\partial^2 \zeta^* / \partial z^2 = 0$. This is equivalent to the condition

$$\frac{\mathrm{d}^2 \alpha}{\mathrm{d}z^2} = 0 . \tag{5.28}$$

According to Eq. (5.10) , the condition (5.28) takes the form

$$\frac{m_1}{(z - d)^3} + \frac{m_2}{(z + d)^3} = 0 \tag{5.29}$$

in the binary case $n = 2$. Solving this equation for z, we obtain

$$z = d \left[\frac{2}{1 - (-m_1/m_2)^{1/3}} - 1 \right] , \tag{5.30}$$

where there are three solutions, corresponding to the cube roots of -1: -1 and $e^{\pm i\pi/3}$. Substituting z from Eq. (5.30) into Eq. (5.25) and solving for d, we obtain

$$\Delta_{\rm WI} \equiv 2d_{\rm WI} = \left(m_1^{1/3} + m_2^{1/3}\right)^{3/2} \tag{5.31a}$$

$$\Delta_{\rm IC} \equiv 2d_{\rm IC} = \left(m_1^{1/3} + m_2^{1/3}\right)^{-3/4} \tag{5.31b}$$

for the roots -1 and $e^{\pm i\pi/3}$, respectively. The subscripts W, I, and C refer to lenses of wide, intermediate, or close separation. When the lens separation Δ satisfies $\Delta > \Delta_{\rm WI}$, there are two critical curves, one surrounding each lens. These critical curves touch at a single point on the negative real axis when $\Delta = \Delta_{\rm WI}$. When $\Delta_{\rm IC} < \Delta < \Delta_{\rm WI}$, there is a single critical curve surrounding both lenses. At $\Delta = \Delta_{\rm IC}$, two critical curves pinch off at two points, whose coordinates are conjugate. When $\Delta < \Delta_{\rm IC}$, there are three critical curves—one surrounding both lenses and two smaller ones inside this curve. Using the lens equation (5.22) to move from the image plane to the source plane, we find that widely separated lenses have two caustics with four cusps each, lenses with intermediate separation have one caustic with six cusps, and closely separated lenses have three caustics, one with four cusps and two with three cusps.

There are two special cases that are important in microlensing: where the lenses have roughly equal masses, or where one lens is much more massive than the other. These correspond to binary stars and a single star orbited by a planet, respectively. When $m_1 = m_2 = 1/2$, the lens equation (5.24) remains quintic in z. However, Eq. (5.26) for the critical curves reduces to a quadratic equation in z^2, which is straightforward to solve analytically. The corresponding lens separations at which mergers occur are $\Delta_{\rm WI} = 2$ and $\Delta_{\rm IC} = 1/\sqrt{2}$.

The critical curves and caustics for $m_1 = m_2 = 0.5$ and for $m_1 = 0.1,\ m_2 = 0.9$ are shown in Figs. 5.2 and 5.3, respectively. In both cases, the caustics are smaller than the critical curves. The caustics and the critical curves are symmetric about the horizontal axis; only for $m_1 = m_2$ is there also symmetry about the vertical axis. The value of the mass ratio $m_1/m_2 \approx 0.11$ used in Fig. 5.3 is far larger than what one would expect for a star-planet system, but better illustrates the differences between the cases of equal and unequal masses. In the extreme case $m_1 \ll m_2$, Eqs. (5.31a) reduce to

$$\Delta_{\rm WI} = 1 + \frac{3}{2} m_1^{1/3} \tag{5.32a}$$

$$\Delta_{\rm IC} = 1 - \frac{3}{4} m_1^{1/3} \tag{5.32b}$$

to leading order in m_1 (cf. Mollerach and Roulet 2002, §6.1.1). Even for a particularly massive planet such as Jupiter, the mass ratio of planet to star is $\sim$0.001. Thus, Eq. (5.32) should be a good approximation for planetary microlensing.

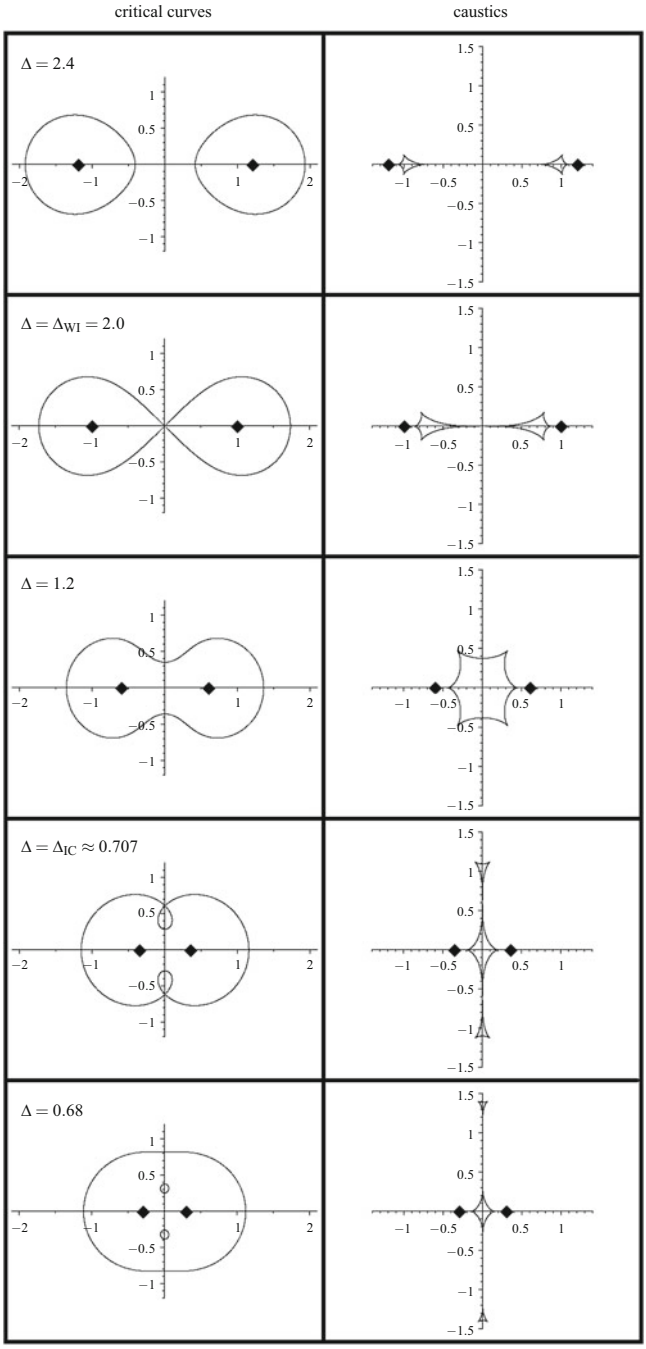


Fig. 5.2 Critical curves and caustics for $m_1 = m_2 = 0.5$. The axes in Figs. 5.2, 5.3, 5.4 and 5.5 are in units of the combined Einstein angle $b = \sqrt{b_1^2 + b_2^2}$. The points in the image plane indicate the locations of the point masses, and they are reproduced in the source plane to guide the eye

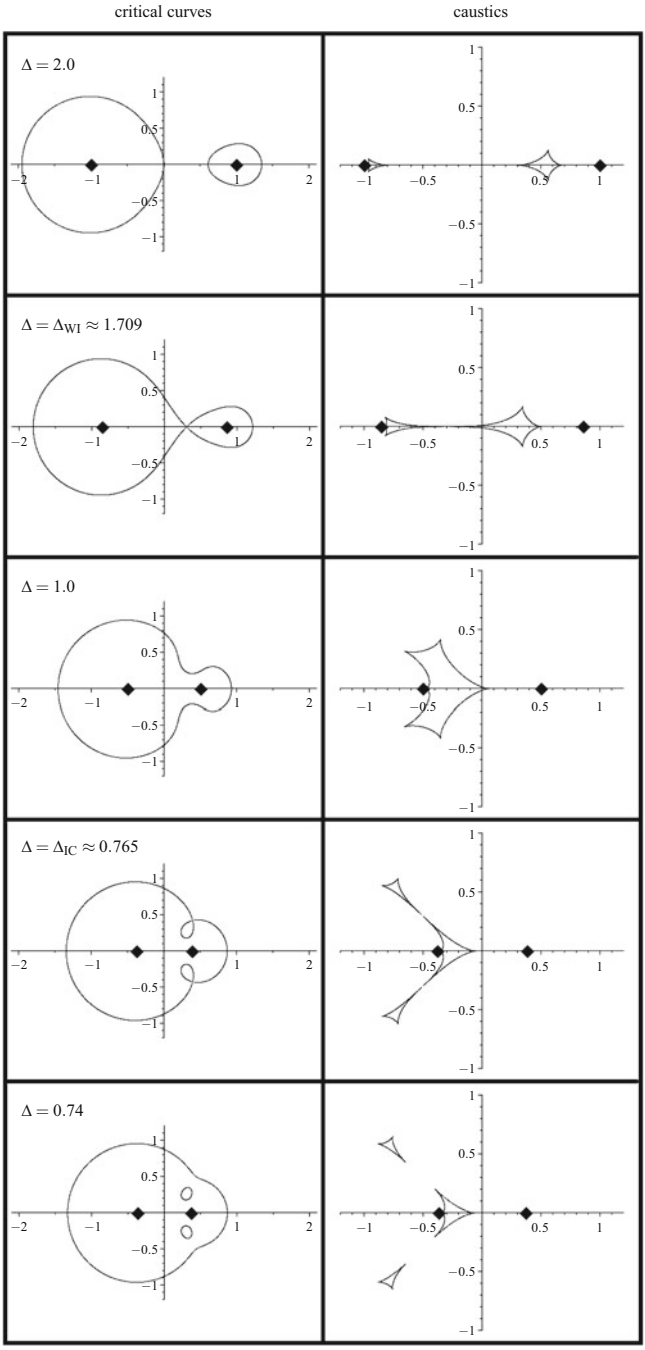


Fig. 5.3 Critical curves and caustics for $m_1 = 0.1$ (right) and $m_2 = 0.9$ (left)

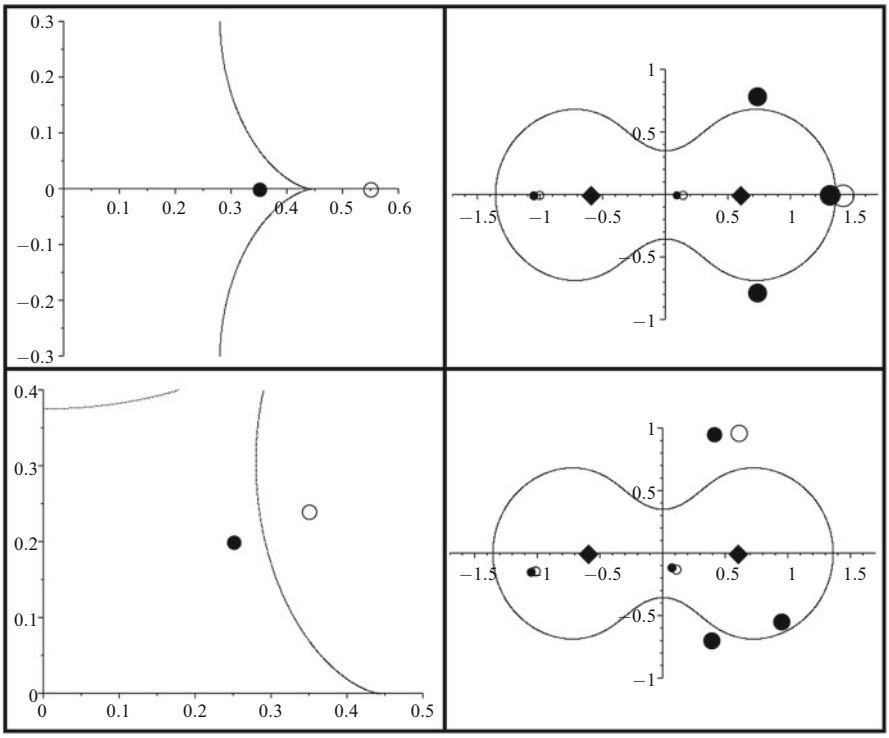


Fig. 5.4 Lensing of sources near a cusp (top) and a fold (bottom). A portion of the caustic, along with two source positions, is shown in each of the left-hand panels, while the associated critical curve and image positions are shown in the right-hand panels. A source outside the caustic (open circle) produces three images (open circles), while a source inside (filled circle) produces five images (filled circles). The ratio of image area to source area gives the (unsigned) magnification of that image. While the images of a circular source are ellipses, this effect is not shown. The lenses (filled diamonds) have equal mass

Figures 5.4 and 5.5 show image configurations for different source positions. Sources outside the caustic produce three images, while those inside give rise to five images. Consistent with our discussion of critical points in Sect. 4.5, we see that the image number changes by two when the source crosses the caustic. The individual images are elliptical, and they differ in brightness from the source. However, the images cannot actually be resolved in microlensing, so the relevant quantity is the total magnification of the source. As the source moves far from the lens, the total magnification approaches unity (cf. Sect. 5.2.1).

As the source crosses from outside to inside the caustic, either a bright pair of images forms or one bright image splits into three bright images. These cases arise for fold and cusp crossings, respectively (cf. Sect. 4.5). This behavior is most readily seen in the light curves shown in Fig. 5.6 with $m_1 = 0.2$ ($m_2 = 0.8$). For a source that crosses a cusp from either direction, the total magnification diverges. By

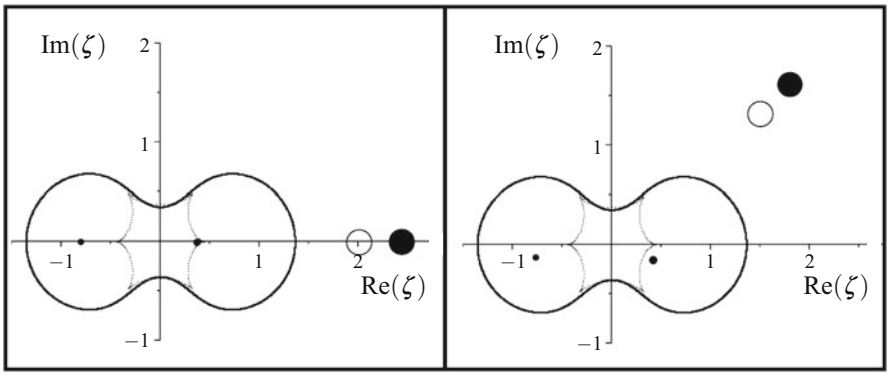


Fig. 5.5 Lensing of sources far from the caustic (thin curve) of an equal-mass binary system in line with a cusp (left) and a fold (right). The source positions are indicated by open circles and the images by filled circles. The area of each circle has the same interpretation as in Fig. 5.4. The critical curve (thick curve) and lens positions (filled diamonds) are also shown. Note that the image near the actual source position has magnification approaching unity, while those near the lens positions become very faint

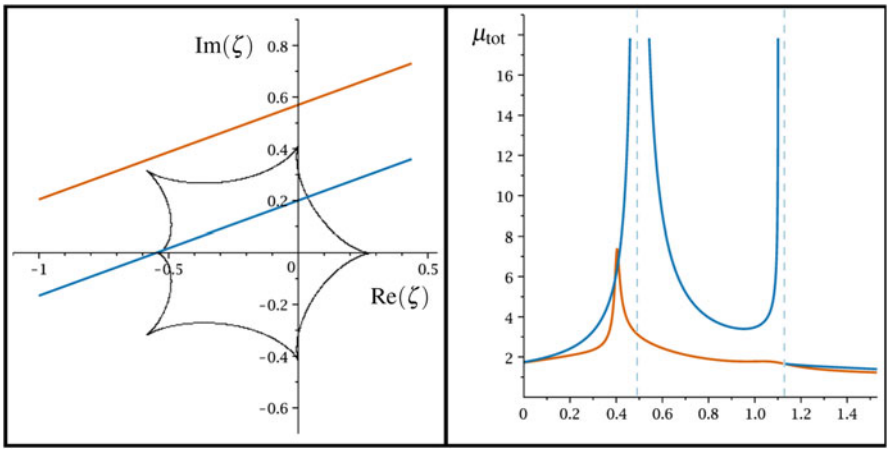


Fig. 5.6 Light curves for a binary system with $m_1 = 0.2$ ($m_2 = 0.8$). Source trajectories (left-hand panel) have the same color as the corresponding light curves (right-hand panel)

contrast, the magnification diverges for a source approaching a fold from the inside, but stays finite for a source approaching a fold from the outside. One can infer the presence of cusps even for a source that doesn't cross the caustic if the light curve possesses spikes of finite height. We would see qualitatively similar behavior for a system with any mass ratio. Thus, we cannot infer the mass ratio from a single light curve (see Sect. 5.4).

5.3 Microlensing of an Extended Source

Thus far, we have considered the microlensing of a point source, where the magnification diverges for a source on a caustic (see Figs. 5.1 and 5.6). For an extended source, there is no such divergence, since the magnification is a normalized convolution of the point-source magnification with the surface brightness of the extended source, as shown in Eq. (4.82). In this section we study the magnification of an extended source using the simple model of a circular disk with uniform surface brightness. The three-dimensional structure of the source star causes the surface brightness to vary in an effect known as **limb darkening** that may need to be taken into account for practical applications.

5.3.1 Single Lens

We first examine finite source effects for the case of a single lens. We work in units scaled by the Einstein radius: u is the angular position of the source scaled by θ_E (as above), and ρ is the angular radius of the source scaled by θ_E. Using the magnification for a point lens from Eq. (5.1) in the convolution integral (4.82), we can write the magnification as

$$\mu_{\rm ext}(u;\rho) = \frac{1}{\pi\rho^2}\int_0^{2\pi} {\rm d}\phi \int_0^{\rho} {\rm d}r\; r\; \mu_{\rm tot}\left(\sqrt{u^2+r^2+2ur\cos\phi}\right) . \tag{5.33}$$

The integral can be evaluated in terms of special functions (Witt and Mao 1994), but it is useful to examine two limits.

First consider a source at the origin, so $u = 0$. Then we can evaluate

$$\mu_{\rm ext}(0;\rho) = \frac{2}{\rho^2}\int_0^{\rho} \frac{r^2+2}{\sqrt{r^2+4}}\,{\rm d}r = \sqrt{1+\frac{4}{\rho^2}} . \tag{5.34}$$

When the source is small ($\rho \ll 1$), the magnification scales as $\mu_{\rm ext}(0;\rho) \approx 2/\rho$; it diverges in the limit of a point source (as expected). As the source size increases, the magnification decreases. In the limit of a large source $\rho \gg 1$ the magnification $\mu_{\rm ext}(0;\rho) \approx 1 + 2\rho^{-2}$ deviates only slightly from unity.

The other interesting case is a source that is small compared with its distance from the origin, so we can make a Taylor series expansion in $\rho/u \ll 1$. Since $0 < r < \rho$, we can first expand the integrand in Eq. (5.33) as a Taylor series in r/u and then evaluate the integral to obtain a series in ρ/u:

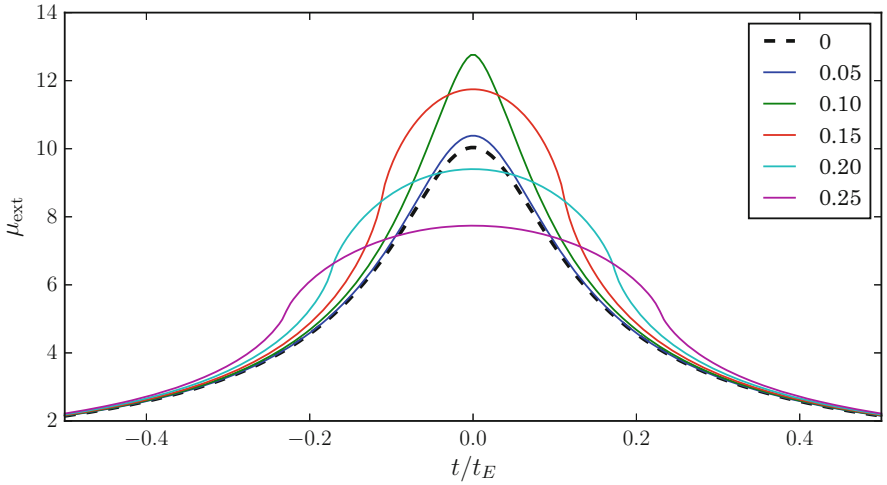


Fig. 5.7 Light curves for extended sources. The trajectory has impact parameter $u_{\rm min} = 0.1$. The thick dashed curve corresponds to a point source, while the curves with different colors correspond to different scaled source sizes ρ as indicated in the legend

$$\begin{aligned}\mu_{\rm ext}(u;\rho) &= \frac{1}{\pi\rho^2}\int_0^{2\pi} {\rm d}\phi \int_0^{\rho} {\rm d}r\, r\, \frac{u^2+2}{u\sqrt{u^2+4}}\left[1 - \frac{8\cos\phi}{8+6u^2+u^4}\left(\frac{r}{u}\right)\right.\\ &\left. + \frac{4(2(1+u^2)+3(2+u^2)\cos 2\phi)}{(2+u^2)(4+u^2)^2}\left(\frac{r}{u}\right)^2 + \mathcal{O}\left(\frac{r}{u}\right)^3\right]\\ &= \frac{u^2+2}{u\sqrt{u^2+4}}\left[1 + \frac{4(1+u^2)}{(2+u^2)(4+u^2)^2}\left(\frac{\rho}{u}\right)^2 + \mathcal{O}\left(\frac{\rho}{u}\right)^3\right]. \quad (5.35)\end{aligned}$$

At lowest order, the magnification reduces to the result for a point source. The first-order correction term vanishes by symmetry in the ϕ integration. Thus, corrections enter only at second order. The coefficient of the correction term is strictly positive, so finite source effects increase the magnification relative to the point-source case.

For intermediate cases, we can evaluate the integral numerically and obtain light curves like those shown in Fig. 5.7. We see that finite source effects raise the wings of the light curve. When the size is small, the magnification increases even at the peak of the light curve. However, when the size is large, finite source effects make the light curve flatter and broader, so the peak winds up being lower than the point source limit. The transition is such that finite source effects increase the magnification when $\rho \lesssim 2u$, while they decrease the magnification when $\rho \gtrsim 2u$ (see Witt and Mao 1994; the transition point can be estimated as the place where Eqs. (5.34) and (5.35) give similar magnifications).

5.3.2 *Fold Caustic*

It is also interesting to examine the case of a source near a caustic like what appears in a binary lens. We actually study the ideal fold caustic introduced in Sect. 4.5.1. According to Eq. (4.63), the total magnification of a point source near a fold is

$$\mu_{\rm tot}(v) = \frac{C}{\sqrt{v}}\Theta(v)\,, \tag{5.36}$$

where the fold caustic lies along the horizontal axis, so v measures the perpendicular distance of the source from the caustic, and the scale factor is $C \equiv \left(K\sqrt{-3h}\right)^{-1}$. Without loss of generality, we can take $h < 0$ such that the region inside the caustic is $v > 0$. This convention, which is opposite to that used in Sect. 4.5.1, allows us to make use of the Heaviside step function. (As a reminder, $\Theta(x)$ is the Heaviside step function that is 1 for $x > 0$ and 0 for $x < 0$.) We can then write the magnification as

$$\begin{aligned}\mu_{\rm ext}(u, v; \rho) &= \frac{C}{\pi\rho^2}\int_{-\infty}^{\infty} {\rm d}u' \int_{-\infty}^{\infty} {\rm d}v' \, \frac{\Theta(v')}{\sqrt{v'}}\Theta\left(\rho - \sqrt{(u'-u)^2 + (v'-v)^2}\right) \\ &= \frac{C}{\pi\rho^2}\int_{-\infty}^{\infty} {\rm d}x \int_{-\infty}^{\infty} {\rm d}y \, \frac{\Theta(y+v)}{\sqrt{y+v}}\Theta\left(\rho - \sqrt{x^2+y^2}\right), \end{aligned} \tag{5.37}$$

where $x = u' - u$, $y = v' - v$. Integrating over x leads to the single integral

$$\mu_{\rm ext}(u, v; \rho) = \frac{2C}{\pi\rho^2}\int_{-\min(\rho, v)}^{\rho} \sqrt{\frac{\rho^2 - y^2}{y+v}}\,{\rm d}y\,. \tag{5.38}$$

We can identify the scaling with source size by changing integration variables to $w = y/\rho$ and then writing

$$\mu_{\rm ext}(u, v; \rho) = \frac{C}{\sqrt{\rho}} f_{\rm fold}\left(\frac{v}{\rho}\right), \tag{5.39}$$

where the function $f_{\rm fold}(x)$ is defined by

$$f_{\rm fold}(x) \equiv \frac{2}{\pi}\int_{-\min(1,x)}^{1} \sqrt{\frac{1-w^2}{w+x}}\,{\rm d}w\,. \tag{5.40}$$

The form of Eq. (5.39) indicates that light curves for different source sizes are all scaled versions of one another; the time (which is implicit in v) scales as $1/\rho$, while the amplitude scales as $1/\sqrt{\rho}$. Figure 5.8 shows both the scaled version of the fold light curve and versions that are properly scaled for different source sizes.

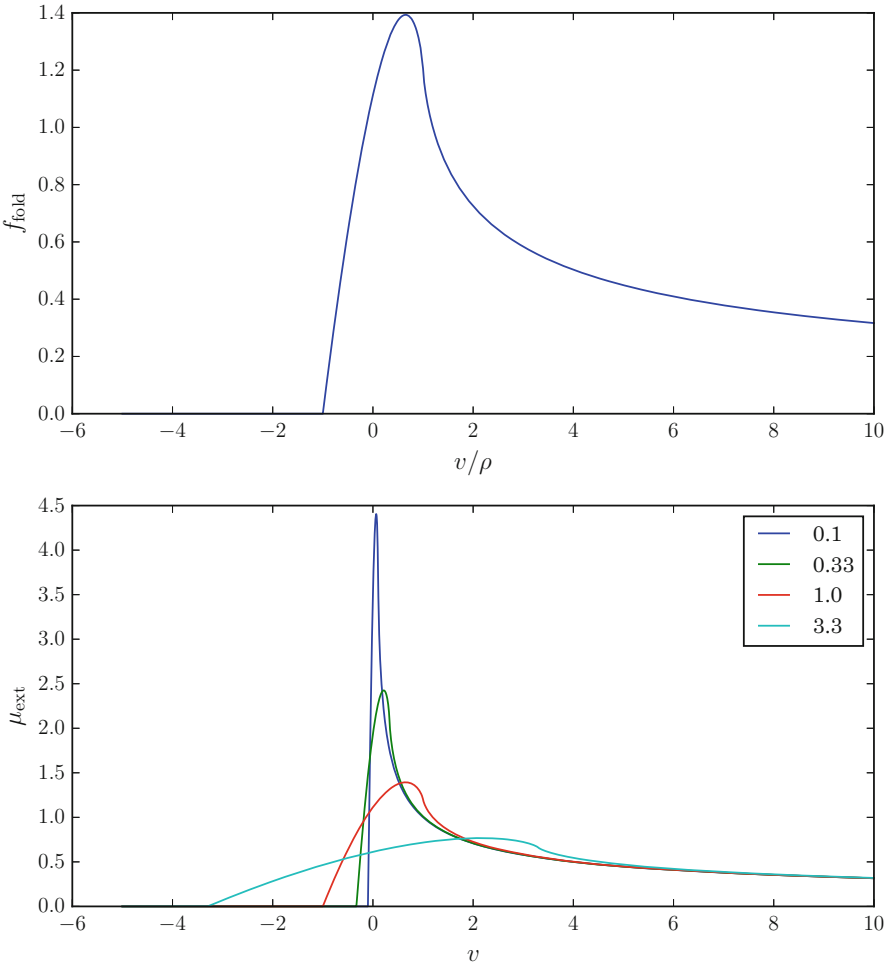


Fig. 5.8 Light curves for an extended source crossing an idealized fold caustic. The top panel shows the scaled light curve $f_{\rm fold}(v/\rho)$. The bottom panel shows the actual light curves for different scaled source sizes ρ as indicated in the legend. Here the fold scaling factor is $C = 1$

5.4 Microlensing Statistics

The first question we must answer in order to put into practice the microlensing theory that we have developed is how likely it is that a given star is being lensed at some time. This is characterized by the **optical depth**,[2] which is the fraction

[2]In stellar and planetary astrophysics, the optical depth measures the degree of photon scattering in a medium. The analogy between scattering and lensing is close enough to make the extension of this concept to the present context worthwhile.

of a region of the source plane covered by stellar Einstein rings. This definition is chosen to be consistent with the notion that a microlensing event occurs when a source star is within one Einstein radius of a lensed star. When the optical depth is low, it is just the probability that a source will be microlensed. Optical depth loses its connection with probability when it reaches a significant fraction of unity, because the stars in that case combine to act as a single lens rather than a collection of individual lenses. In other words, the sky is so crowded with lenses that their Einstein rings overlap (cf. Sect. 6.4.2). For the purposes of including binary lenses in our determination of optical depth, we define the effective Einstein radius of such a system to be the quadrature sum of the stars' Einstein radii. If the minimum magnification (i.e., maximum lens-source separation) for which microlensing is said to occur is lowered (raised), the optical depth will likewise increase (decrease) by a multiplicative factor.

Consider a source at distance D_s. Conceptually, the optical depth is the fraction of the source plane that is covered by the Einstein areas of intervening lenses. For specificity, consider an area that subtends solid angle Ω. We are interested in lenses within the same solid angle. Consider lenses in a slab between D_l and $D_l + \mathrm{d}D_l$; its volume is $\Omega D_l^2\,\mathrm{d}D_l$. Suppose $(\mathrm{d}n/\mathrm{d}m)\,\mathrm{d}m$ is the number density of lenses in the slab with masses between m and $m + \mathrm{d}m$. Then the number of lenses in question is

$$\mathrm{d}N = \Omega D_l^2 \frac{\mathrm{d}n}{\mathrm{d}m}\,\mathrm{d}m\,\mathrm{d}D_l\,. \tag{5.41}$$

The Einstein area of each lens (in physical units in the source plane) is

$$\pi R_E^2 = \frac{4\pi G m}{c^2}\frac{D_{ls} D_s}{D_l}\,. \tag{5.42}$$

The total Einstein area is then $\pi R_E^2\,\mathrm{d}N$. To obtain the optical depth, we divide this by the area of the region of the source plane, ΩD_s^2:

$$\begin{aligned}\mathrm{d}\tau &= \frac{1}{\Omega D_s^2} \times \Omega D_l^2 \frac{\mathrm{d}n}{\mathrm{d}m}\,\mathrm{d}m\,\mathrm{d}D_l \times \frac{4\pi G m}{c^2}\frac{D_{ls} D_s}{D_l} \\ &= \frac{4\pi G}{c^2}\,\frac{D_l D_{ls}}{D_s}\,\mathrm{d}D_l\, m \frac{\mathrm{d}n}{\mathrm{d}m}\,\mathrm{d}m\,. \end{aligned} \tag{5.43}$$

We obtain the total optical depth now by integrating over masses and distances:

$$\begin{aligned}\tau(D_s) &= \frac{4\pi G}{c^2}\int_0^{D_s} \frac{D_l(D_{ls})}{D_s}\int_0^\infty m\frac{\mathrm{d}n}{\mathrm{d}m}\mathrm{d}m\,\mathrm{d}D_l \\ &= \frac{4\pi G}{c^2}\int_0^{D_s} \frac{D_l(D_s - D_l)}{D_s}\,\rho\,\mathrm{d}D_l\,, \end{aligned} \tag{5.44}$$

where $D_{ls} = D_s - D_l$, and the mass density is given by

$$\rho = \int_0^\infty m \frac{\mathrm{d}n}{\mathrm{d}m}\,\mathrm{d}m\,. \tag{5.45}$$

Let us see how the optical depth depends on ρ. In the special case of uniform density, we can evaluate the integral to obtain

$$\tau_{\mathrm{uni}} = \frac{2\pi G\rho D_s^2}{3c^2}\,. \tag{5.46}$$

We can write the density in terms of the rotation speed, v_{rot}, at a distance r from the Galactic Center by means of the relations

$$\frac{GM(r)}{r^2} = \frac{v_{\mathrm{rot}}^2}{r}\,; \quad M(r) = \frac{4}{3}\pi r^3\rho\,, \tag{5.47}$$

where $M(r)$ is the mass interior to r. For an observer at Earth and a source at the Galactic Center, we have $r = r_\oplus = D_s$, with the galactocentric radial coordinate of Earth denoted by $r_\oplus$. Substituting for ρ then leads to the simple expression

$$\tau_{\mathrm{uni}} = \frac{v_{\mathrm{rot}}^2}{2c^2} \sim 10^{-6}\,. \tag{5.48}$$

Thus, we must observe a million or so stars to have a reasonable chance to observe a microlensing event!

If the density is not uniform, we must exercise some care. In particular, galactic density profiles are given as functions of galactocentric coordinates. For a spherically symmetric galaxy, the density profile can be written as $\varrho(r)$. The density profile as measured from Earth would then be a function of the distance to the lens and the angle of the lens relative to the line of sight to the Galactic Center. For a lens at distance D_l and angle b, its galactocentric distance is determined from the law of cosines as $r^2 = D_l^2 + r_\oplus^2 - 2D_l r_\oplus \cos b$. Thus, the mass density with respect to Earth is

$$\rho(D_l, b) = \varrho\sqrt{(D_l^2 + r_\oplus^2 - 2D_l r_\oplus \cos b)}\,. \tag{5.49}$$

For a nonspherical mass distribution, the function ρ also depends on the galactic longitude.

Suppose that the galaxy is described by the singular isothermal profile of Eq. (2.42). The optical depth is then

$$\tau_{\mathrm{iso}} = \frac{4\pi G}{c^2}\int_0^{D_s} \frac{D_l(D_s - D_l)}{D_s}\,\rho(D_l, b)\,\mathrm{d}D_l$$

$$= \frac{v_{\rm rot}^2}{c^2 D_s} \int_0^{D_s} \frac{D_l(D_s - D_l)}{D_l^2 + r_\oplus^2 - 2D_l r_\oplus \cos b} \, {\rm d}D_l \,, \tag{5.50}$$

where $v_{\rm rot}$ is a constant. For a source at a distance $D_s = r_\oplus$, the optical depth reduces to

$$\tau_{\rm iso} = \frac{v_{\rm rot}^2}{c^2} \int_0^1 \frac{x(1-x)}{x^2 - 2x\cos b + 1} \, {\rm d}x \,, \tag{5.51}$$

where $x \equiv D_l/D_s$. In general, this integral can only be evaluated numerically. For a source at the Galactic Center ($b = 0$), we have

$$\tau_{\rm iso} = \frac{v_{\rm rot}^2}{c^2} \int_0^1 \frac{x \, {\rm d}x}{1-x} \to \infty \tag{5.52}$$

(cf. Dodelson 2017, p. 94). This divergence is an artifact of the singularity in ϱ at the Galactic Center. We can obtain a finite result for $D_s < r_\oplus$ and $b = 0$:

$$\begin{aligned} \tau_{\rm iso} &= \frac{v_{\rm rot}^2}{c^2 D_s} \int_0^{D_s} \frac{D_l(D_s - D_l)}{(D_l - r_\oplus)^2} {\rm d}D_l \\ &= \frac{v_{\rm rot}^2}{c^2 D_s} \left[(2r_\oplus - D_s) \log\left(\frac{r_\oplus}{r_\oplus - D_s} \right) - 2D_s \right]. \end{aligned} \tag{5.53}$$

For a source in galactic bulge, we find that $\tau_{\rm iso} \sim \tau_{\rm uni}$. Since the divergence in $\tau_{\rm iso}$ as $D_s \to r_\oplus$ is logarithmic, this rough equality holds almost all the way to the Galactic Center.

The optical depth constrains the abundance of microlenses, but tells us nothing further about their properties. Inferring lens parameters from a single event is not possible, so we must observe an ensemble of events from which statistical conclusions can be drawn. The distributions of event duration (Mollerach and Roulet 2002) and magnification are often used for this purpose. Deriving them does not strictly involve lensing, so we move directly to the astrophysical applications of microlensing.

5.5 Applications

5.5.1 Probing Dark Matter

As noted in the opening of this chapter, microlensing was initially used to search for MACHO dark matter in the Milky Way. Paczyński (1986) developed the theory of microlensing by a point mass and pointed out that by monitoring a few million

stars in the Large and Small Magellanic Clouds,[3] one could detect or place strong upper limits on the presence of MACHOs in the dark matter halo of the Milky Way. Such a program was demanding but feasible thanks to the advent of digital detectors (charge-coupled devices, or CCDs) and computer processing. In the early 1990s, three teams began to search for microlensing: the MACHO Project (e.g., Alcock et al. 2000), the EROS Experiment (Expérience pour la Recherche d'Objets Sombres; e.g., Tisserand et al. 2007), and OGLE (Optical Gravitational Lensing Experiment; e.g., Wyrzykowski et al. 2011a,b).

The teams sought to identify stars with variable brightness due to microlensing, and not other types of variability, such as variable stars and supernovae in background galaxies.[4] They also developed criteria for selecting events consistent with expectations for microlensing, e.g., events consistent with the Paczyński light curve, and which look the same in red and blue light. In 5.7 years, the MACHO Project identified 13–17 microlensing events in observing 11.9 million stars. Making detailed simulations of their detection efficiency, they found that the optical depth for events with time scales between 2 and 400 days is $\tau = 1.2^{+0.4}_{-0.3} \times 10^{-7}$ (Alcock et al. 2000). EROS and OGLE found the smaller values $\tau < 0.36 \times 10^{-7}$ and $\tau = (0.16 \pm 0.12) \times 10^{-7}$, respectively (Tisserand et al. 2007; Wyrzykowski et al. 2011a). The observed optical depth and event rate are much smaller than what was predicted for a dark matter halo made entirely of MACHOs, so the results place strong upper limits on the fraction of dark matter in the form of MACHOs (at least in the mass range from $\sim 10^{-7}\ M_\odot$ to a few tens of $M_\odot$).

The MACHO and EROS programs were completed, but OGLE continues to monitor about 300 million stars in 51 fields (Udalski et al. 2015a). They detect around 2000 microlensing events per year.[5]

5.5.2 *Finding Planets*

As discussed in Sect. 5.2, planets orbiting a lens star can create detectable features in the light curve. Microlensing therefore offers a unique way to find extrasolar planets, which is sensitive to a different combination of planet parameters (mass and orbit) than traditional methods.[6] In particular, the mass of a planet mainly affects the *duration* of the perturbations to the Paczyński light curve, not the *amplitude*.

[3] The Magellanic Clouds are small galaxies that orbit the Milky Way and are visible in the southern hemisphere.

[4] In fact, microlensing campaigns have yielded large catalogs of variable stars (see, e.g., Alcock et al. 1996; Soszyński et al. 2016).

[5] http://ogle.astrouw.edu.pl/ogle4/ews/ews.html.

[6] The "transit method" detects a planet when it passes between us and its host star, causing the star's light to dim briefly. The "radial velocity method" detects a star moving because a planet pulls on it.

Microlensing can detect low-mass planets if the monitoring is frequent enough to catch perturbations that last a few hours to a few days. Initial searches for microlensing planets relied on a two-step system: survey teams looked at large numbers of stars every few days, and once they detected a new event, they sent a community alert so that follow-up teams could begin more intensive monitoring. Next-generation searches combine a large field of view with a rapid observing cadence to enable both discovery and follow-up in the same data set.

Gaudi (2012) gives a nice discussion of both the methods and results of microlensing searches for planets. Here we can only scratch the surface and mention a few particularly interesting discoveries. Gaudi et al. (2008) discovered two planets with masses $M_{p1} \sim 0.71 M_J$ and $M_{p2} \sim 0.27 M_J$, where $M_J = 9.5 \times 10^{-4} M_\odot$ is the mass of Jupiter. These planets orbit a host star with mass $M_s \sim 0.50 M_\odot$. The masses, orbital sizes, and implied temperatures make the planets analogues of Jupiter and Saturn. At a time when other methods were finding planets quite unlike any in our own solar system, this discovery indicated that analogues of our solar system do exist and can be found with microlensing. In a different system, Batista et al. (2011) observed complicated caustic-crossing features in the light curve. To explain the features, they had to consider finite source effects, orbital motion of the planet around the star, and Earth's motion around the Sun. To handle degeneracies in the lens models, Batista et al. (2011) did a Bayesian analysis using a model of the Milky Way and argued that the planet ($M_p \sim 2.6 M_J$) is unusually massive relative to the star ($M_s \sim 0.19 M_\odot$). For a third system, Udalski et al. (2015b) took advantage of the parallax effect (Sect. 5.1.2) and used simultaneous observations from the ground and the infrared Spitzer Space Telescope to break degeneracies and learn more about a planetary microlensing event. It is quite striking to see how a seemingly small change in the observing position can lead to a large change in the shape and timing of the microlensing light curve.

Based on the success so far, microlensing searches for planets have become a principal component of NASA's proposed Wide Field Infrared Survey Telescopes (WFIRST).

5.5.3 *Characterizing Compact Objects*

Microlensing can also be used to search for compact objects that remain after stars die: white dwarfs, neutron stars, and black holes. These objects—especially black holes—can be more massive than the typical stars that cause microlensing, so they lead to long-duration events that are affected by Earth's orbital motion. Wyrzykowski et al. (2016) searched the OGLE database for microlensing events that exhibited parallax effects from Earth's orbital motion and found 13 candidates of which the most massive is $9.3 M_\odot$ at a distance of 2.4 kpc. Such searches offer a unique opportunity to find isolated black holes (whereas searches that use X-rays or gravitational waves can only find black holes in binary systems).

All of the events discussed so far have been found by watching source stars and hoping that an unseen lens happens to pass in the foreground. In rare circumstances it is possible to invert the process—to find lens stars that will move in front of source stars—and thus to predict microlensing events. Sahu et al. (2017) noted that the white dwarf Stein 2051 B had a proper motion that would carry it in front of a star from late 2013 through early 2014. They measured the shift in the apparent position of the background star (the impact parameter was too large to yield a detectable change in magnification) and used it to measure the white dwarf's mass. McGill et al. (2018) have predicted a similar event for the white dwarf LAWD 37 in November 2019. Measuring white dwarf masses with microlensing provides important constraints on models of white dwarfs, which involves the physics of degenerate matter as well as white dwarf atmospheres.

Problems

5.1 Figure 5.1 shows how the magnification changes as a function of time in a microlensing event. Make an analogous plot to show how the centroid changes with time in astrometric microlensing. Hint: recall Eq. (2.33).

5.2 Consider the lens potential

$$\psi(x, y) = \ln\left(\sqrt{x^2 + y^2}\right) + \frac{\kappa}{2}\left(x^2 + y^2\right) + \frac{\gamma_+}{2}\left(x^2 - y^2\right) + \gamma_\times xy\,. \tag{5.54}$$

In the case $\kappa = 0$, this potential models a point mass in the presence of external shear, resulting from massive objects far from the origin. For $\kappa \neq 0$, the model consists of a lensed macro-image perturbed by a point mass. In this case, the single lensed macroimage is split by the point mass into a number of microimages. This is called the **Chang-Refsdal lens** (Chang and Refsdal 1979). The mass is chosen so that the coefficient of the first term in Eq. (5.54) is 1.

(a) Write the complex deflection function the complex image position (see Sect 5.2.1) and complex shear (see Sect 7.2.1).
(b) What is the maximum number of images that this model can produce?
(c) Find the critical curves and caustics for different values of κ and γ_+. For each set of values of κ and γ_+, plot light curves that illustrate the behavior of this lens.

In the context of quasar lensing, the relative motion of observer, lens, and source cannot be detected. However, such relative motion could be observed if we picture the Chang-Refsdal model as describing a stellar microlensing event.

5.3 Consider an equal-mass binary lens (cf. Sect. 5.2.2) in which the separation Δ is small enough that there is a single critical curve and caustic, as in the middle row of Fig. 5.2.

(a) Solve the lens equation for a source at the origin to show that there are five images in a symmetric configuration.
(b) Solve the lens equation for a source on the separation axis, and show that a source inside the caustic produces five images, while a source outside the caustic produces three images. (Hint: the top row of Fig. 5.4 might be helpful.)
(c) Show that when the source moves far from the caustics, there is one image close to the source whose magnification is close to unity and one faint image close to each point mass.

5.4 In Sect. 5.2.1 we noted that a system of $n > 1$ point masses in a plane can create as many as $5n - 5$ images. Bleher et al. (2014) explicitly construct lens configurations that achieve this limit. For $n \geq 4$, let $n - 1$ of the lenses have Einstein radius $\theta_E = 1$, and distribute them equally around a circle of radius $a = (n-1)^{-1/n} \left(\frac{n-1}{n}\right)^{1/2}$, and let the remaining lens have a small mass, and place it at the origin. For $n = 2$ or 3, place all the masses on the circle and omit the mass at the origin. Use the computational techniques discussed in Appendix E to construct the critical curves and caustics, and verify that the configuration can produce $5n - 5$ images. Hint: recall that a source far from the lenses produces $n + 1$ images (Petters et al. 2001, Theorem 11.7).

5.5 In Sect. 5.5.2 we noted that planets can create features in microlensing light curves that last from a few hours to a few days. Beaulieu et al. (2006) reported the discovery of a planet of 5.5 Earth masses. Making reasonable assumptions, estimate the duration of the planetary perturbations, and compare your answer with the light curve shown in Fig. 1 of the paper by Beaulieu et al. (2006).

5.6 In Sect. 5.4 we computed the microlensing optical depth assuming a singular isothermal sphere mass distribution. Repeat the analysis for a nonsingular isothermal sphere, assuming a source at the Galactic Center. As the NIS core radius goes to zero, do your results match the SIS analysis?

5.7 Our analysis has assumed that all of the observed light comes from the source star. In practice, observations often target crowded fields and have resolution that is limited by atmospheric blurring and/or telescope optics. In that case, light from the lens star and/or other stars along the line of sight may be *blended* with the light from the source star. Let the blend fraction, $f_{\rm bl}$, be the fraction of light that comes from stars other than the source star long before or after a microlensing event. Make a new version of Fig. 5.1 that includes different values for $f_{\rm bl}$ as well as different values of $u_{\rm min}$. Given a well-measured light curve, should it be easy to measure both $f_{\rm bl}$ and $u_{\rm min}$, or are they likely to be degenerate?

References

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., Bennett, D. P., et al. (2000). The MACHO Project: Microlensing results from 5.7 years of large magellanic cloud observations. *Astrophysical Journal, 542*, 281.

Alcock, C., Allsman, R. A., Axelrod, T. S., Bennett, D. P., Cook, K. H., Freeman, K. C., et al. (1996). The MACHO Project LMC variable star inventory. II. LMC RR Lyrae Stars- Pulsational characteristics and indications of a global youth of the LMC. *Astronomical Journal, 111*, 1146.

Batista, V., Gould, A., Dieters, S., Dong, S., Bond, I., Beaulieu, J. P., et al. (2011). MOA-2009-BLG-387Lb: A massive planet orbiting an M dwarf. *Astronomy & Astrophysics, 529*, A102.

Beaulieu, J.-P., Bennett, D. P., Fouqué, P., Williams, A., Dominik, M., Jørgensen, U. G., et al. (2006). Discovery of a cool planet of 5.5 Earth masses through gravitational microlensing. *Nature, 439*, 437.

Bleher, P. M., Homma, Y., Ji, L. L., & Roeder, R. K. W. (2014). Counting zeros of harmonic rational functions and its application to gravitational lensing. *International Mathematics Research Notices, 2014*, 2245.

Chang, K., & Refsdal, S. (1979). Flux variations of QSO 0957+561 A, B and image splitting by stars near the light path. *Nature, 282*, 561.

Dodelson, S. (2017). *Gravitational lensing*. Cambridge: Cambridge University Press.

Gaudi, B. S. (2012). Microlensing surveys for exoplanets. *Annual Review of Astronomy and Astrophysics, 50*, 411.

Gaudi, B. S., Bennett, D. P., Udalski, A., Gould, A., Christie, G. W., Maoz, D., et al. (2008). Discovery of a Jupiter/Saturn analog with gravitational microlensing. *Science, 319*, 927.

Khavinson, D., & Neumann, G. (2006). On the number of zeros of certain rational harmonic functions. *Proceedings of the American Mathematical Society, 134*, 1077.

McGill, P., Smith, L. C., Evans, N. W., Belokurov, V., & Smart, R. L. (2018). A predicted astrometric microlensing event by a nearby white dwarf. *Monthly Notices of the Royal Astronomical Society, 478*, L29.

Mollerach, S., & Roulet, E. (2002). *Gravitational lensing and microlensing*. Singapore: World Scientific.

Paczyński, B. (1986). Gravitational microlensing by the galactic halo. *Astrophysical Journal, 304*, 1.

Petters, A. O., Levine, H., & Wambsganss, J. (2001). *Singularity theory and gravitational lensing*. Boston: Birkhäuser.

Rhie, S. H. (2003). n-Point gravitational lenses with $5(n-1)$ images. astro-ph/0305166.

Sahu, K. C., Anderson, J., Casertano, S., Bond, H. E., Bergeron, P., Nelan, E. P., et al. (2017). Relativistic deflection of background starlight measures the mass of a nearby white dwarf star. *Science, 356*, 1046.

Soszyński, I., Pawlak, M., Pietrukowicz, P., Udalski, A., Szymański, M. K., Wyrzykowski, Ł., et al. (2016). The OGLE collection of variable stars. Over 450 000 eclipsing and ellipsoidal binary systems toward the galactic bulge. *Acta Astronomica, 66*, 405.

Tisserand, P., Le Guillou, L., Afonso, C., Albert, J. N., Andersen, J., Ansari, R., et al. (2007). Limits on the Macho content of the galactic halo from the EROS-2 survey of the magellanic clouds. *Astronomy & Astrophysics, 469*, 387.

Udalski, A., Szymański, M. K., & Szymański, G. (2015a). OGLE-IV: Fourth phase of the optical gravitational lensing experiment. *Acta Astronomica, 65*, 1.

Udalski, A., Yee, J. C., Gould, A., Carey, S., Zhu, W., Skowron, J., et al. (2015b). *Spitzer* as a microlens parallax satellite: Mass measurement for the OGLE-2014-BLG-0124L planet and its host star. *Astrophysical Journal, 799*, 237.

Witt, H. J., & Mao, S. (1994). Can lensed stars be regarded as pointlike for microlensing by MACHOs? *Astrophysical Journal, 430*, 505.

Wyrzykowski, Ł., Kostrzewa-Rutkowska, Z., Skowron, J., Rybicki, K. A., Mróz, P., Kozłowski, S., et al. (2016). Black hole, neutron star and white dwarf candidates from microlensing with OGLE-III. *Monthly Notices of the Royal Astronomical Society, 458*, 3012.

Wyrzykowski, Ł., Kozłowski, S., Skowron, J., Udalski, A., Szymański, M. K., Kubiak, M., et al. (2011a). The OGLE view of microlensing towards the magellanic clouds - III. Ruling out subsolar MACHOs with the OGLE-III LMC data. *Monthly Notices of the Royal Astronomical Society, 413*, 493.

Wyrzykowski, L., Skowron, J., Kozłowski, S., Udalski, A., Szymański, M. K., Kubiak, M., et al. (2011b). The OGLE view of microlensing towards the magellanic clouds - IV. OGLE-III SMC data and final conclusions on MACHOs. *Monthly Notices of the Royal Astronomical Society, 416*, 2949.

Chapter 6
Strong Lensing by Galaxies

The remainder of this book is concerned with applications of gravitational lensing of cosmological relevance. We proceed by considering ever larger mass and distance scales, starting with galaxy-scale strong lensing in the present chapter. The background sources in this context are point-like quasars or extended galaxies. The axisymmetric models of Chap. 2 can be readily adapted to the more general setting of Chap. 4. Only such non-axisymmetric mass distributions can be expected to describe observed lens systems.

The isothermal sphere of Sect. 2.3, in both its singular and nonsingular forms, offers a good starting point for modeling galaxy lenses, but it suffers from some notable deficiencies. The most basic problem is that the image configurations of observed lenses do not have the collinear appearance that axisymmetric models require. We consider two non-axisymmetric lens models in Sect. 6.1, which reduce to the axisymmetric case for appropriate parameter values. Another issue is that the density of a physical object must be finite everywhere, in contrast to the singular isothermal sphere of Sect. 2.3.2. We saw in the case of an isothermal model that it is straightforward to introduce a core of finite density that leaves the qualitative behavior of singular density profiles largely intact. Section 6.2 extends the discussion in Sect. 2.3.3 to the present context.

Even all of these refinements to the models and methods of Chap. 2 are insufficient for understanding the full variety of observed galaxy lenses. For example, we must account for extended sources (Sect. 6.3) and for lenses that cannot be described by smooth density profiles (Sect. 6.4).

Determining the properties of lensed images for a given source position only gets us so far. What is of interest in practice is to infer the mass distribution that gives rise to observed images. We illustrate this approach in Sect. 6.5, first considering special cases in which the lens parameters can be found analytically, before resorting to statistical methods that the lensing researcher must inevitably confront. Applications of the models and methods described in this chapter are outlined in Sect. 6.6.

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_6

6.1 Singular Isothermal Lens Models

The lens models presented in Chap. 2 illustrate how multiple images can be produced, as well as the dependence of image number on the position of the source. In the axisymmetric case, it is just the angular distance of the source from the lens center that is important; the azimuthal coordinate, ϕ, can be defined so that $\phi = 0$ and $\phi = \pi$ are the only allowed values. Upon introducing asymmetry, however, parameterizing the caustics and critical curves requires both the radial and azimuthal coordinates. In this section, we introduce two causes of asymmetry frequently encountered in observed lenses. Firstly, real galaxies are not circular in projection, owing either to intrinsic ellipticity in the case of early-type galaxies or inclination in the case of late-type galaxies. Secondly, galaxies are often found within groups or clusters, so that the lens system consists of a primary deflector and one or more perturbing objects.

6.1.1 Spherical Lenses with External Shear

We begin by considering a singular isothermal sphere (SIS) that is perturbed by other deflectors. If the perturbers are sufficiently far from the primary lens, we can use a Taylor series expansion and characterize their effects through convergence and shear (as in Sect. 4.1.3). The additional terms in the lens potential are often referred to as **external convergence** and **external shear** because they are produced by objects outside of the main lens galaxy. External convergence is subject to the mass-sheet degeneracy (cf. Sect. 4.7.2), but external shear has distinct effects because it breaks axisymmetry.

In Sect. 2.3.2, we found that the magnitude of the deflection angle for an SIS is constant. This no longer holds in the absence of circular symmetry. Moreover, the tangential caustic, which is a single point in the axisymmetric case, becomes a closed curve, allowing for the possibility of four images for certain source positions (see Fig. 6.1). Recall that the *non*-singular isothermal sphere (NIS), discussed in Sect. 2.3.3, gives rise to a more complicated form for the deflection angle. Adding shear to that model complicates matters still further! For this reason, we confine quantitative analysis to the singular case and offer a conceptual treatment of the NIS in Sect. 6.2.

Without loss of generality, we choose coordinates aligned with the direction of shear ($\phi_\gamma = 0$). Then the lens potential associated with the external shear can be written as

$$\psi_\gamma(x, y) = -\frac{\gamma}{2}(x^2 - y^2)\,, \tag{6.1}$$

where x and y are coordinates in the *lens* plane, and the components of shear are $(\gamma_+, \gamma_\times) \equiv (-\gamma, 0)$, with $\gamma > 0$. The reason we choose $\gamma_+ < 0$ is explained in Sect. 6.1.2. The potential of the SIS takes the form

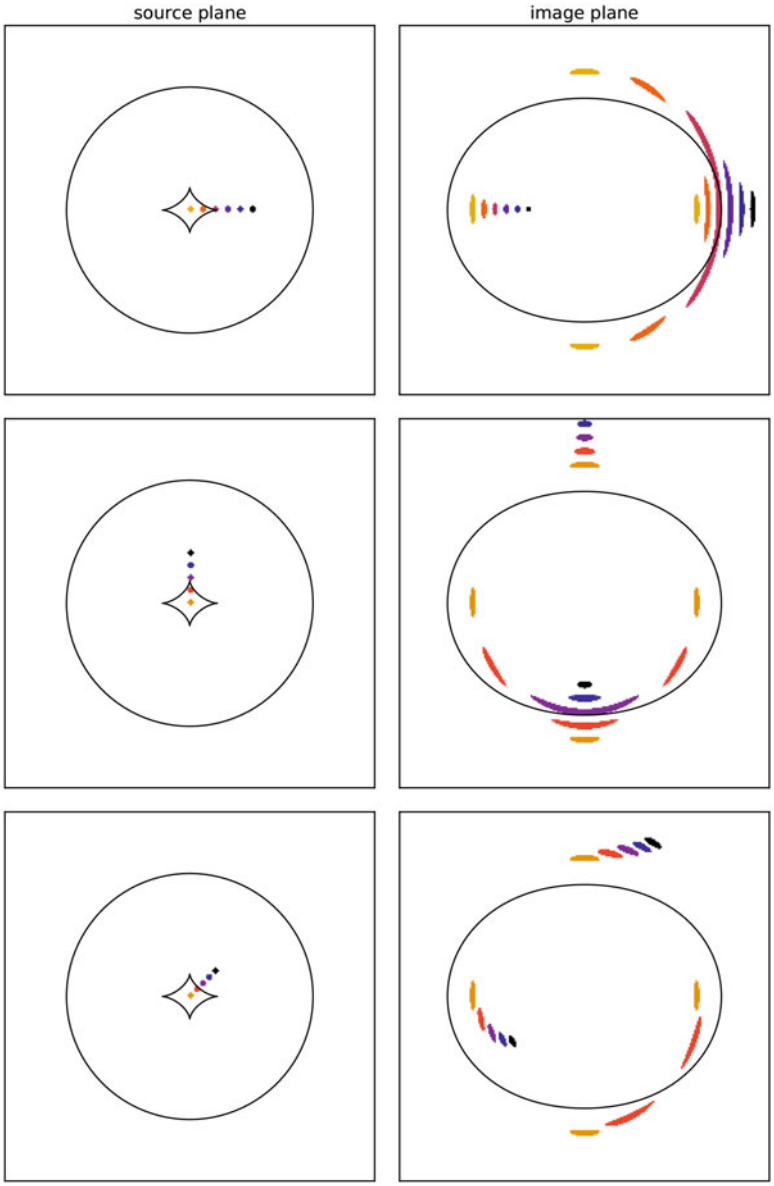


Fig. 6.1 Sample image configurations for an SIS with external shear. In each row, the left panel shows a set of sources along with the tangential caustic and radial pseudocaustic, while the right panel shows the corresponding images along with the tangential critical curve

$$\psi_{\rm SIS}(x,\,y) = b\sqrt{x^2+y^2}\,, \tag{6.2}$$

which comes from integrating Eq. (2.47) over the coordinate θ. The potential of the SIS with external shear is the sum $\psi = \psi_{\rm SIS} + \psi_\gamma$.

The lens equation then reads

$$u = x - \frac{b\,x}{\sqrt{x^2+y^2}} + \gamma\,x \tag{6.3a}$$

$$v = y - \frac{b\,y}{\sqrt{x^2+y^2}} - \gamma\,y\,, \tag{6.3b}$$

where b is the Einstein radius of the SIS *without* shear and $(u,\,v)$ are coordinates in the *source* plane. For a source at the origin, we find four images with positions

$$(x,\,y) = \left(0, \pm\frac{b}{1-\gamma}\right) \quad \text{and} \quad (x,\,y) = \left(\pm\frac{b}{1+\gamma}, 0\right)\,. \tag{6.4}$$

We see that the breaking of circular symmetry leads to four distinct images instead of an Einstein ring. Images resembling Einstein rings can be formed if the source is spatially extended.

Let us now consider a source away from the origin. For the sake of analytic tractability, we assume that the source lies on one of the coordinate axes in the source plane. When the source is on the positive u-axis,[1] we find the following solutions:

$$(x,\,y) = \left(\frac{u+b}{1+\gamma}, 0\right) \qquad \text{for } 0 \le u < \infty \tag{6.5a}$$

$$(x,\,y) = \left(\frac{u-b}{1+\gamma}, 0\right) \qquad \text{for } 0 \le u < u_r \tag{6.5b}$$

$$(x,\,y) = \left(\frac{u}{2\gamma}, \pm\sqrt{\frac{b^2}{(1-\gamma)^2} - \frac{u^2}{4\gamma^2}}\right) \qquad \text{for } 0 \le u < u_t, \tag{6.5c}$$

where $u_r \equiv b$ and $u_t \equiv 2\gamma b/(1-\gamma)$. The meaning of the subscripts will become apparent. For $\gamma < 1/3$, we have $u_t < u_r$, and we see that one image is produced if $u_r < u < \infty$, two images if $u_t < u < u_r$, and four images if $0 \le u < u_t$. For $\gamma > 1/3$, we have $u_r < u_t$, so one image is produced if $u_t < u < \infty$, three images if $u_r < u < u_t$, and four images if $0 \le u < u_r$.

[1] For a source on the negative u-axis, we let $(x,\,y) \to (-x,\,y)$ in Eq. (6.5a) while leaving the sign of u unchanged.

Since the number of images changes by two as the source crosses the value $u = u_t$, and by one as it crosses $u = u_r$, we suspect that u_t and u_r mark the points where the tangential and radial caustics, respectively, intersect the positive u-axis. (By symmetry, the caustics intersect the negative u-axis at $-u_t$ and $-u_r$.) When $u = u_t$, three of the images converge to $\boldsymbol{\theta}_t \equiv \begin{bmatrix} x_t & y_t \end{bmatrix}^T = \begin{bmatrix} b/(1-\gamma) & 0 \end{bmatrix}^T$. When $u = u_r$, one of the images is at $\boldsymbol{\theta}_r \equiv \begin{bmatrix} x_r & y_r \end{bmatrix}^T = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$. To confirm that $\boldsymbol{\beta}_t \equiv \begin{bmatrix} u_t & 0 \end{bmatrix}^T$ and $\boldsymbol{\beta}_r \equiv \begin{bmatrix} u_r & 0 \end{bmatrix}^T$ are caustic points, we must show that $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_r$ lie on critical curves. To do this, we compute the magnification at $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_r$. The inverse amplification tensor for an SIS with shear at an arbitrary point $\boldsymbol{\theta} = \begin{bmatrix} x & y \end{bmatrix}^T$ is given by

$$\mathsf{A}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} 1 - \frac{by^2}{(x^2+y^2)^{3/2}} + \gamma & \frac{bxy}{(x^2+y^2)^{3/2}} \\ \frac{bxy}{(x^2+y^2)^{3/2}} & 1 - \frac{bx^2}{(x^2+y^2)^{3/2}} - \gamma \end{bmatrix} \tag{6.6}$$

$$= \begin{bmatrix} 1 - \frac{b\sin^2\phi}{R} + \gamma & \frac{b\sin\phi\cos\phi}{R} \\ \frac{b\sin\phi\cos\phi}{R} & 1 - \frac{b\cos^2\phi}{R} - \gamma \end{bmatrix}. \tag{6.7}$$

The corresponding inverse magnification is

$$\mu^{-1}(R, \phi) = 1 - \gamma^2 - \frac{b}{R}\left[1 + \gamma\cos(2\phi)\right]. \tag{6.8}$$

In terms of the polar coordinates (R, ϕ), both $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_r$ have $\phi = 0$, while their respective radial coordinates are $R = |x_t| = b/(1-\gamma)$ and $R = |x_r| = 0$. Thus, the magnification diverges ($\mu^{-1} = 0$) for $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ and vanishes ($\mu^{-1} \to -\infty$) for $\boldsymbol{\theta} = \boldsymbol{\theta}_r$. Note that the magnification will be zero whenever $\mu^{-1} \to \pm\infty$. The sign corresponds to the parity of the image. Finally, we conclude that $\boldsymbol{\beta}_t$ is a cusp of the tangential caustic and $\boldsymbol{\beta}_r$ lies on the radial pseudocaustic.

We now turn to the case of a source on the positive v-axis.[2] We find images at

$$(x, y) = \left(0, \frac{v+b}{1-\gamma}\right) \qquad \text{for } 0 \le v < \infty \tag{6.9a}$$

$$(x, y) = \left(0, \frac{v-b}{1-\gamma}\right) \qquad \text{for } 0 \le v < v_r \tag{6.9b}$$

$$(x, y) = \left(\pm\sqrt{\frac{b^2}{(1+\gamma)^2} - \frac{v^2}{4\gamma^2}}, -\frac{v}{2\gamma}\right) \qquad \text{for } 0 \le v < v_t, \tag{6.9c}$$

[2]For a source on the negative v-axis, we let $(x, y) \to (x, -y)$ in Eq. (6.9a) while leaving the sign of v unchanged.

where tangential caustics intersect the positive v-axis at $v_r \equiv b$ and $v_t \equiv 2\gamma b/(1+\gamma)$, respectively. We see from the above solutions for (x, y) that for $\gamma < 1$, one image is produced if $v_r < v < \infty$, two images if $v_t < v < v_r$, and four images if $0 \leq v < v_t$. Shear amplitudes exceeding unity would indicate that terms in the perturbing potential beyond the quadrupole term cannot be neglected (i.e., the shear approximation breaks down). Even $\gamma > 1/3$ pushes the envelope of realistic shear values (Holder and Schechter 2003). Note that a source on the v-axis must be closer to the origin than a source on the u-axis to produce four images. For reasonable shear values, the case of three images does not arise.

For a source that is not on a symmetry axis, we must solve the lens equation numerically (see Appendix E). Nevertheless, our analytic approach provides useful information. In particular, we can derive parametric equations for the caustics. The first step is to express the critical curve in the form $R = R(\phi)$. We then obtain the corresponding caustic by writing the lens equation (6.3a) in the parametric form

$$u(\phi) = [(1+\gamma)R(\phi) - b]\cos\phi \tag{6.10a}$$

$$v(\phi) = [(1-\gamma)R(\phi) - b]\sin\phi\,. \tag{6.10b}$$

Recall that the tangential critical curve is the locus of points (R, ϕ) for which $\mu \to \infty$. From Eq.(6.8), we find the following relation between R and ϕ:

$$R_t(\phi) = \theta_E \left[\frac{1+\gamma\cos(2\phi)}{1-\gamma^2}\right]. \tag{6.11}$$

Substituting this expression into Eq. (6.10a) gives

$$u_t(\phi) = \frac{2\gamma b}{1-\gamma}\cos^3\phi \tag{6.12a}$$

$$v_t(\phi) = -\frac{2\gamma b}{1+\gamma}\sin^3\phi\,. \tag{6.12b}$$

This equation describes a curve known as an **astroid** with semimajor and semiminor axes $a_1 = 2\gamma b/(1-\gamma)$ and $a_2 = 2\gamma b/(1+\gamma)$, respectively. Meanwhile, the radial critical curve reduces to the single point $R_r(\phi) = 0$, so that Eq. (6.10a) implies a circular radial caustic of radius b. If a source is far outside both caustics, then $R \gg b$, and Eq. (6.8) yields $\mu \to (1-\gamma^2)^{-1}$. This is just the magnification due to a constant shear field with vanishing convergence.

The tangential and radial caustics are shown in Fig. 6.2 for several values of the shear. In the absence of shear, the tangential caustic reduces to a point, and a source at that point appears as an Einstein ring. As the shear increases, the tangential caustic expands horizontally. For sufficiently large shear, two cusps poke through the radial caustic, thus becoming "naked." Without solving the lens equation, we deduce the number of images for a given source position, as discussed in the figure caption.

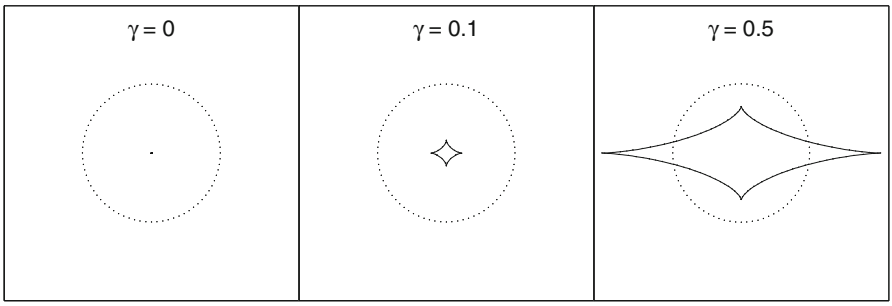


Fig. 6.2 Caustics for a singular isothermal sphere with shear along the horizontal axis. From left to right, the panels show cases with no shear ($\gamma = 0$), moderate shear ($\gamma = 0.1$), and large shear ($\gamma = 0.5$). The tangential and radial caustics are indicated by solid and dotted curves, respectively. Typically the tangential caustic is enclosed by the radial caustic (center panel). In that case, a source outside the radial caustic produces one image, a source between the radial and tangential caustics produces two images, and a source within both caustics produces four images. If $\gamma > 1/3$ (e.g., right panel), the source may lie outside the radial caustic but inside the tangential caustic. Such naked cusps give rise to three images

6.1.2 *Elliptical Lenses*

Moving beyond axisymmetry, it is natural to consider elliptical lenses. If we assume that the *density* has elliptical symmetry, we obtain the singular isothermal elliptical mass distribution (SIEMD; Kassiola and Kovner 1993; Kormann et al. 1994; Keeton and Kochanek 1998). In coordinates aligned with the major axis of the ellipse, the convergence can be written as

$$\kappa_{\rm SIEMD} = \frac{b}{2\sqrt{q^2x^2 + y^2}} = \frac{b}{2R\sqrt{q^2\cos^2\phi + \sin^2\phi}}, \tag{6.13}$$

where q is the axis ratio of the isodensity contours ($0 < q \leq 1$) and b is the Einstein radius of the corresponding SIS ($q = 1$). In order to obtain the lens equation, we must solve Poisson's equation for the lens potential and take its gradient to obtain the deflection. Problem 6.4 gives the components of the deflection vector, which allows one to study the allows one to study the magnification, critical curves, caustics, and image configurations. The SIEMD and its generalization to nonsingular models can be used to study elliptical galaxies and spiral galaxies with arbitrary inclination with respect to the line of sight.

Alternatively, if we assume that the *potential* has elliptical symmetry, we obtain the singular isothermal elliptical potential model (SIEP; Kassiola and Kovner 1993):

$$\psi_{\rm SIEP}(x, y) = b\sqrt{q^2x^2 + y^2} \tag{6.14a}$$

$$\psi_{\rm SIEP}(R,\phi) = bR\sqrt{q^2\cos^2\phi + \sin^2\phi}\,, \tag{6.14b}$$

where q is now the axis ratio of the isopotential contours. Since $\nabla^2\psi_{\rm SIEP} = 2\kappa_{\rm SIEP}$, the convergence is given in terms of polar coordinates by

$$\kappa_{\rm SIEP}(R,\phi) = \frac{q^2 b}{2R\left(q^2\cos^2\phi + \sin^2\phi\right)^{3/2}}\,. \tag{6.15}$$

Note that both the SIEMD and SIEP models reduce to the SIS when $q = 1$. While the convergence of the SIEP has a different angular dependence than that of the SIEMD, the SIEP is easier to work with and has qualitatively similar properties, for axis ratios close to unity. We therefore focus on the SIEP in this section, but it is important to bear in mind that it is the SIEMD that is used in lens modeling (see Problem 6.3). The lens equation for the SIEP has the form

$$u = x - \frac{q^2 b\, x}{\sqrt{q^2x^2 + y^2}} \tag{6.16a}$$

$$v = y - \frac{b\, y}{\sqrt{q^2x^2 + y^2}}\,. \tag{6.16b}$$

We can solve these equations analytically if the source lies along one of the coordinate axes in the source plane.

For a source on the positive u-axis,[3] we find the solutions

$$(x,y) = (u + q\,b, 0) \qquad \text{for } 0 \le u < \infty \tag{6.17a}$$

$$(x,y) = (u - q\,b, 0) \qquad \text{for } 0 \le u < u_r \tag{6.17b}$$

$$(x,y) = \left(\frac{u}{1-q^2}, \pm\sqrt{b^2 - \frac{q^2u^2}{(1-q^2)^2}}\right) \qquad \text{for } 0 \le u < u_t, \tag{6.17c}$$

where $u_r \equiv q\,b$ and $u_t \equiv (1-q^2)b/q$. We see from these solutions that for $q > 1/\sqrt{2}$, one image is produced if $u_r < u < \infty$, two images if $u_t < u < u_r$, and four images if $0 \le u < u_t$. When $q < 1/\sqrt{2}$, we find that $u_r < u_t$, and naked cusps are produced.

For a source on the positive v-axis,[4] we find

$$(x,y) = (0, v + b) \qquad \text{for } 0 \le v < \infty \tag{6.18a}$$

[3] Same as Footnote 1, but for Eq. (6.17a).

[4] Same as Footnote 2, but for Eq. (6.18a).

$$(x, y) = (0, v - b) \qquad \text{for } 0 \le v < v_r \tag{6.18b}$$

$$(x, y) = \left(\pm q\sqrt{b^2 - \frac{v^2}{(1-q^2)^2}}\ ,\ -\frac{q^2 v}{1-q^2} \right) \qquad \text{for } 0 \le v < v_t, \tag{6.18c}$$

where $v_r \equiv b$ and $v_t \equiv (1-q^2)b$. We see that for all allowed values q, one image is produced if $v_r < v < \infty$, two images if $v_t < v < u_r$, and four images if $0 \le v < v_t$. Notice that naked cusps cannot form on the v-axis.

To compute the magnification, we find it convenient to write the lens potential and convergence in the generalized isothermal form

$$\Psi(R, \phi) = RF(\phi) \tag{6.19a}$$

$$\kappa(R, \phi) = \frac{G(\phi)}{2R}\,, \tag{6.19b}$$

where (R, ϕ) are polar coordinates. From Poisson's equation, we find that

$$G(\phi) = F(\phi) + F''(\phi)\,. \tag{6.20}$$

For an elliptical potential, the angular part of the potential takes the form

$$F(\phi) = b\sqrt{q^2 \cos^2\phi + \sin^2\phi}\,. \tag{6.21}$$

The angular part of the convergence is then

$$G(\phi) = \frac{bq^2}{\left(q^2\cos^2\phi + \sin^2\phi\right)^{3/2}}\,. \tag{6.22}$$

In terms of $G(\phi)$, the inverse magnification is given by

$$\mu^{-1} = 1 - \frac{G(\phi)}{R}\,. \tag{6.23}$$

The tangential critical curve can be parametrized as $R_t(\phi) = G(\phi)$, from which it follows that the tangential caustic has the parameterization

$$u_t(\phi) = \frac{bq^2(1-q^2)\cos^3\phi}{\left(q^2\cos^2\phi + \sin^2\phi\right)^{3/2}} \tag{6.24a}$$

$$v_t(\phi) = -\frac{b(1-q^2)\sin^3\phi}{\left(q^2\cos^2\phi + \sin^2\phi\right)^{3/2}}\,. \tag{6.24b}$$

Like the SIS with shear, the SIEP and all singular isothermal models for that matter have a radial critical point at the origin: $R_r(\phi) = 0$. The radial caustic for the SIEP can be parametrized as

$$u_r(\phi) = -\frac{q^2 b \cos\phi}{\sqrt{q^2\cos^2\phi + \sin^2\phi}} \tag{6.25a}$$

$$v_r(\phi) = -\frac{b \sin\phi}{\sqrt{q^2\cos^2\phi + \sin^2\phi}}\,. \tag{6.25b}$$

When $q = 1$, both caustics reduce to those for an SIS. If q differs only slightly from unity, the tangential caustic takes the form of an astroid and the radial caustic that of a circle. Recall that the SIS with shear gives rise to caustics of the same form. The caustic structure of the SIEMD lens is addressed in Problem 6.4.

To conclude our discussion of the SIEP without shear, we wish to show that the caustic parameterizations of Eqs. (6.24a) and (6.25a) reduce to the particular values u_t, v_t, u_r, and v_r following Eqs. (6.17a) and (6.18a). Recall that we took the source coordinates to be positive in the on-axis cases. One may be tempted simply to evaluate the expressions (6.24a) and (6.25a) at $\phi = 0$, but that would cause u_r and v_r to be negative. We must instead evaluate Eqs. (6.25a) at $\phi = \pi$. This is because a source that approaches u_r or v_r from below corresponds to images on the opposite side of the lens from the source.

In realistic lens systems, we must include the effects of both intrinsic or apparent ellipticity in the lens galaxy and external shear from perturbing bodies. Figure 6.3 shows caustics for a SIEMD with shear. We vary the ellipticity, $e \equiv 1 - q$, from top to bottom, and the *modified* shear angle, $\varphi_\gamma \equiv \phi_\gamma - \pi/2$, from left to right. We fix the shear amplitude at $\gamma = 0.1$. The shear angle appearing in Eqs. (4.35) is defined so that $\phi_\gamma = 0$ for an *image* that is stretched horizontally while $\varphi_\gamma = 0$ for a *caustic* that is stretched horizontally, that is, in direction of the major axis of the galaxy. For a fixed value of φ_γ, increasing the ellipticity enlarges the tangential caustic, similar to what we found when increasing γ for an SIS with shear. For $e = 0.9$, which could be appropriate for a nearly edge-on disk (see Keeton and Kochanek 1998), naked cusps are produced for all three values of φ_γ. Meanwhile, the aspect ratio of the radial caustic decreases with increasing ellipticity, in agreement with our analysis of the SIEP. For an SIEP, the height of the radial caustic is constant, in contrast to its behavior seen in Fig. 6.3. This is not surprising, since the SIEP and SIEMD make different predictions when the ellipticity is large.

We now consider the effects of shear angle on the structure of the caustics, again assuming an amplitude of $\gamma = 0.1$. When $\varphi_\gamma = 0$, we find a small naked-cusp region when $e = 0.5$. In the absence of shear, naked cusps do not form for this value of e. In other words, the shear conspires with the ellipticity to create a larger tangential caustic than would be possible with ellipticity alone. When $\varphi_\gamma = \pi/4$, the tangential caustic undergoes a counterclockwise rotation, its size decreases, and

its shape is altered. Nevertheless, it continues to resemble an astroid, in the sense that it is composed of four cusps connected by fourfolds (Schneider et al. 1992; Petters et al. 2001). We see that the reduction in size of the tangential caustic for $\varphi_\gamma = \pi/4$ removes the naked cusps we found when $\varphi_\gamma = 0$ and $e = 0.5$. When the shear and ellipticity are perpendicular ($\varphi_\gamma = \pi/2$), the shear and ellipticity partially cancel out, leading to a substantial decrease in the size of the tangential caustic.

Although the values of ellipticity and shear angle we have considered do not represent an exhaustive sampling of the (e, γ) plane, the nine parameter combinations whose caustics are shown in Fig. 6.3 provide a fair representation of caustic topologies expected for realistic lens populations. We note, however, that it is possible, given an appropriate combination of shear and ellipticity, to generate

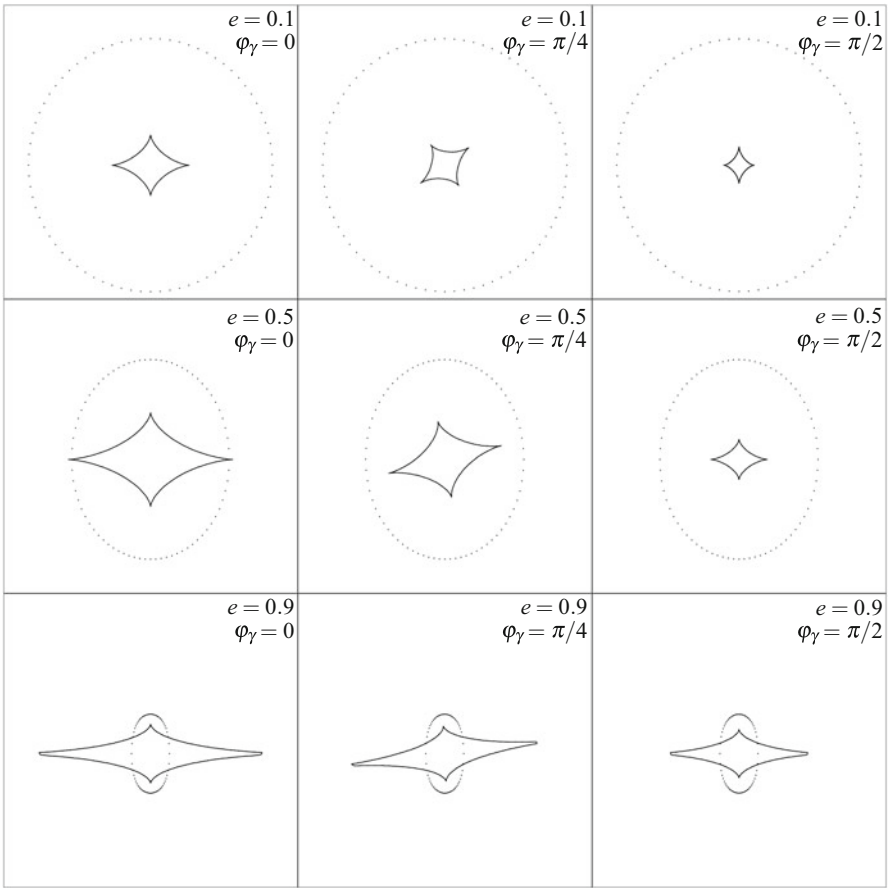


Fig. 6.3 Caustics for a SIEMD with shear for various values of ellipticity e and modified shear angle φ_γ. The shear amplitude is fixed at $\gamma = 0.1$

caustics for which six, or even eight, lensed images can be formed (Keeton et al. 2000b). Such exotic caustics are discussed in more detail by Schneider et al. (1992) and Petters et al. (2001).

6.2 Lenses with a Core of Finite Density

An important property of singular isothermal models is that the central density diverges, which does not occur in nature. One simple way to remedy this is to replace the radial coordinate $\varrho_q \equiv \sqrt{q^2x^2 + y^2}$, where $0 < q \leq 1$, by $\varrho_q^* \equiv \sqrt{q^2(x^2 + s^2) + y^2}$ for some $s > 0$. This gives rise to a radial caustic with a critical curve that is no longer a single point. The image number changes by two as the source crosses this radial caustic rather than changing by one, as in the case of a pseudocaustic. Since a source that is far from the lens is singly imaged, a lens with finite central density always creates an odd number of images. If the central core is small ($s \ll b$), the additional image we find for a source inside the radial caustic is too faint to observe. This is because it is close to the center of the lens galaxy and the brightness of the galaxy obscures the image. As it happens, even a lens that is singular at its center behaves like a nonsingular lens as long as its convergence $\kappa \propto R^{-\eta}$ is "shallower" than isothermal, i.e., $0 < \eta < 1$.

6.2.1 Emergence of Radial Caustics

Recall that a caustic forms whenever the magnification diverges. In the axisymmetric case, this occurs only for a source at the origin. For a point mass, a radial caustic is not produced. A singular isothermal lens gives rise to a radial pseudocaustic, where the magnification vanishes. To determine when a radial caustic forms, we consider a power law lens with potential

$$\psi(R) = \frac{b^2}{2-\eta}\left(\frac{R}{b}\right)^{\eta-2} . \tag{6.26}$$

The inverse magnification works out to be

$$\mu^{-1}(R) = \left[1 - \left(\frac{b}{R}\right)^{\eta}\right]\left[1 - (1-\eta)\left(\frac{b}{R}\right)^{\eta}\right] . \tag{6.27}$$

The first factor vanishes when $R = b$. This critical curve maps to a source at the origin. The second factor vanishes when

$$R = (1-\eta)^{\frac{1}{\eta}}\, b\,. \tag{6.28}$$

This is the equation for the radial critical curve we have been seeking. The corresponding caustic is described by

$$|\beta| = \eta(1-\eta)^{(1-\eta)/\eta}\, b\,, \tag{6.29}$$

i.e., a circle of radius $|\beta|$. It turns out that β itself is negative which means that images merging on the radial critical curve are on the opposite side of the lens from the source.

6.2.2 *Central Images*

When the density profile is shallow enough to admit a radial caustic, a source interior to that caustic will be observed as an odd number of images (cf. Sect. 4.4), one of which will be faint and close to the center of the lens. It is not generally possible to ascertain the position and magnification of this "central" image analytically, so we offer a qualitative discussion here. We can use the presence or absence of a detectable central image to constrain the central slope of a singular lens or the core radius of a nonsingular lens. If a central image is not observed, we can place an upper limit on its brightness, which corresponds to a lower limit on the central slope or core radius of the lens. If a central image is observed, we can use its brightness to determine an actual value of the slope or core radius (see Problem 6.8).

In practice, using a central image to constrain the mass distribution of a lens galaxy requires observation at wavelengths where the lens galaxy itself is not bright. This is because we expect central images to be much fainter than the cores of galaxies at optical wavelengths. The absence of central images in radio observations of the lenses B0739+366 and B1030+074 allowed Rusin and Ma (2001) to constrain the central power law slopes of the surface mass densities to be no smaller than $\eta = 0.85$ and $\eta = 0.91$, respectively. Turning the problem around, Keeton (2003) used stellar distributions in nearby elliptical galaxies to determine the expected range of central power law density profile slopes, from which the magnification of the central image can be inferred. While magnifications as large as 0.1 are possible, typical values are closer to 0.001, making detection extremely difficult even at radio wavelengths.

Since the centers of most if not all galaxies harbor supermassive black holes, it is important to understand how such objects alter the caustic structure of lens galaxies. Mao et al. (2001) showed that central images cannot form in nonsingular isothermal lenses that contain black holes more massive than a certain threshold. Even for masses below this value, the central image may become unobservably faint. Alternatively, a second central image may form, as Rusin et al. (2005) demonstrated for more general lens models. In practice, the case of two central images is indistinguishable from that of one central image. The important point is that central images are demagnified by black holes. Thus a given magnification is degenerate to some extent with respect to black hole mass and inner density profile.

To date, a central image has been observed only in the lensed quasar PMN J1632−0033. Winn et al. (2004), who discovered this image at radio wavelengths, found a logarithmic slope of $\eta = 0.91$ in the surface mass density profile. Their model included a black hole of $\sim 10^8 M_\odot$. Varying this mass did not noticeably change the value of η. Rusin and Ma (2001) found that increasing the black hole mass led to a mild decrease in the profile slope. Both groups concluded that suppression of the central image by a massive black hole is required in order for theory and observation to agree. We note that a fifth image was also detected in the quasar lens SDSS J1004+4112 (Inada et al. 2005), but the lensing object in that case is a galaxy cluster. Since clusters are known to have shallow inner density profiles, the detection of a central image is not so surprising, though it is still useful for modeling purposes (see Sect. 7.1).

6.2.3 *Caustic Metamorphosis*

Referring to Eq. (2.57), we find that the magnitude of the deflection angle at a distance R from the center of a nonsingular isothermal sphere is given by

$$\alpha_{\rm NIS}(R) = \frac{b}{R}\left(\sqrt{R^2+s^2} - s\right) . \tag{6.30}$$

The lens equation for an NIS with external shear is then

$$u = x - \left(\frac{\sqrt{x^2+y^2+s^2} - s}{x^2+y^2}\right) b\, x + \gamma\, x \tag{6.31a}$$

$$v = y - \left(\frac{\sqrt{x^2+y^2+s^2} - s}{x^2+y^2}\right) b\, y - \gamma\, y\,, \tag{6.31b}$$

which reduces to Eq. (6.3a) in the limit $s \to 0$. Rather than writing down the amplification tensor, we compute the magnification directly from the deflection vector, $\boldsymbol{\alpha}$, the partial derivatives of which are

$$\begin{aligned}
\frac{\partial \alpha_x}{\partial x} &= \frac{\partial}{\partial x}(\alpha_{\rm NIS}\cos\phi - \gamma\, x) \\
&= \frac{\partial R}{\partial x}\frac{{\rm d}\alpha_{\rm NIS}}{{\rm d}R}\cos\phi - \frac{\partial \phi}{\partial x}\alpha_{\rm NIS}\sin\phi - \gamma \\
&= \left(\frac{b}{\sqrt{R^2+s^2}} - \frac{\alpha_{\rm NIS}}{R}\right)\cos^2\phi + \frac{\alpha_{\rm NIS}}{R}\sin^2\phi - \gamma \\
&= \frac{b\cos^2\phi}{\sqrt{R^2+s^2}} - \frac{\alpha_{\rm NIS}}{R}\cos(2\phi) - \gamma
\end{aligned} \tag{6.32a}$$

$$\begin{aligned}\frac{\partial\alpha_x}{\partial y} &= \frac{\partial\alpha_y}{\partial x} = \frac{\partial}{\partial y}(\alpha_{\rm NIS}\cos\phi - \gamma\, x)\\ &= \frac{\partial R}{\partial y}\frac{{\rm d}\alpha_{\rm NIS}}{{\rm d}R}\cos\phi - \frac{\partial\phi}{\partial y}\alpha_{\rm NIS}\sin\phi\\ &= \left(\frac{b}{\sqrt{R^2+s^2}} - 2\frac{\alpha_{\rm NIS}}{R}\right)\sin\phi\cos\phi\\ &= \left(\frac{b}{2\sqrt{R^2+s^2}} - \frac{\alpha_{\rm NIS}}{R}\right)\sin(2\phi) \qquad (6.32b)\end{aligned}$$

$$\begin{aligned}\frac{\partial\alpha_y}{\partial y} &= \frac{\partial}{\partial y}(\alpha_{\rm NIS}\sin\phi + \gamma\, y)\\ &= \frac{\partial R}{\partial y}\frac{{\rm d}\alpha_{\rm NIS}}{{\rm d}R}\sin\phi + \frac{\partial\phi}{\partial y}\alpha_{\rm NIS}\cos\phi + \gamma\\ &= \left(\frac{b}{\sqrt{R^2+s^2}} - \frac{\alpha_{\rm NIS}}{R}\right)\sin^2\phi + \frac{\alpha_{\rm NIS}}{R}\cos^2\phi + \gamma\\ &= \frac{b\sin^2\phi}{\sqrt{R^2+s^2}} + \frac{\alpha_{\rm NIS}}{R}\cos(2\phi) + \gamma\,. \qquad (6.32c)\end{aligned}$$

The corresponding convergence is

$$\kappa = \frac{1}{2}\left(\frac{\partial\alpha_x}{\partial x} + \frac{\partial\alpha_y}{\partial y}\right) = \frac{b}{2\sqrt{R^2+s^2}}\,. \qquad (6.33)$$

The components of the *total* shear are

$$\Gamma_+ = \frac{1}{2}\left(\frac{\partial\alpha_x}{\partial x} - \frac{\partial\alpha_y}{\partial y}\right) = \left(\kappa - \frac{\alpha_{\rm NIS}}{R}\right)\cos(2\phi) - \gamma \qquad (6.34a)$$

$$\Gamma_\times = \frac{\partial\alpha_x}{\partial y} = \left(\kappa - \frac{\alpha_{\rm NIS}}{R}\right)\sin(2\phi)\,, \qquad (6.34b)$$

so that the shear amplitude is

$$\begin{aligned}\Gamma &= \sqrt{\Gamma_+^2 + \Gamma_\times^2}\\ &= \sqrt{\left(\kappa - \frac{\alpha_{\rm NIS}}{R}\right)^2 - 2\gamma\left(\kappa - \frac{\alpha_{\rm NIS}}{R}\right)\cos(2\phi) + \gamma^2}\,. \qquad (6.35)\end{aligned}$$

Our goal is to determine the maximum number of images for a given set of parameter values. The required information is contained in the signs of the eigenvalues of the amplification tensor or equivalently its inverse. The inverse

magnification is given in terms of the relevant eigenvalues as $\mu^{-1} = \lambda_+\lambda_-$, where $\lambda_\pm = 1 - \kappa \mp \Gamma$. Recall that both eigenvalues are positive when the source is sufficiently far from the lens. Each time a source crosses a caustic on its way toward the center of the lens, a pair of images is created on a critical curve, and one of the eigenvalues changes sign. Therefore, the signs of the eigenvalues at the origin tell us how many images are produced. If both λ_+ and λ_- are positive, no caustics will form, and the total number of images will be one; if either λ_+ or λ_- is negative, there will be one caustic, and the total number of images will be three. If both λ_+ and λ_- are negative, there will be two caustics, and the total number of images will be five.

In order to compute $\lambda_\pm$ at the origin, we begin by setting $\phi = 0$. The shear amplitude then reduces to

$$\Gamma = \kappa - \frac{\alpha_{\rm NIS}}{R} - \gamma\,. \tag{6.36}$$

This implies that

$$\lambda_+ = 1 - 2\kappa + \frac{\alpha_{\rm NIS}}{R} + \gamma \tag{6.37a}$$

$$\lambda_- = 1 - \frac{\alpha_{\rm NIS}}{R} - \gamma\,. \tag{6.37b}$$

Letting $R \to 0$, we obtain

$$\lambda_\pm^{(0)} = 1 - \frac{b}{2s} \pm \gamma\,. \tag{6.38}$$

These eigenvalues vanish when

$$s_\pm = \frac{b}{2(1 \pm \gamma)}\,. \tag{6.39}$$

Note that $0 < s_+ < s_-$. If $s < s_+$, then both eigenvalues are negative at the origin; if $s_+ < s < s_-$, then $\lambda_+ > 0$ and $\lambda_- < 0$ at the origin; and if $s > s_-$, then both eigenvalues are positive at the origin. In other words, as an NIS with shear becomes less concentrated, its lensing power diminishes. For sufficiently large s, multiple images do not form.

Figure 6.4 shows critical curves and caustics for various core radii. We assume $b = 1$ and $\gamma = 0.1$, in which case $s_+ = 5/11 \approx 0.455$ and $s_- = 5/9 \approx 0.556$. We see that as s increases, the inner critical curve increases until it touches the outer critical curve (see Problem 6.4), and then it shrinks and disappears at $s = s_+$. As s increases further, the remaining critical curve shrinks until it disappears at $s = s_-$.

Note that this lens can produce multiple images for any $s < s_- = b/(2(1-\gamma))$. Compare this to what we found for an NIS without shear in Sect. 2.3.3, which can have multiple images only for $s < b/2$. This difference is explained by the presence

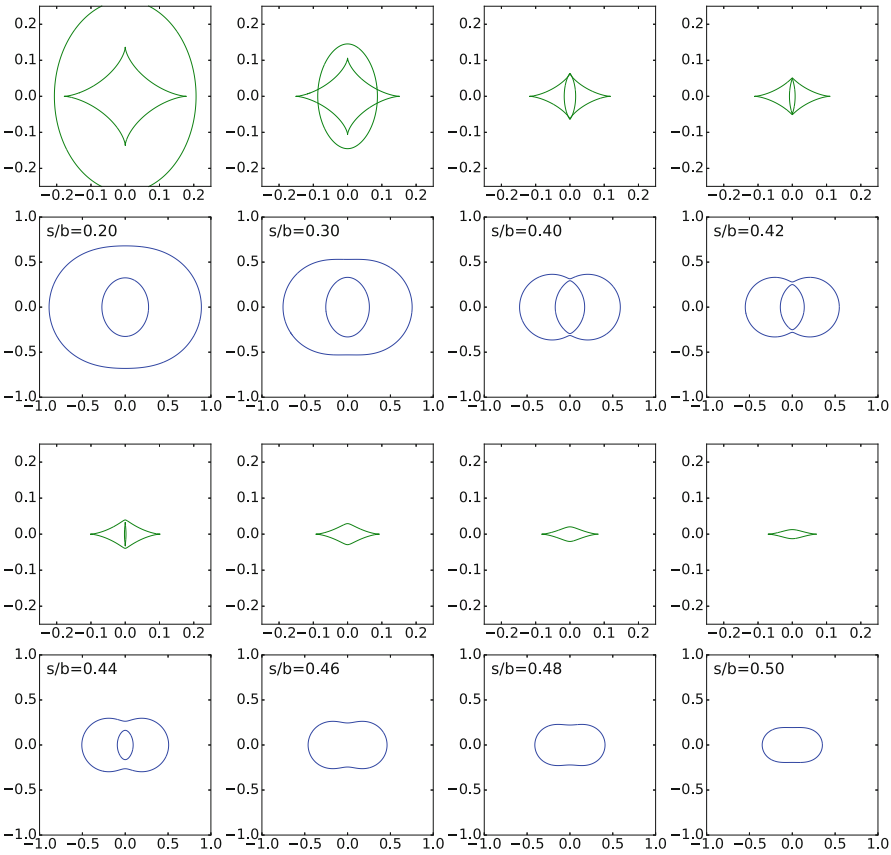


Fig. 6.4 Illustration of caustic metamorphosis for an NIS with Einstein radius b in the presence of external shear with magnitude $\gamma = 0.1$ and orientation angle $\phi_\gamma = \frac{\pi}{2}$. In each row, the top panels show the caustics (green) for different values of the core radius, while the bottom panels show the corresponding critical curves (blue)

of external shear. For a circular lens, the condition for strong lensing is $\kappa > 1$. In the general case, on the other hand, the condition is $\lambda_- < 0$, i.e., $\kappa + \gamma > 1$.

Varying s alters not only the structure of the caustics but also their relative sizes. In particular, Fig. 6.4 shows that naked cusps emerge between $s = 0.2b$ and $s = 0.3b$. In the limit $s \to 0$, a shear of $\gamma = 0.1$ is insufficient to produce naked cusps. For a nonsingular lens, as s increases, the radial caustic shrinks more than the tangential caustic, whence the naked cusps. A similar effect is observed for a singular lens as the central density profile becomes shallower. This argument holds whether a lens is noncircular due to external shear or intrinsic ellipticity. Indeed, Oguri and Keeton (2004) found that naked cusps can form for systems with larger axis ratios if the inner density profile is shallower than isothermal. One such lens is the system SDSS J1029+2623 (Oguri et al. 2008).

6.3 Ring Images of Extended Sources

We have seen that noncircular lenses cause a point source to appear as one or more discrete images. In particular, Einstein rings do not form, even if the source and lens are perfectly aligned. For an extended source, each point within it is lensed into discrete images. If none of the source lies on a caustic, each of its points has the same number of images. Consequently, the source will appear as a number of extended images. If, on the other hand, any point of the source lies on the tangential caustic, two or three of its images lie on the tangential critical curve. If the source covers the entire tangential caustic, its images cover the critical curve and are highly magnified. However, a ring image need not coincide with the critical curve, because the magnification of an extended source involves a convolution integral. The advantage of ring images is that they offer more observational constraints than the two or four images typically seen in quasar lenses. Note that the term "Einstein ring" is often used in the context of extended sources, whether or not the lens is axisymmetric.

Suppose the lens surface brightness can be described by a continuous function $I(\boldsymbol{\theta})$. Kochanek et al. (2001) pointed out that it is useful to define a "ring curve" as the set of points where the surface brightness is maximized in the radial direction; this does not characterize the radial thickness of the ring, but it does describe the angular structure. Sample ring curves are shown in Fig. 6.5. The point on the ring curve at azimuth ϕ is found by solving

$$\nabla_{\boldsymbol{\theta}} I(\boldsymbol{\theta}) \cdot \mathbf{e}_R(\phi) = [\nabla_{\boldsymbol{\theta}} I(\boldsymbol{\theta})]^T \, \mathbf{e}_R(\phi) = 0 \,. \tag{6.40}$$

Here $\mathbf{e}_R(\phi)$ is the outward unit vector in the radial direction, and $\boldsymbol{\theta} = \begin{bmatrix} R\cos\phi & R\sin\phi \end{bmatrix}^T$. For elliptical sources, it proves more convenient to work with the *source*-plane gradient

$$\nabla_{\boldsymbol{\beta}} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} \frac{\partial}{\partial \boldsymbol{\theta}} = \mathsf{A} \nabla_{\boldsymbol{\theta}} \,, \tag{6.41}$$

where $\mathsf{A} = \mathsf{A}(\boldsymbol{\theta})$ is the amplification tensor. Using this expression and conservation of surface brightness, Eq. (6.40) can be recast as

$$\left[\mathsf{A}^{-1} \nabla_{\boldsymbol{\beta}} I^{(s)}(\boldsymbol{\beta}) \right]^T \mathbf{e}_R(\phi) = 0 \,, \tag{6.42}$$

where $I^{(s)}(\boldsymbol{\beta})$ is the surface brightness of the source. If the source has elliptical symmetry, we can write

$$I^{(s)}(\boldsymbol{\beta}) = J^{(s)}(\eta) \quad \text{where} \quad \eta = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mathsf{S} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \,, \tag{6.43}$$

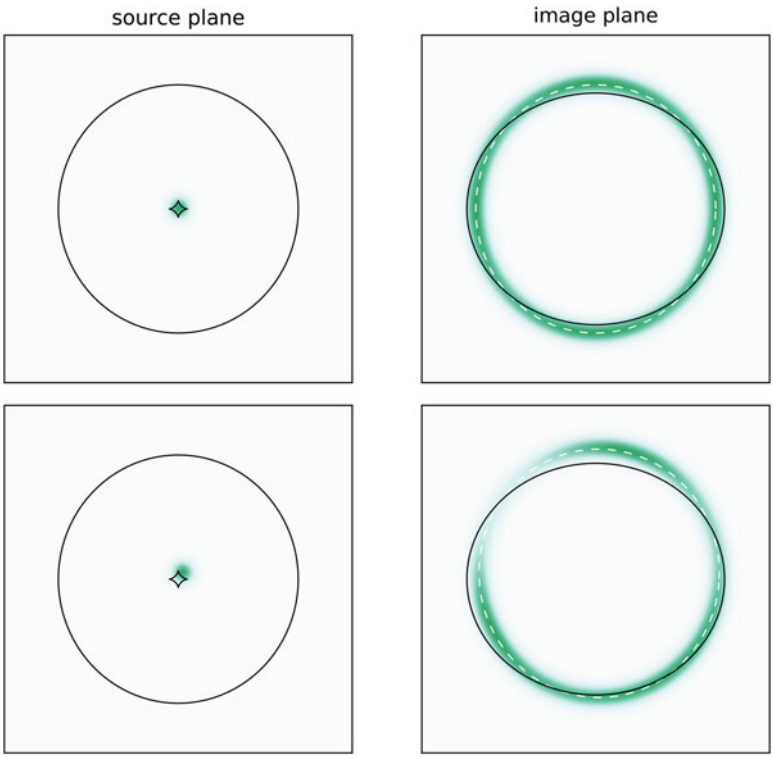


Fig. 6.5 Sample Einstein rings produced by an isothermal elliptical mass distribution. In each row, the left panel shows the extended source along with the tangential caustic and radial pseudocaustic, while the right panel shows the corresponding ring image along with the tangential critical curve. The "ring curve" is shown with a dashed line. The top row has a source at the origin, while the bottom row has the source offset slightly

for a source centered at $\boldsymbol{\beta}_0$. The tensor

$$\mathsf{S} = \begin{bmatrix} \cos^2\varphi_s + q_s^{-2}\sin^2\varphi_s & (1-q_s^{-2})\cos\varphi_s\sin\varphi_s \\ (1-q_s^{-2})\cos\varphi_s\sin\varphi_s & q_s^{-2}\cos^2\varphi_s + \sin^2\varphi_s \end{bmatrix} \tag{6.44}$$

describes an ellipse with axis ratio $0 < q_s \leq 1$ whose major axis makes the angle $0 \leq \phi_s < 2\pi$ with the horizontal. Equation (6.42) can now be written as

$$\begin{aligned} 0 &= 2\frac{\partial \boldsymbol{J}^{(s)}}{\partial \eta}\left[\mathsf{A}^{-1}\mathsf{S}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]^T \mathbf{e}_R \\ &= 2\frac{\partial \boldsymbol{J}^{(s)}}{\partial \eta}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathsf{S}\mathsf{A}^{-1}\mathbf{e}_R\,. \end{aligned} \tag{6.45}$$

If we assume that the source brightness profile $J^{(s)}(\eta)$ is strictly decreasing, the derivative never vanishes, so it can be factored out. Thus the ring curve depends only on the angular structure of the source, not on its brightness profile, and it can be found as the solution of

$$[\boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta}) - \boldsymbol{\beta}_0]^T \, \mathsf{S}\mathsf{A}^{-1}\mathbf{e}_R = 0\,, \tag{6.46}$$

where we used the lens equation to replace $\boldsymbol{\beta}$ with $\boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})$.

As a specific example, consider the generalized isothermal lens defined by Eqs. (6.19a). It turns out that $\mathbf{e}_R$ is an eigenvector of A^{-1} with eigenvalue 1,[5] so $\mathsf{A}^{-1}\mathbf{e}_R = \mathbf{e}_R$. If the source is circular, then S is the identity matrix. With these simplifications, Eq. (6.46) has the solution

$$R_{\rm ring}(\phi) = F(\phi) + \boldsymbol{\beta}_0^T \mathbf{e}_R(\phi)\,. \tag{6.47}$$

Note that the ring curve depends on the angular structure of the potential (6.19a), whereas the critical curve depends on that of the convergence (6.19b): $R_{\rm crit}(\phi) = G(\phi)$. Thus, the ring and critical curves generally have different shapes (see Fig. 6.5). See Kochanek et al. (2001) for discussion of more general geometries.

6.4 Perturbations Due to Small-Scale Structure

So far we have supposed that the lens galaxy has a smooth mass distribution described by a simple function. In reality, however, the mass can be concentrated on a variety of scales: galaxies contain tens of billions of individual stars; they are orbited by up to a few dozen small, satellite galaxies; and simulations of galaxy formation suggest that the visible satellites are a subset of large population of "subhalos," or gravitationally bound clumps of dark matter that orbit within the galaxy's overall dark matter halo. While both stars and subhalos may be small compared to the overall lens galaxy, they may nevertheless be important for lensing, especially if their Einstein radii are comparable to or larger than the angular size of the source. In this section we briefly discuss the main qualitative effects of small-scale structure.

6.4.1 Millilensing by Dark Matter Substructure

First consider a galaxy that contains $N_{\rm sub}$ subhalos, where the number could be a few dozen to a few hundred. For lensing purposes we do not necessarily need to

[5] The isothermal model does not have any radial magnification.

distinguish between visible satellite galaxies and dark matter subhalos. We can use the principle of superposition to write the lens potential as a sum of contributions from the large-scale, smooth component of the lens and the many subhalos:

$$\psi(\boldsymbol{\theta}) = \psi_{\rm smooth}(\boldsymbol{\theta}) + \sum_{i=1}^{N_{\rm sub}} \psi_i(\boldsymbol{\theta} - \boldsymbol{\theta}_i)\,, \tag{6.48}$$

where $\psi_{\rm smooth}$ is the lens potential for the smooth mass distribution, while ψ_i is the lens potential for subhalo i, which is centered at $\boldsymbol{\theta}_i$. Different models have been used for subhalos (see, e.g., Nierenberg et al. 2014), but one simple model is an isothermal sphere whose outer parts have been stripped away by tidal interactions with the large galaxy.[6] One way to construct a tidally truncated isothermal model is to write the 3-D density as

$$\rho(r) = \frac{\sigma^2 a^2}{2\pi G(s^2 + r^2)(a^2 + r^2)}\,. \tag{6.49}$$

As with the nonsingular isothermal sphere discussed in Sect. 2.3.3, σ is the velocity dispersion, and s is the core radius. The new parameter a is referred to as the truncation radius: the extra factor of $a^2/(a^2 + r^2)$ has little effect for $r \ll a$ but serves to suppress the density beyond $r \sim a$, such that the total mass is $M = \pi\sigma^2 a^2/(G(s + a))$. The projected surface mass density can be written as

$$\Sigma(R) = \frac{\sigma^2}{2\pi G(1 - s^2/a^2)}\left(\frac{1}{\sqrt{s^2 + R^2}} - \frac{1}{\sqrt{a^2 + R^2}}\right)\,. \tag{6.50}$$

This has the form of the difference between two nonsingular isothermal spheres, one with core radius s and the other with core radius a.[7]

Figure 6.6 illustrates how subhalos affect lensing by showing magnification maps (i.e., plotting $|\mu|$ as a function of position $\boldsymbol{\theta}$). The left panel shows a smooth SIEMD, while the middle and right panels show models in which 2% of the total mass of the galaxy has been placed in truncated isothermal spheres. The overall structure of the magnification map, including the location of the critical curve, does not change very much. However, the subhalos create small perturbations on scales comparable to their Einstein radii. A point image that forms near a subhalo can have its position, magnification, and time delay all perturbed (relative to the original smooth model), while an arc or ring image can have its shape distorted. The spatial scale for such perturbations is typically a few to tens of milliarcseconds, so the phenomenon is often referred to as "millilensing." Its implications are discussed in Sect. 6.6.

[6] Another model for subhalos is the NFW model that we will encounter in Sect. 7.1.1.

[7] In some literature, especially for cluster lensing, this model is referred to as a pseudo-isothermal elliptical mass distribution, or PIEMD, after Kassiola and Kovner (1993). It has also been called a pseudo-Jaffe model (Keeton et al. 2000a).

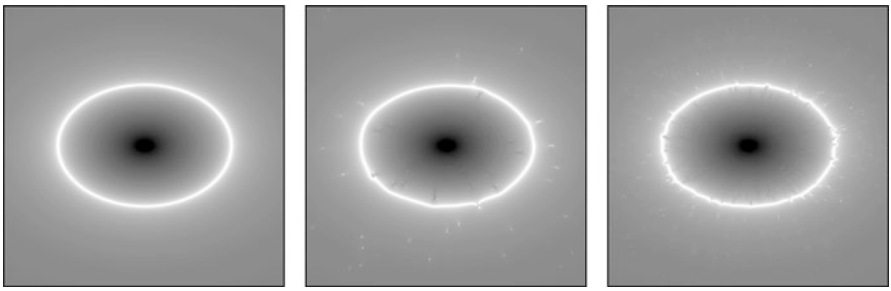


Fig. 6.6 Examples of magnification maps in the image plane for a smooth lens (left) and one with substructure (middle and right). The smooth component here is an SIEMD. In the middle panel, 2% of the mass has been replaced by dark matter subhalos. In this example all of the subhalos have the same mass, but in reality they would have a range of masses. In the right panel, the subhalos are 10× less massive but there are 10× as many, so the total mass in subhalos remains the same

6.4.2 Microlensing by Stars

Now consider a galaxy that contains $N_{\rm star}$ stars. Formally the lens potential can again be written using the superposition principle:

$$\psi(\boldsymbol{\theta}) = \psi_{\rm smooth}(\boldsymbol{\theta}) + \sum_{i=1}^{N_{\rm star}} \theta_{E,i}^2 \ln|\boldsymbol{\theta} - \boldsymbol{\theta}_i| \ , \tag{6.51}$$

where $\theta_{E,i}$ is the Einstein radius for star i. In practice this expression is difficult to use because it contains billions of terms. It is more useful to "zoom in" and examine a local region where the number of stars is more modest. Formally, we write

$$\psi(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}) = \psi_{\rm smooth}(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}) + \sum_{i=1}^{N_{\rm star}} \theta_{E,i}^2 \ln|\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta} - \boldsymbol{\theta}_i| \ , \tag{6.52}$$

and then make two approximations. First, we make a Taylor series expansion of $\psi_{\rm smooth}$ and keep only the lowest order terms, which correspond to constant convergence and shear (cf. Sect. 4.1.3). Second, we limit the sum to stars in some "local" region where $\boldsymbol{\theta}_i - \boldsymbol{\theta}_0$ is comparable to $\Delta\boldsymbol{\theta}$.[8] Thus, we approximate the lens potential as[9]

[8] This local region needs to be defined with some care because the deflection can be non-negligible even at many times the star's Einstein radius (see, e.g., Wambsganss 1999).

[9] This is a generalization of the Chang-Refsdal lens mentioned in Problem 5.2 to multiple stars.

$$\psi(x, y) \approx \frac{1}{2}\kappa_{\rm smooth}(x^2 + y^2) + \frac{1}{2}\gamma_{\rm smooth}(x^2 - y^2) + \sum_{i=1}^{N_{\rm loc}} \theta_{E,i}^2 \ln\sqrt{(x - x_i)^2 + (y - y_i)^2}\,, \tag{6.53}$$

where (x, y) are the components of $\Delta\boldsymbol{\theta}$ in a frame aligned with the direction of the shear and (x_i, y_i) is the position of star i in this frame. Also, the number of stars in this local region is $N_{\rm loc}$. They contribute convergence $\kappa_{\rm stars} = \sum_{i=1}^{N_{\rm loc}} \pi\theta_{E,i}^2/A_{\rm loc}$ where $A_{\rm loc}$ is the area of this local region.

It is important to understand the scales involved. Lensed images typically appear about an arcsecond from the center of a lens galaxy. The Einstein radii of the individual stars are of order a few microarcseconds, so the local regions we need to consider are typically a few tens of microarcseconds across (and centered on each of the main images). Within each region, there can be a complicated caustic network as shown in Fig. 6.7. Since the observed magnification is a convolution of the source surface brightness with the lensing magnification (cf. Sect. 4.6.2), large sources effectively average over the caustic network and experience a magnification that is consistent with a smooth lens model, while sources whose angular size is microarcseconds or smaller can experience magnifications that are very different from what a smooth lens model would predict. The stars and source typically move relative to one another at speeds that correspond to a fraction of a microarcsecond per year, so the magnification of a small source can change over the course of months and years. Section 6.6 discusses how this "microlensing" can be used to study accretion disks around supermassive black holes in distant galaxies.

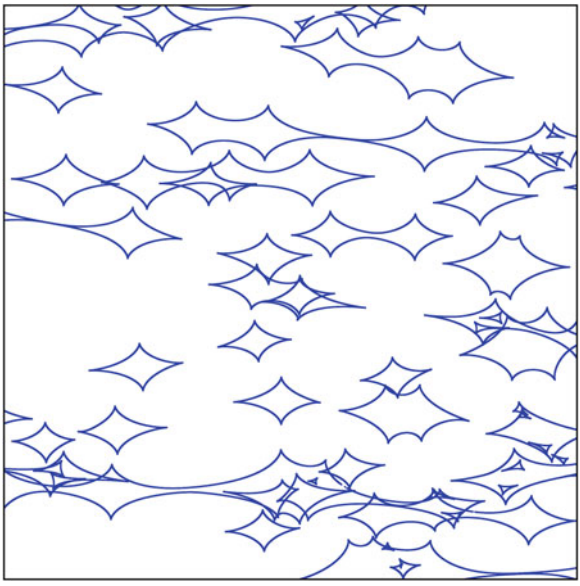


Fig. 6.7 Example of a microlensing caustic network. This figure shows a local region that is 20 stellar Einstein radii on a side. The total convergence is $\kappa_{\rm tot} = 0.4$, which is divided into a smooth piece with $\kappa_{\rm smooth} = 0.3$ and a contribution $\kappa_{\rm stars} = 0.1$ from stars. The shear is $\gamma_{\rm smooth} = 0.4$, and the coordinates are chosen to be aligned with the shear

6.5 Lens Modeling

To this point we have focused on understanding the properties of images produced by a given lens. In many applications, the images are observed, and we need to find the mass distribution that produced them. Galaxies are complicated objects, so we should not expect to recover every last detail of their mass distributions, but we can hope to make a model that captures the key features. In this section we consider how to solve for parameters of a lens model given a set of lensed images, focusing on smooth models (see Sect. 6.6 for discussion of models that include small-scale structure). We examine lenses with point images in detail and discuss some key points about handling lenses with extended images in Sect. 6.5.3.

6.5.1 Analytic Determination of Lens Parameters

It is useful to start with examples that can be solved by hand. An axisymmetric lens and a non-axisymmetric lens in a "cross" configuration both have symmetries that simplify the analysis.

Einstein Radius for an Axisymmetric Lens

From the lens equation $\beta = \theta - \theta_E \,\mathrm{sgn}(\theta)$ for an SIS lens, we see that the image positions are $\theta_\pm = \beta \pm \theta_E$, where the image at θ_- only exists for $\beta < \theta_E$. In the case of two images, it is straightforward to solve for the two lens parameters:

$$\beta = \frac{1}{2}\left(\theta_+ + \theta_-\right) , \tag{6.54a}$$

$$\theta_E = \frac{1}{2}\left(\theta_+ - \theta_-\right) . \tag{6.54b}$$

For a point mass, on the other hand, $\beta = \theta - \theta_E^2/\theta$, which implies the positions

$$\theta_\pm = \frac{1}{2}\left(\beta \pm \sqrt{\beta^2 + 4\theta_E^2}\right) . \tag{6.55}$$

In terms of these positions, the lens parameters are (recall that θ_+ and θ_- have opposite signs)

$$\beta = \theta_+ + \theta_- , \tag{6.56a}$$

$$\theta_E = \sqrt{-\theta_+\,\theta_-} . \tag{6.56b}$$

Notice that the lens parameters depend on our choice of model. If we define $d_\pm \equiv |\theta_\pm|$, we see that the Einstein radius of an SIS lens is the arithmetic mean of d_+ and d_-, while for a point mass it is the geometric mean of these distances. If the image configuration is roughly symmetric about the lens ($d_+ \approx d_-$), the Einstein radii inferred from the SIS and point mass models will be similar.

Einstein Radius and Shear Magnitude for a Non-axisymmetric Lens

Galaxies are rarely isolated, so we should include shear in our model, at the very least. (Also, galaxies are rarely spherical, but generalizing to elliptical models usually precludes solving for parameters analytically.) Suppose that we observe a lens with a "cross" configuration, where images form at $\pm\theta_1$ on the x-axis and a second pair at $\pm\theta_2$ on the y-axis. We immediately rule out a spherical lens, since the images and lens are not collinear. Equation (6.3a) describes an SIS lens, so we assume that model for convenience. Letting $x = \theta_1$ and $y = 0$, we find $u = 0$. If $x = 0$ and $y = \theta_2$, we find $v = 0$. Thus, a cross configuration results from a source directly behind the lens.

We have two remaining nontrivial equations from which we can determine the Einstein radius and shear:

$$0 = \theta_1 - b\,\mathrm{sgn}(\theta_1) + \gamma\theta_1\,, \tag{6.57a}$$

$$0 = \theta_2 - b\,\mathrm{sgn}(\theta_2) - \gamma\theta_2\,. \tag{6.57b}$$

Without loss of generality, we assume $\theta_{1,2} > 0$. Taking the difference of the above equations, we find a shear amplitude of

$$\gamma = -\left(\frac{\theta_1 - \theta_2}{\theta_1 + \theta_2}\right)\,. \tag{6.58}$$

If $\theta_1 > \theta_2$, we have $\gamma < 0$. This is equivalent to a rotation of the shear tensor by $\pi/2$. Finally, the Einstein radius of the primary lens is

$$b = \frac{2\theta_1\theta_2}{\theta_1 + \theta_2}\,. \tag{6.59}$$

The equivalent analysis for a point mass gives

$$\gamma = -\left(\frac{\theta_1^2 - \theta_2^2}{\theta_1^2 + \theta_2^2}\right)\,, \tag{6.60a}$$

$$b^2 = \frac{2\theta_1^2\theta_2^2}{\theta_1^2 + \theta_2^2}\,. \tag{6.60b}$$

Again we find that the source must be directly behind the lens.

6.5.2 Statistical Determination of Lens Parameters

In typical cases, the lens is not circular, and the source is not directly behind the lens. One important consequence is that the number of model parameters generally differs from the number of observational constraints. A common approach is the method of least squares, where we begin by defining a quantity known as the "goodness of fit" that has the form

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{\ell} \left[\frac{q_j^{\rm mod}(\mathbf{p}) - q_j^{\rm obs}}{\sigma_j} \right]^2 \\ &= \left[\mathbf{q}^{\rm mod}(\mathbf{p}) - \mathbf{q}^{\rm obs} \right]^T \mathsf{S}^{-1} \left[\mathbf{q}^{\rm mod}(\mathbf{p}) - \mathbf{q}^{\rm obs} \right] , \end{aligned} \tag{6.61}$$

where $\mathbf{q}^{\rm obs}$ is the ℓ-dimensional vector of observational constraints and $\mathbf{q}^{\rm mod}(\mathbf{p})$ is the corresponding vector of model values as a function of the m-dimensional vector $\mathbf{p}$ of parameters we wish to determine. We take the covariance matrix S to be diagonal, i.e., we assume that the n measurement uncertainties $\{\sigma_j\}$ are statistically independent. Specifically, we have $S_{ij} = \sigma_j^2 \delta_{ij}$. Note that the covariance matrix is always symmetric, even if it is not diagonal. The form of χ^2 is chosen so that more weight is given to terms with lower uncertainty. Our goal is to have $\mathbf{q}^{\rm mod}$ as close as possible to $\mathbf{q}^{\rm obs}$, which is accomplished by minimizing χ^2 with respect to $\mathbf{p}$.

Einstein Radius for an Axisymmetric Lens

To gain some experience working with the least squares fitting method, we reconsider the axisymmetric lenses of Sect. 6.5.1. We focus on the image positions and introduce a goodness of fit evaluated in the source plane:

$$\chi^2_{\rm src} = \sum_{i=1}^{n} \left(\boldsymbol{\beta}^{\rm mod} - \boldsymbol{\beta}_i^{\rm obs} \right)^T \mathsf{S}_i^{-1} \left(\boldsymbol{\beta}^{\rm mod} - \boldsymbol{\beta}_i^{\rm obs} \right) , \tag{6.62}$$

where n is the number of images. The model source position $\boldsymbol{\beta}^{\rm mod}$ is the same for all images. The values $\boldsymbol{\beta}_i^{\rm obs}$ are not observed directly but rather are computed from the observed image positions according to $\boldsymbol{\beta}_i^{\rm obs} = \boldsymbol{\beta}(\boldsymbol{\theta}_i^{\rm obs}) = \boldsymbol{\theta}_i^{\rm obs} - \boldsymbol{\alpha}(\boldsymbol{\theta}_i^{\rm obs})$. The quantity $\boldsymbol{\beta}_i^{\rm obs}$ is independent of the two-component source position $\boldsymbol{\beta}^{\rm mod}$ but does depend on the $m-2$ lens parameters $\mathbf{p}_{\rm lens}$. Specializing to the case of two collinear images with independent uncertainties (so the covariance matrix is diagonal), we can write

$$\chi^2_{\rm src} = \left[\frac{\beta^{\rm mod} - \beta_+^{\rm obs}(\mathbf{p}_{\rm lens})}{\sigma_+} \right]^2 + \left[\frac{\beta^{\rm mod} - \beta_-^{\rm obs}(\mathbf{p}_{\rm lens})}{\sigma_-} \right]^2 , \tag{6.63}$$

where $\sigma_\pm$ are the standard deviations of the measurements $\theta_\pm$. Our goal is to find the values of the parameters that minimize $\chi^2_{\rm src}$.

As before, the model parameters are the source position $\beta^{\rm mod}$ and the Einstein radius θ_E. To find the source position, we minimize $\chi^2_{\rm src}$ with respect to $\beta^{\rm mod}$. Setting the derivative

$$\frac{\partial \chi^2_{\rm src}}{\partial \beta^{\rm mod}} = \frac{2}{\sigma_+^2}\left(\beta^{\rm mod} - \beta_+^{\rm obs}\right) + \frac{2}{\sigma_-^2}\left(\beta^{\rm mod} - \beta_-^{\rm obs}\right) \tag{6.64}$$

equal to zero, we obtain

$$\beta^{\rm mod} = \frac{\sigma_-^2\beta_+^{\rm obs} + \sigma_+^2\beta_-^{\rm obs}}{\sigma_-^2 + \sigma_+^2}\,. \tag{6.65}$$

To minimize $\chi^2_{\rm src}$ with respect to the Einstein radius, we consider the derivative

$$\frac{\partial \chi^2_{\rm src}}{\partial \theta_E} = -2\left[\frac{1}{\sigma_+^2}\left(\beta^{\rm mod} - \beta_+^{\rm obs}\right)\frac{\partial \beta_+^{\rm obs}}{\partial \theta_E} + \frac{1}{\sigma_-^2}\left(\beta^{\rm mod} - \beta_-^{\rm obs}\right)\frac{\partial \beta_-^{\rm obs}}{\partial \theta_E}\right]. \tag{6.66}$$

Since we must specify a model in order to ascertain the dependence of the right-hand side on θ_E, we first consider the SIS. In that case, the lens equation yields

$$\beta_\pm^{\rm obs} = \theta_\pm \mp \theta_E\,. \tag{6.67}$$

Setting $\partial\chi^2_{\rm src}/\partial\theta_E = 0$ then gives

$$\theta_E = \left(\frac{1}{\sigma_+^2 + \sigma_-^2}\right)\left[\sigma_-^2\theta_+ - \sigma_+^2\theta_- + \beta^{\rm mod}(\sigma_+^2 - \sigma_-^2)\right]. \tag{6.68}$$

In order to compare with our results from Sect. 6.5.1, we set $\sigma_+ = \sigma_- \equiv \sigma$ and take the limit as $\sigma \to 0$. Then Eqs. (6.65) and (6.68) reduce to

$$\beta^{\rm mod} = \frac{1}{2}\left(\beta_+^{\rm obs} + \beta_-^{\rm obs}\right) \tag{6.69a}$$

$$\theta_E = \frac{1}{2}\left(\theta_+ - \theta_-\right) \tag{6.69b}$$

as before.

We now turn to the point mass, for which

$$\beta_\pm^{\rm obs} = \theta_\pm - \frac{\theta_E^2}{\theta_\pm}\,. \tag{6.70}$$

Equation (6.66) then reads

$$\frac{\partial \chi^2_{\rm src}}{\partial \theta_E} = 4\theta_E \left[\frac{1}{\sigma_+^2} \left(\frac{\beta^{\rm mod}}{\theta_+} - 1 + \frac{\theta_E^2}{\theta_+^2} \right) + \frac{1}{\sigma_-^2} \left(\frac{\beta^{\rm mod}}{\theta_-} - 1 + \frac{\theta_E^2}{\theta_-^2} \right) \right] . \tag{6.71}$$

Setting this to zero gives three solutions for θ_E. One of these is $\theta_E = 0$, which we can discard since it corresponds to the case of no lensing. As for the other two, we have

$$\theta_E^2 = \theta_+ \theta_- \left[\frac{\theta_+ \theta_- (\sigma_+^2 + \sigma_-^2) - \beta^{\rm mod} (\sigma_+^2 \theta_+ + \sigma_-^2 \theta_-)}{\theta_+^2 \sigma_+^2 + \theta_-^2 \sigma_-^2} \right] . \tag{6.72}$$

Combining this with Eq. (6.65) and again taking the limit $\sigma_+ = \sigma_- = \sigma \to 0$, we recover Eqs. (6.56).

The General Case

More generally, the observational constraints can include the fluxes and time delays as well as the positions of lensed images. Thus, we use the decomposition

$$\chi^2_{\rm tot} = \chi^2_{\rm pos} + \chi^2_{\rm flux} + \chi^2_{\rm tdel} . \tag{6.73}$$

The observational uncertainties are measured in the image plane, so it is best to define the χ^2 contribution from the image positions in that plane by writing

$$\chi^2_{\rm pos} = \sum_{i=1}^{n} \left[\boldsymbol{\theta}_i^{\rm mod}(\mathbf{p}) - \boldsymbol{\theta}_i^{\rm obs} \right]^T \mathsf{S}_i^{-1} \left[\boldsymbol{\theta}_i^{\rm mod}(\mathbf{p}) - \boldsymbol{\theta}_i^{\rm obs} \right] , \tag{6.74}$$

where there are n images and S_i is the covariance matrix of image i.[10] This expression requires that we solve the lens equation for each set of lens parameters that we wish to evaluate, which can be computationally demanding. In order to avoid solving the lens equation, we make a Taylor series expansion of the lens equation as

$$\begin{aligned} \boldsymbol{\beta}^{\rm mod} &\approx \boldsymbol{\beta}^{\rm mod}\Big|_{\boldsymbol{\theta}_i^{\rm mod} = \boldsymbol{\theta}_i^{\rm obs}} + \left. \frac{\partial \boldsymbol{\beta}^{\rm mod}}{\partial \boldsymbol{\theta}_i^{\rm mod}} \right|_{\boldsymbol{\theta}_i^{\rm mod} = \boldsymbol{\theta}_i^{\rm obs}} \left(\boldsymbol{\theta}_i^{\rm mod} - \boldsymbol{\theta}_i^{\rm obs} \right) \\ &\approx \boldsymbol{\beta}_i^{\rm obs} + \mathsf{A}_i^{-1} \left(\boldsymbol{\theta}_i^{\rm mod} - \boldsymbol{\theta}_i^{\rm obs} \right) , \end{aligned} \tag{6.75}$$

[10] If the images are blurred together, $\chi^2_{\rm pos}$ can be generalized to account for covariances between different images.

where $\mathsf{A}_i \equiv \mathsf{A}(\boldsymbol{\theta}_i^{\rm obs})$ is the amplification tensor at image i. This allows us to express $\chi^2_{\rm pos}$ as

$$\chi^2_{\rm pos} \approx \sum_{i=1}^{n} \left[\boldsymbol{\beta}^{\rm mod} - \boldsymbol{\beta}_i^{\rm obs}\right]^T \mathsf{T}_i^{-1} \left[\boldsymbol{\beta}^{\rm mod} - \boldsymbol{\beta}_i^{\rm obs}\right] . \tag{6.76}$$

This is similar to Eq. (6.62), except that the covariance matrices S_i in the image plane have been replaced by their counterparts T_i in the source plane. The two are related by

$$\mathsf{T}_i = \mathsf{A_i}^{-1} \mathsf{S}_i \left(\mathsf{A}_i^{-1}\right)^T . \tag{6.77}$$

The dependence of T_i on A_i means that although S_i is diagonal, T_i need not be. Moreover, the T_i are in general different from one another, even if all S_i are equal.

Keeton (2010) discusses the relative merits of the exact and approximate expressions for $\chi^2_{\rm pos}$. For now, we proceed with Eq. (6.76) for illustration. Since this expression is quadratic in $\boldsymbol{\beta}^{\rm mod}$, we can solve for $\boldsymbol{\beta}^{\rm mod}$ analytically. As a reminder of how to take the gradient of $\chi^2_{\rm pos}$ with respect to a vector $\boldsymbol{\beta}^{\rm mod}$, consider a scalar function of the form $\Phi(\mathbf{x}) \equiv \mathbf{x}^T \mathsf{M} \mathbf{x}$, where $\mathbf{x}$ has components x_i and M is a square, symmetric matrix with constant elements. The kth component of $\boldsymbol{\nabla}\Phi$ is

$$\begin{aligned} \frac{\partial \Phi}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i,j} x_i M_{ij} x_j \\ &= \sum_{i,j} \left(\delta_{ik} M_{ij} x_j + x_i M_{ij} \delta_{jk}\right) \\ &= \sum_j M_{kj} x_j + \sum_i x_i M_{ik} \\ &= 2 \sum_\ell M_{k\ell} x_\ell \, , \end{aligned} \tag{6.78}$$

where the last line follows from the symmetry of M. In matrix form, we have $\boldsymbol{\nabla}\Phi = 2\mathsf{M}\mathbf{x}$. Applying this result to Eq. (6.76), we can write

$$\boldsymbol{\nabla} \chi^2_{\rm pos} = 2 \sum_{i=1}^{n} \mathsf{T}_i^{-1} \left(\boldsymbol{\beta}^{\rm mod} - \boldsymbol{\beta}_i^{\rm obs}\right) , \tag{6.79}$$

where the gradient is taken with respect to $\boldsymbol{\beta}^{\rm mod}$. Note that T_i and its inverse are symmetric since the covariance matrix, S_i, is symmetric. Setting this to zero gives

$$\mathsf{C}\boldsymbol{\beta}^{\rm mod} = \mathbf{d} \, , \tag{6.80}$$

where

$$\mathsf{C} = \sum_{i=1}^{n} \mathsf{T}_i^{-1} = \sum_{i=1}^{n} \mathsf{A}_i^T \mathsf{S}_i^{-1} \mathsf{A}_i \tag{6.81}$$

$$\mathbf{d} = \sum_{i=1}^{n} \mathsf{T}_i^{-1} \boldsymbol{\beta}_i^{\rm obs} = \sum_{i=1}^{n} \mathsf{A}_i^T \mathsf{S}_i^{-1} \mathsf{A}_i \boldsymbol{\beta}_i^{\rm obs} \,. \tag{6.82}$$

For a single lens plane, A_i is symmetric, but we have not used this in our derivation, in order to be as general as possible.

The flux and time delay terms in Eq. (6.73) are also written as sums of squared differences between model and data:

$$\chi^2_{\rm flux} = \sum_{i=1}^{n} \left[\frac{\mu_i(\mathbf{p}) F_s^{\rm mod} - F_i^{\rm obs}}{\sigma_{F_i}} \right]^2 , \tag{6.83a}$$

$$\chi^2_{\rm tdel} = \sum_{ij} \left[\frac{\Delta T_{ij}^{\rm mod}(\mathbf{p}) - \Delta T_{ij}^{\rm obs}}{\sigma_{\Delta T_{ij}}} \right]^2 , \tag{6.83b}$$

where σ_{F_i} is the measurement uncertainty for the flux of image i, $\mu_i(\mathbf{p})$ is the predicted magnification for image i, which depends on the lens model parameters, and $F_s^{\rm mod}$ is the flux of the source, which is another parameter of the model. In Eq. (6.83b), $\Delta T_{ij}^{\rm obs}$ is the differential time delay measured between images i and j, $\sigma_{\Delta T_{ij}}$ is its uncertainty, and the sum runs over all image pairs for which the time delay has been measured. Note that $\chi^2_{\rm flux}$ is quadratic in $F_s^{\rm mod}$, so the optimal value can be found analytically, following a procedure analogous to what we did for $\boldsymbol{\beta}^{\rm mod}$ above.

6.5.3 *Modeling Lenses with Extended Sources*

The preceding discussion focused on lenses with point-like images for which the constraints are the positions and fluxes of the images as well as differential time delays. We now consider lenses in which the source is extended and the images are arcs or rings. At many wavelengths,[11] observations use charge-coupled devices to collect light into discrete pixels, so the data consist of the brightness in each pixel. We can collect them into a data vector $\mathbf{d}^{\rm obs} = \{d_i^{\rm obs}\}$ where the index i runs over pixels. Let $\boldsymbol{\theta}_i$ be the angular position of pixel i. Suppose we have a lens

[11] The exception is radio wavelengths, where observations usually use interferometry. The modeling methods discussed here can be generalized to handle interferometric data.

model characterized by parameters $\mathbf{p}_{\rm lens}$. We also have to build a model for the extended source, which is described by a surface brightness distribution $f_{\rm src}(\boldsymbol{\beta}|\mathbf{p}_{\rm src})$ that depends on some parameters $\mathbf{p}_{\rm src}$. Then we can use conservation of surface brightness (cf. Sect. 4.6.1) to write the model prediction as a vector $\mathbf{d}^{\rm mod}$ with components

$$d_i^{\rm mod}(\mathbf{p}_{\rm lens}, \mathbf{p}_{\rm src}) = f_{\rm src}\left[\boldsymbol{\theta}_i - \boldsymbol{\alpha}(\boldsymbol{\theta}_i|\mathbf{p}_{\rm lens})|\mathbf{p}_{\rm src}\right] . \tag{6.84}$$

(The formalism can be generalized to include blurring due to the atmosphere and telescope optics.) Then we can define a goodness of fit

$$\chi^2 = (\mathbf{d}^{\rm mod} - \mathbf{d}^{\rm obs})^T \mathsf{S}^{-1} (\mathbf{d}^{\rm mod} - \mathbf{d}^{\rm obs}) , \tag{6.85}$$

where S is the covariance matrix for the pixel brightnesses.

The modeling can proceed in a couple of different ways. One approach is to make assumptions about the structure of the source such that $f_{\rm src}$ can be described by a modest number of parameters. Then the lens parameters, $\mathbf{p}_{\rm lens}$, and source parameters, $\mathbf{p}_{\rm src}$, can be varied together in an attempt to find the best overall fit to the data.

A different approach is to decompose the source into a set of basis functions and fit for the amplitude of each mode. For example, one possibility is to define pixels in the source plane and let the set of source plane brightnesses, $\mathbf{s} = \{s_j\}$ where j runs over source plane pixels, be free parameters. In this case, conservation of surface brightness means that the predicted image can be written in the form $\mathbf{d}^{\rm mod} = \mathsf{L}(\mathbf{p}_{\rm lens})\,\mathbf{s}$, where the "lensing operator" L is a matrix that, in its simplest form, has entry $L_{ij} = 1$ if the lens model connects image pixel i to source pixel j, and $L_{ij} = 0$ otherwise. (The matrix can be generalized to allow interpolation between pixels.) In this case the goodness of fit has the form

$$\chi^2 = (\mathsf{L}\mathbf{s} - \mathbf{d}^{\rm obs})^T \mathsf{S}^{-1} (\mathsf{L}\mathbf{s} - \mathbf{d}^{\rm obs}) . \tag{6.86}$$

This is quadratic in $\mathbf{s}$, so the optimal set of source brightnesses can be found analytically following a procedure similar to that in Sect. 6.5.2. The same idea can be applied to other choices of basis functions, including smooth functions known as *shapelets*. Depending on the number of modes used, the source may have a considerable amount of freedom that can lead to unrealistic source models (especially in the presence of noisy data). It is therefore common to add a term to χ^2 that penalizes models for not conforming to certain assumptions about the source structure. Extensive discussions of this "free-form" approach to extended source modeling can be found in the literature (e.g., Warren and Dye 2003; Suyu et al. 2006; Vegetti and Koopmans 2009; Tagore and Keeton 2014; Birrer et al. 2015; Tagore and Jackson 2016).

6.6 Applications

Strong lensing by galaxies has been used for a wide range of astrophysical applications. While we cannot give an exhaustive review of the literature, we can introduce the different types of applications and provide references for readers who wish to dig deeper into particular topics.

One set of applications involves the physical properties of lens galaxies. Lensing is useful here because it provides a unique way to measure masses. Lens models constrain the Einstein radius in a way that is not very sensitive to the choice of model,[12] and the mass within the Einstein radius is among the most precise mass measurements available for distant galaxies. Mass measured with lensing is generally larger than the mass that can be ascribed to the visible stars and gas (even accounting for uncertainties in how much stellar mass is associated with a given luminosity), which provides strong support for the hypothesis that galaxies contain significant amounts of dark matter. Yet lensing is often limited in its ability to probe the mass distribution far from the Einstein radius. To learn more about the full density profile of a lens galaxy, three approaches have been taken. One is to combine lensing with other observations, such as the motions of stars within the lens galaxy, which probe the mass distribution on smaller scales. A second approach is to combine a number of lenses with different Einstein radii, making some (reasonable) assumptions about scaling relations in order to compare different galaxies. A third approach is to look for special lenses whose images span a range of distances from the center of the lens, offering mass constraints across that range. The general finding is that stars and dark matter combine to create a density profile that is close to isothermal. Lensing supports growing evidence that stellar populations can vary from one galaxy to another in a way that is not fully understood. (See the review by Treu (2010) for more discussion and an extensive set of references.)

A second application focuses on finer details in the mass distributions of lens galaxies. As discussed in Sect. 6.4.1, substructure in the lens galaxy can lead to images that are inconsistent with a smooth mass distribution. Such millilensing was first noted in four-image lenses with flux ratio "anomalies" that could not be explained by simple, smooth lens models (Mao and Schneider 1998; Metcalf and Madau 2001; Dalal and Kochanek 2002; Chiba 2002). The concept of substructure-induced anomalies has been extended to positions and time delays in quasar lenses (Chen et al. 2007; Keeton and Moustakas 2009), as well as to the shapes of arcs and rings (Koopmans 2005; Vegetti and Koopmans 2009; Hezaveh et al. 2013a). If the anomalies are thought to be dominated by a single subhalo, its properties can be constrained through direct modeling as in Sect. 6.5 (Vegetti et al. 2010a,b, 2012; Nierenberg et al. 2014; Hezaveh et al. 2016b). If the plethora of other subhalos throughout the lens cannot be ignored, more sophisticated statistical analysis

[12] Section 6.5.1 shows that the inferred Einstein radius does depend on the choice of model, but the quantitative difference between results from various models is typically no more than a few percent.

is needed (Hezaveh et al. 2016a; Birrer et al. 2017; Chatterjee and Koopmans 2018; Bayer et al. 2018); in particular, Cyr-Racine et al. (2016, 2018) present a comprehensive theory of lensing with an extended population of subhalos. Small halos projected along the line of sight need to be considered as well (Metcalf 2005; Inoue 2016; Despali et al. 2018). Current results certainly indicate that lensing has a unique ability to detect small-scale structure and may ultimately help determine the fundamental physical properties of dark matter particles.

A third application exploits a fortuitous match between interesting scales in the lens and source. As discussed in Sect. 6.4.2, stars in the lens galaxy create spatial variations in the lensing magnification on scales of microarcseconds. The optical, ultraviolet (UV), and X-ray emission from quasars originates from an accretion disk around a supermassive black hole that is small enough to be affected by microlensing. Thus, in quasar lenses the magnification can vary with both wavelength (because the emission region is smaller for X-rays than for UV and optical light) and time (as the stars and source move relative to one another), and observations of these effects can be used to measure the sizes of accretion disks as well as the density of stars responsible for microlensing. The analysis requires sophisticated statistical techniques that include generating many realizations of caustic networks (like the one shown in Fig. 6.7; see Kochanek (2004) and Mediavilla et al. (2006, 2009) for more discussion), but the effort is rewarded by unique constraints on structures that are far too small to observe directly. Overall, microlensing shows that accretion disk sizes generally scale with black hole mass in a way that follows theoretical predictions, but the optical sizes are larger than expected. It also supports the dark matter picture by indicating that stars cannot account for 100% of the mass density in the vicinity of lensed images. (See Kochanek et al. (2007) and Chartas et al. (2016) for more discussion and references.)

A fourth application likewise gives attention to the sources, but when they are galaxies rather than quasars. One interesting source population is dusty star-forming galaxies (DSFGs). Surveys at far-infrared and submillimeter wavelengths have revealed that a substantial fraction of star formation, especially in the early universe, occurs in regions that are rich in dust, which absorbs light from the young stars and reemits it at longer wavelengths. Studying DSFGs is therefore critical for a complete view of galaxy evolution (see the review by Casey et al. 2014). Most of the brightest observed DSFGs are lensed by intervening galaxies (they would not be so bright without lensing magnification; see Negrello et al. 2010; Hezaveh and Holder 2011; Hezaveh et al. 2013b). Determining their intrinsic properties therefore relies on fitting lens models that jointly reconstruct the light distribution of the source and the mass distribution of the lens. Because of lensing magnification, such models can reveal the intrinsic properties of DSFGs with a higher effective resolution than would be possible without lensing. As new facilities and instruments map spectral lines in increasing detail, lens modeling that uses the full spectral information offers a three-dimensional view of the kinematics of these distant star-forming galaxies (e.g., Geach et al. 2018; Rivera et al. 2018).

The last application we mention considers lenses in their cosmological context. Because the light rays cover cosmic distances, lensing can be used to probe the

global geometry of the universe, provided that we understand how the rays are bent by the lens. Using Eqs. (4.24) and (4.25), we can write the differential time delay between two lensed images at positions $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ as

$$\Delta T_{ij} = \frac{1+z_l}{c}\,\frac{D_l D_s}{D_{ls}}\left[\tau(\boldsymbol{\theta}_i|\boldsymbol{\beta}) - \tau(\boldsymbol{\theta}_j|\boldsymbol{\beta})\right], \tag{6.87}$$

where $\tau(\boldsymbol{\theta}|\boldsymbol{\beta}) = \frac{1}{2}|\boldsymbol{\theta}-\boldsymbol{\beta}|^2 - \psi(\boldsymbol{\theta})$ is the Fermat potential. If the source varies with time, the temporal offset between two light curves reveals the differential time delay ΔT_{ij}. If we use a lens model to determine the lens potential ψ and the source position $\boldsymbol{\beta}$, we can combine the time delay and Fermat potential to constrain the "time delay distance" as

$$D_{\Delta T}(z_l, z_s) \equiv \frac{1+z_l}{c}\,\frac{D_l D_s}{D_{ls}} = \frac{\Delta T_{ij}}{\tau(\boldsymbol{\theta}_i|\boldsymbol{\beta}) - \tau(\boldsymbol{\theta}_j|\boldsymbol{\beta})}\,. \tag{6.88}$$

The time delay distance depends on the lens and source redshifts, obviously, but it also depends on cosmological parameters through Eq. (3.152) for the angular diameter distance. The upshot is that gravitational lens time delays can be used to constrain cosmological parameters. This "time delay cosmography" provides a valuable complement to other cosmological probes (such as the cosmic microwave background, discussed in Chap. 9, as well as Type Ia supernovae) because it involves different physics and different data. There are many details that must be handled with care, which are illustrated by a project known as H0LiCOW for "H_0 Lenses in COSMOGRAIL's Wellspring" (Suyu et al. 2017). These include measuring lens light curves in order to determine the time delays (Bonvin et al. 2017) and then fitting a lens model (Wong et al. 2017). They also include understanding the distribution of matter around the lens and along the line of sight (Sluse et al. 2017; Rusu et al. 2017) in order to avoid a degeneracy associated with the mass-sheet transformation (cf. Sect. 4.7.2; but see McCully et al. (2014, 2017) for an approach that accounts for multiplane lensing). With such attention to detail, lensing can provide cosmological constraints that are competitive with other probes. (See the review by Treu and Marshall (2016) for more discussion.)

Problems

6.1 For a given source position, we can think of the time delay function $T(\boldsymbol{\theta}|\boldsymbol{\beta})$ as a surface and use Fermat's principle (Sect. 4.1.2) to find the images as stationary points on the surface. Plot the time delay surface for an SIS lens with Einstein radius $\theta_E = 1''$ and a source at position $(0.3'', 0'')$. Comment on its topography (i.e., identify the stationary points).

6.2 We can visualize image configurations without solving the lens equation by using conservation of surface brightness (cf. Sect. 4.6.1): if the surface brightness in the source plane is $S(\boldsymbol{\beta})$, then the surface brightness in the image plane is $I(\boldsymbol{\theta}) = S\big(\boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})\big)$. This is the method that was used to create Fig. 6.1. Make your own version of the figure for an SIS with external shear.

6.3 For the SIEP model, plot convergence maps for different values of the axis ratio. Based on what you see, explain why the SIEP is not considered to be a good model for galaxies.

6.4 For the SIEMD model,

$$\kappa = \frac{b}{2\sqrt{q^2x^2 + y^2}}\,,$$

the two components of deflection are (see Kassiola and Kovner 1993; Keeton and Kochanek 1998)

$$\alpha_x = \frac{b}{\sqrt{1-q^2}} \tan^{-1}\left(\frac{\sqrt{1-q^2}\,x}{\sqrt{q^2x^2+y^2}}\right)$$
$$\alpha_y = \frac{b}{\sqrt{1-q^2}} \tanh^{-1}\left(\frac{\sqrt{1-q^2}\,y}{\sqrt{q^2x^2+y^2}}\right),$$

and the lens potential can be written as $\psi = x\,\alpha_x + y\,\alpha_y$.

(a) Find the tangential critical curve, and show that it is an isodensity contour. Write a parametric form for the tangential caustic.
(b) The pseudocaustic is the set of points in the source plane that map to the origin in the image plane. To find it, consider an image plane circle with radius R, map it to the source plane, and take the limit $R \to 0$.
(c) For what range of axis ratios does the SIEMD have naked cusps?
(d) Use the method from Problem 6.2 to make a plot analogous to Fig. 6.1 but now for an SIEMD lens.
(e) For one of the fold lenses from part (d), use the time delay surface or eigenvalues of the amplification tensor to classify the images as minima, maxima, or saddles.

6.5 Use the method from Problem 6.2 to plot image configurations for a nonsingular isothermal sphere with shear.

(a) Consider a small value of the core radius. How is this lens similar to or different from an SIS with shear?
(b) Consider several "large" values of the core radius that correspond to different stages of the caustic metamorphosis shown in Fig. 6.4.

6.6 In Sect. 6.2.3 we considered caustic metamorphosis for an NIS with shear, focusing on the conditions for critical curves and caustics to disappear. Now find the value of s for which the two critical curves intersect.

6.7 Suppose you observe three different cross lenses with the following image positions and brightnesses. (The positions are measured relative to the center of the lens, and the brightnesses are given relative to the first image.)

Lens	x ($''$)	y ($''$)	Brightness
#1	0.679	−1.177	$\equiv 1$
	−0.679	1.177	1.000
	0.993	0.574	0.621
	−0.993	−0.574	0.621
# 2	−0.622	0.741	$\equiv 1$
	0.622	−0.741	1.000
	−0.644	−0.541	0.869
	0.644	0.541	0.869
#3	−0.752	0.907	$\equiv 1$
	0.752	−0.907	1.000
	−0.779	−0.590	0.695
	0.779	0.590	0.695

Fit the lenses using an isothermal model that has shear, ellipticity, or both. For each lens, determine which is the correct model.

Note: You can work analytically, in which case it may be helpful to rotate into a reference frame aligned with the images. Or you can work numerically using the methods discussed in Sect. 6.5.2.

6.8 Suppose you observe three different circular lenses with central images. (As in Problem 6.7, the positions are measured relative to the center of the lens, and the brightnesses are given relative to the first image.)

Lens	x ($''$)	y ($''$)	Brightness
A	1.268	0	$\equiv 1$
	−0.765	0	0.7697
	−0.103	0	0.0437
B	1.362	0	$\equiv 1$
	−0.922	0	0.7306
	−0.040	0	0.0060
C	1.406	0	$\equiv 1$
	−0.988	0	0.7254
	−0.018	0	0.0012

Fit the lenses using a nonsingular isothermal sphere. How would your results change if you had an upper limit on the brightness of each central image rather than a measurement?

References

Bayer, D., Chatterjee, S., Koopmans, L. V. E., Vegetti, S., McKean, J. P., Treu, T., et al. (2018). Observational constraints on the sub-galactic matter-power spectrum from galaxy-galaxy strong gravitational lensing. arXiv e-prints, arXiv:1803.05952.

Birrer, S., Amara, A., & Refregier, A. (2015). Gravitational lens modeling with basis sets. *The Astrophysical Journal, 813*, 102.

Birrer, S., Amara, A., & Refregier, A. (2017). Lensing substructure quantification in RX J1131−1231: A 2 keV lower bound on dark matter thermal relic mass. *Journal of Cosmology and Astroparticle Physics, 5*, 037.

Bonvin, V., Courbin, F., Suyu, S. H., Marshall, P. J., Rusu, C. E., Sluse, D., et al. (2017). H0LiCOW - V. New COSMOGRAIL time delays of HE 0435−1223: H_0 to 3.8 per cent precision from strong lensing in a flat ΛCDM model. *Monthly Notices of the Royal Astronomical Society, 465*, 4914.

Casey, C. M., Narayanan, D., & Cooray, A. (2014). Dusty star-forming galaxies at high redshift. *Physics Reports, 541*, 45.

Chartas, G., Rhea, C., Kochanek, C., Dai, X., Morgan, C., Blackburne, J., et al. (2016). Gravitational lensing size scales for quasars. *Astronomische Nachrichten, 337*, 356.

Chatterjee, S., & Koopmans, L. V. E. (2018). The inner mass power spectrum of galaxies using strong gravitational lensing: Beyond linear approximation. *Monthly Notices of the Royal Astronomical Society, 474*, 1762.

Chen, J., Rozo, E., Dalal, N., & Taylor, J. E. (2007). Astrometric perturbations in substructure lensing. *The Astrophysical Journal, 659*, 52.

Chiba, M. (2002). Probing dark matter substructure in lens galaxies. *The Astrophysical Journal, 565*, 17.

Cyr-Racine, F.-Y., Keeton, C. R., & Moustakas, L. A. (2018). Beyond subhalos: Probing the collective effect of the universe's small-scale structure with gravitational lensing. arXiv e-prints, arXiv:1806.07897.

Cyr-Racine, F.-Y., Moustakas, L. A., Keeton, C. R., Sigurdson, K., & Gilman, D. A. (2016). Dark census: Statistically detecting the satellite populations of distant galaxies. *Physical Review D, 94*, 043505.

Dalal, N., & Kochanek, C. S. (2002). Direct detection of cold dark matter substructure. *The Astrophysical Journal, 572*, 25.

Despali, G., Vegetti, S., White, S. D. M., Giocoli, C., & van den Bosch, F. C. (2018). Modelling the line-of-sight contribution in substructure lensing. *Monthly Notices of the Royal Astronomical Society, 475*, 5424.

Geach, J. E., Ivison, R. J., Dye, S., & Oteo, I. (2018). A magnified view of circumnuclear star formation and feedback around an AGN at $z = 2.6$. *The Astrophysical Journal Letters, 866*, L12.

Hezaveh, Y., Dalal, N., Holder, G., Kisner, T., Kuhlen, M., & Perreault Levasseur, L. (2016a). Measuring the power spectrum of dark matter substructure using strong gravitational lensing. *Journal of Cosmology and Astroparticle Physics, 11*, 048.

Hezaveh, Y., Dalal, N., Holder, G., Kuhlen, M., Marrone, D., Murray, N., et al. (2013a). Dark matter substructure detection using spatially resolved spectroscopy of lensed dusty galaxies. *The Astrophysical Journal, 767*, 9.

Hezaveh, Y. D., Dalal, N., Marrone, D. P., Mao, Y.-Y., Morningstar, W., Wen, D., et al. (2016b). Detection of lensing substructure using ALMA Observations of the Dusty Galaxy SDP.81. *The Astrophysical Journal, 823*, 37.

Hezaveh, Y. D., & Holder, G. P. (2011). Effects of strong gravitational lensing on millimeter-wave galaxy number counts. *The Astrophysical Journal, 734*, 52.

Hezaveh, Y. D., Marrone, D. P., Fassnacht, C. D., Spilker, J. S., Vieira, J. D., Aguirre, J. E., et al. (2013b). ALMA Observations of SPT-discovered, strongly lensed, dusty, star-forming galaxies. *The Astrophysical Journal, 767*, 132.

Holder, G. P., & Schechter, P. L. (2003). External shear in quadruply imaged lens systems. *The Astrophysical Journal, 589*, 688.

Inada, N., Oguri, M., Keeton, C. R., Eisenstein, D. J., Castander, F. J., Chiu, K., et al. (2005). Discovery of a fifth image of the large separation gravitationally lensed quasar SDSS J1004+4112. *Publications of the Astronomical Society of Japan, 57*, L7.

Inoue, K. T. (2016). On the origin of the flux ratio anomaly in quadruple lens systems. *Monthly Notices of the Royal Astronomical Society, 461*, 164.

Kassiola, A., & Kovner, I. (1993). Elliptic mass distributions versus elliptic potentials in gravitational lenses. *The Astrophysical Journal, 417*, 450.

Keeton, C. R. (2003). Lensing and the centers of distant early-type galaxies. *The Astrophysical Journal, 582*, 17.

Keeton, C. R. (2010). On modeling galaxy-scale strong lens systems. *General Relativity and Gravitation, 42*, 2151.

Keeton, C. R., Falco, E. E., Impey, C. D., Kochanek, C. S., Lehár, J., McLeod, B. A., et al. (2000a). The host galaxy of the lensed quasar Q0957+561. *The Astrophysical Journal, 542*, 74.

Keeton, C. R., & Kochanek, C. S. (1998). Gravitational lensing by spiral galaxies. *The Astrophysical Journal, 495*, 157.

Keeton, C. R., Mao, S., & Witt, H. J. (2000b). Gravitational lenses with more than four images. I. Classification of caustics. *The Astrophysical Journal, 537*, 697.

Keeton, C. R., & Moustakas, L. A. (2009). A new channel for detecting dark matter substructure in galaxies: Gravitational lens time delays. *The Astrophysical Journal, 699*, 1720.

Kochanek, C. S. (2004). Quantitative interpretation of quasar microlensing light curves. *The Astrophysical Journal, 605*, 58.

Kochanek, C. S., Dai, X., Morgan, C., Morgan, N., & Poindexter, G., S. C. (2007). Turning AGN microlensing from a curiosity into a tool. In G. J. Babu & E. D. Feigelson (Eds.), *Statistical challenges in modern astronomy IV. Astronomical society of the Pacific conference series* (Vol. 371, p. 43). New York: Springer.

Kochanek, C. S., Keeton, C. R., & McLeod, B. A. (2001). The importance of Einstein rings. *The Astrophysical Journal, 547*, 50.

Koopmans, L. V. E. (2005). Gravitational imaging of cold dark matter substructures. *Monthly Notices of the Royal Astronomical Society, 363*, 1136.

Kormann, R., Schneider, P., & Bartelmann, M. (1994). Isothermal elliptical gravitational lens models. *Astronomy and Astrophysics, 284*, 285

Mao, S., & Schneider, P. (1998). Evidence for substructure in lens galaxies? *Monthly Notices of the Royal Astronomical Society, 295*, 587.

Mao, S., Witt, H. J., & Koopmans, L. V. E. (2001). The influence of central black holes on gravitational lenses. *Monthly Notices of the Royal Astronomical Society, 323*, 301.

McCully, C., Keeton, C. R., Wong, K. C., & Zabludoff, A. I. (2014). A new hybrid framework to efficiently model lines of sight to gravitational lenses. *Monthly Notices of the Royal Astronomical Society, 443*, 3631.

McCully, C., Keeton, C. R., Wong, K. C., & Zabludoff, A. I. (2017). Quantifying environmental and line-of-sight effects in models of strong gravitational lens systems. *The Astrophysical Journal, 836*, 141.

Mediavilla, E., Muñoz, J. A., Falco, E., Motta, V., Guerras, E., Canovas, H., et al. (2009). Microlensing-based estimate of the mass fraction in compact objects in lens galaxies. *The Astrophysical Journal, 706*, 1451.

Mediavilla, E., Muñoz, J. A., Lopez, P., Mediavilla, T., Abajas, C., Gonzalez-Morcillo, C., et al. (2006). A fast and very accurate approach to the computation of microlensing magnification patterns based on inverse polygon mapping. *The Astrophysical Journal, 653*, 942.

Metcalf, R. B. (2005). The importance of intergalactic structure to gravitationally lensed quasars. *The Astrophysical Journal, 629*, 673.

Metcalf, R. B., & Madau, P. (2001). Compound gravitational lensing as a probe of dark matter substructure within galaxy halos. *The Astrophysical Journal, 563*, 9.

Negrello, M., Hopwood, R., De Zotti, G., Cooray, A., Verma, A., Bock, J., et al. (2010). The detection of a population of submillimeter-bright, strongly lensed galaxies. *Science, 330*, 800.

Nierenberg, A. M., Treu, T., Wright, S. A., Fassnacht, C. D., & Auger, M. W. (2014). Detection of substructure with adaptive optics integral field spectroscopy of the gravitational lens B1422+231. *Monthly Notices of the Royal Astronomical Society, 442*, 2434.

Oguri, M., & Keeton, C. R. (2004). Effects of triaxiality on the statistics of large-separation gravitational lenses. *The Astrophysical Journal, 610*, 663.

Oguri, M., Ofek, E. O., Inada, N., Morokuma, T., Falco, E. E., Kochanek, C. S., et al. (2008). The third image of the large-separation lensed quasar SDSS J1029+2623. *The Astrophysical Journal Letters, 676*, L1.

Petters, A. O., Levine, H., & Wambsganss, J. (2001). *Singularity theory and gravitational lensing*. Boston: Birkhäuser.

Rivera, J., Baker, A. J., Gallardo, P. A., Gralla, M., Harris, A. I., Huffenberger, K. M., et al. (2018). The Atacama Cosmology Telescope: CO($J = 3 - 2$) mapping and lens modeling of an ACT-selected dusty star-forming galaxy. arXiv e-prints, arXiv:1807.08895.

Rusin, D., Keeton, C. R., & Winn, J. N. (2005). Measuring supermassive black holes in distant galaxies with central lensed images. *The Astrophysical Journal Letters, 627*, L93.

Rusin, D., & Ma, C.-P. (2001). Constraints on the inner mass profiles of lensing galaxies from missing odd images. *The Astrophysical Journal Letters, 549*, L33.

Rusu, C. E., Fassnacht, C. D., Sluse, D., Hilbert, S., Wong, K. C., Huang, K.-H., et al. (2017). H0LiCOW - III. Quantifying the effect of mass along the line of sight to the gravitational lens HE 0435−1223 through weighted galaxy counts. *Monthly Notices of the Royal Astronomical Society, 467*, 4220.

Schneider, P., Ehlers, J., & Falco, E. E. (1992). *Gravitational lenses*. Berlin: Springer.

Sluse, D., Sonnenfeld, A., Rumbaugh, N., Rusu, C. E., Fassnacht, C. D., Treu, T., et al. (2017). H0LiCOW - II. Spectroscopic survey and galaxy-group identification of the strong gravitational lens system HE 0435−1223. *Monthly Notices of the Royal Astronomical Society, 470*, 4838.

Suyu, S. H., Bonvin, V., Courbin, F., Fassnacht, C. D., Rusu, C. E., Sluse, D., et al. (2017). H0LiCOW - I. H_0 lenses in COSMOGRAIL's Wellspring: Program overview. *Monthly Notices of the Royal Astronomical Society, 468*, 2590.

Suyu, S. H., Marshall, P. J., Hobson, M. P., & Blandford, R. D. (2006). A Bayesian analysis of regularized source inversions in gravitational lensing. *Monthly Notices of the Royal Astronomical Society, 371*, 983.

Tagore, A. S., & Jackson, N. (2016). On the use of shapelets in modelling resolved, gravitationally lensed images. *Monthly Notices of the Royal Astronomical Society, 457*, 3066.

Tagore, A. S., & Keeton, C. R. (2014). Statistical and systematic uncertainties in pixel-based source reconstruction algorithms for gravitational lensing. *Monthly Notices of the Royal Astronomical Society, 445*, 694.

Treu, T. (2010). Strong lensing by galaxies. *Annual Review of Astronomy and Astrophysics, 48*, 87.

Treu, T., & Marshall, P. J. (2016). Time delay cosmography. *The Astronomy and Astrophysics Review, 24*, 11.

Vegetti, S., Czoske, O., & Koopmans, L. V. E. (2010a). Quantifying dwarf satellites through gravitational imaging: The case of SDSSJ120602.09+514229.5. *Monthly Notices of the Royal Astronomical Society, 407*, 225.

Vegetti, S., & Koopmans, L. V. E. (2009). Bayesian strong gravitational-lens modelling on adaptive grids: Objective detection of mass substructure in Galaxies. *Monthly Notices of the Royal Astronomical Society, 392*, 945.

Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T., & Gavazzi, R. (2010b). Detection of a dark substructure through gravitational imaging. *Monthly Notices of the Royal Astronomical Society, 408*, 1969.

Vegetti, S., Lagattuta, D. J., McKean, J. P., Auger, M. W., Fassnacht, C. D., & Koopmans, L. V. E. (2012). Gravitational detection of a low-mass dark satellite galaxy at cosmological distance. *Nature, 481*, 341.

Wambsganss, J. (1999). Gravitational lensing: Numerical simulations with a hierarchical tree code. *Journal of Computational and Applied Mathematics, 109*, 353.

Warren, S. J., & Dye, S. (2003). Semilinear gravitational lens inversion. *The Astrophysical Journal, 590*, 673.

Winn, J. N., Rusin, D., & Kochanek, C. S. (2004). The central image of a gravitationally lensed quasar. *Nature, 427*, 613.

Wong, K. C., Suyu, S. H., Auger, M. W., Bonvin, V., Courbin, F., Fassnacht, C. D., et al. (2017). H0LiCOW - IV. Lens mass model of HE 0435−1223 and blind measurement of its time-delay distance for cosmology. *Monthly Notices of the Royal Astronomical Society, 465*, 4895.

Chapter 7
Strong and Weak Lensing by Galaxy Clusters

In some parts of the Universe, hundreds of galaxies are bound by their mutual gravity into enormous structures known as clusters. The lensing properties of galaxy clusters differ from those of individual galaxies for two reasons: a cluster is a hundred to a thousand times more massive than a galaxy, and the inner density profile of a cluster is shallower than isothermal. Since a larger mass means a larger Einstein radius, the angular separation between the lens and source for which lensing can be observed is larger for a cluster than for a galaxy. Consequently, there is a non-negligible probability that two or more sources will be lensed by a single cluster. The sources are likely to have different redshifts and hence different values of the distance ratio D_{ls}/D_s that enters the reduced deflection angle. Figure 7.1 shows how the distance ratio affects the locations of lensed images.

In our discussion of cored galaxy models in Sect. 6.2, we saw that radial caustics emerge for density profiles that are less concentrated than the singular isothermal sphere. This allows the creation of radial arcs for a source sufficiently close to the radial caustic. The bottom panels of Fig. 7.1 illustrate radial arcs and show that they, too, depend on D_{ls}/D_s. Section 7.1 extends our previous discussion of strong lensing to clusters.

Since they are so massive, clusters can produce distortions well beyond the strong lensing region. These distortions are too small to be detected individually, but they can be extracted from statistical analysis of thousands of galaxies. Section 7.2 introduces the theory of **weak lensing** and describes how measurements of shape distortions can be used to constrain lens mass distributions. Section 7.3 then discusses applications of both strong and weak lensing by clusters.

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_7

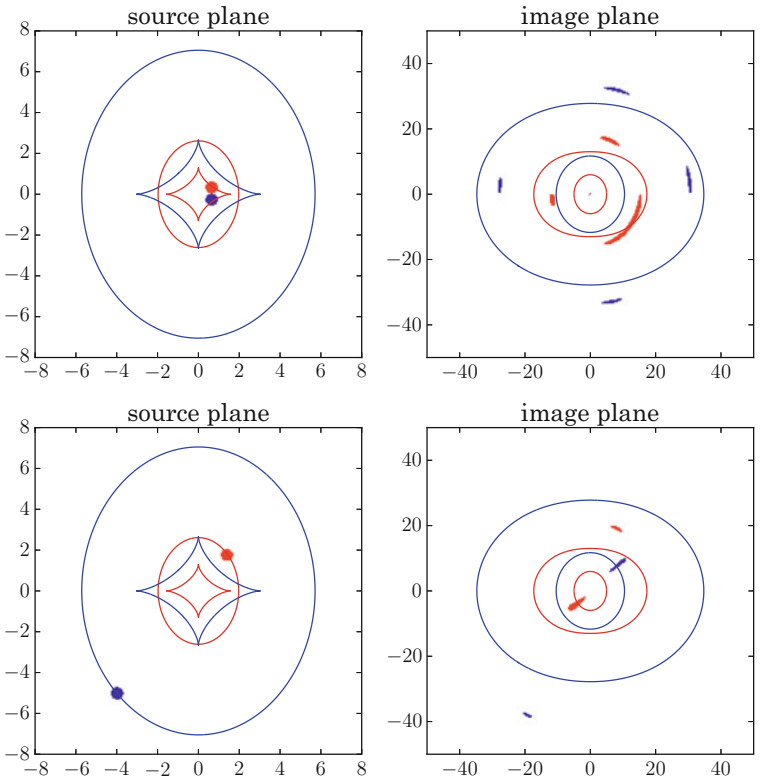


Fig. 7.1 Illustration of cluster strong lensing with sources at different distances behind the lens: red shows $D_{ls}/D_s = 0.5$, while blue shows $D_{ls}/D_s = 0.7$. The left panels show caustics, while the right panels show the corresponding critical curves. In the top row, the red source yields a tangential arc, while the blue source yields a four-image configuration because this source is well inside its caustic (remember that the red caustic corresponds to a different source plane). In the bottom row, both sources yield radial arcs. We see that sources with larger values of D_{ls}/D_s create images that tend to lie farther from the center of the lens. The lens here is an elliptical NFW model (see Sect. 7.1.1)

7.1 Strong Lensing by Clusters

7.1.1 Navarro-Frenk-White (NFW) Model

While a power law of the form $\rho \propto r^{-\eta}$ works well for galaxies (e.g., the singular isothermal sphere), it does not apply to clusters. A natural generalization is to allow the power law index to take on one value for small radii and another for large radii. To ensure a smooth transition between these regimes, we write the density profile as

$$\rho(r) = \frac{\rho_s r_s^q}{r^p (r_s + r)^{q-p}} , \tag{7.1}$$

where r_s is a scale radius such that $\rho \propto r^{-p}$ when $r \ll r_s$ and $\rho \propto r^{-q}$ when $r \gg r_s$. Navarro et al. (1995, 1996, 1997) found $p = 1$ and $q = 3$ from cosmological simulations of cluster-sized dark matter halos. Compared to the isothermal profile, the **NFW profile** is shallower near the center and steeper farther out.

To find its lensing properties, we begin by calculating the mass within a sphere of radius r:

$$M_{3D}(r) = 4\pi \int_0^r \rho(r')(r')^2 dr' = 4\pi\rho_s r_s^3 \left[\ln\left(1 + \frac{r}{r_s}\right) - \frac{r}{r_s + r} \right]. \tag{7.2}$$

The projected (two-dimensional) mass within a circle of radius R is then (cf. Problem 2.1)

$$M_{2D}(R) = 4\pi\rho_s r_s^3 \left[\ln\left(\frac{R}{2r_s}\right) + F_{NFW}\left(\frac{R}{r_s}\right) \right], \tag{7.3}$$

where

$$F_{NFW}(x) = \begin{cases} \frac{1}{\sqrt{x^2-1}} \tan^{-1}\sqrt{x^2-1} & (x > 1) \\ \frac{1}{\sqrt{1-x^2}} \tanh^{-1}\sqrt{1-x^2} & (x < 1) \\ 1 & (x = 1)\,. \end{cases} \tag{7.4}$$

The reduced deflection angle and convergence work out to be

$$\alpha(\theta) = 4\kappa_s\theta_s \frac{\ln(x/2) + F_{NFW}(x)}{x} \tag{7.5a}$$

$$\kappa(\theta) = 2\kappa_s \frac{1 - F_{NFW}(x)}{x^2 - 1}\,, \tag{7.5b}$$

where $x = \theta/\theta_s$, $\theta_s = r_s/D_l$ is the scale radius in angular units, and $\kappa_s = \rho_s r_s/\Sigma_{crit}$ is a characteristic convergence that is also known as the lensing strength of an NFW halo. These functions are shown in Fig. 7.2. Bartelmann (1996) first performed this analysis in order to demonstrate that NFW lenses can produce radial as well as tangential arcs.

Clusters usually do not have perfect spherical symmetry, so it is interesting to consider a projected mass distribution with elliptical symmetry. To define an elliptical lens model, let (x, y) be Cartesian coordinates in the lens plane, and let us choose a reference frame that is centered on the lens and has the x-axis aligned with the major axis of the ellipse. (Any other reference frame can be obtained by rotation and translation.) We can take the projected mass distribution described by Eq. (7.5b) and make it elliptical by replacing θ with $\sqrt{x^2 + y^2/q^2}$, where $0 < q \leq 1$ is the minor/major axis ratio of the ellipse. Then the two components of the deflection vector can be computed as follows (Schramm 1990; Keeton 2001):

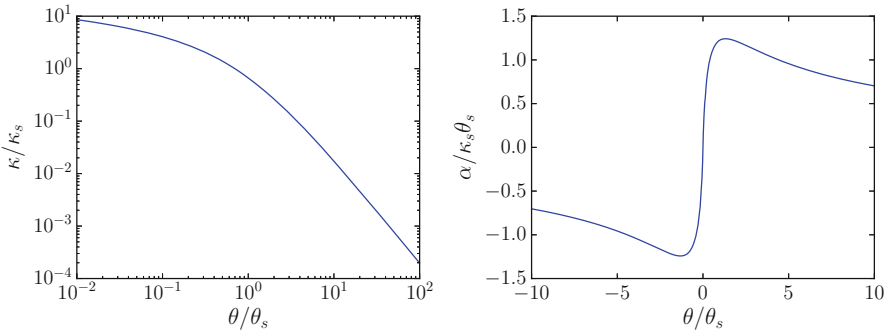


Fig. 7.2 Convergence (left) and deflection (right) for an NFW lens model, from Eq. (7.5a)

$$\alpha_x(x, y) = q\, x\, J_0(x, y) \tag{7.6a}$$

$$\alpha_y(x, y) = q\, y\, J_1(x, y)\,, \tag{7.6b}$$

where the integrals have the form[1]

$$J_n(x, y) = \int_0^1 \frac{\kappa(\xi(u))}{[1 - (1 - q^2)u]^{n+1/2}}\, \mathrm{d}u\,,$$

$$\text{where} \quad \xi(u) = \sqrt{u\left[x^2 + \frac{y^2}{1 - (1 - q^2)u}\right]}\,. \tag{7.7}$$

(Schramm (1990) and Keeton (2001) also give corresponding expressions for the lens potential and the components of the amplification tensor.) These integrals cannot be evaluated analytically, but they can be computed numerically.

7.1.2 Cluster Lens Modeling

Once lensed images are observed, they can be used to constrain the cluster's mass distribution in a generalization of the lens modeling introduced in Sect. 6.5. Cluster mass distributions are sufficiently complicated that any detailed analysis must be computational, but we can highlight some key principles (for more discussion, see Kneib and Natarajan 2011).

A simple axisymmetric lens illustrates what can be learned from cluster arcs. Recall from Sect. 2.2 that the reduced deflection angle can be written as

[1]These functions should not be confused with cylindrical Bessel functions encountered in Appendix C.2.2 and elsewhere.

$$\alpha(\theta) = \frac{4GD_{ls}}{c^2 D_l D_s} \frac{M_{2D}(D_l\theta)}{\theta} . \tag{7.8}$$

Tangential arcs form near the Einstein radius, θ_E, which solves $\alpha(\theta_E) = \theta_E$. Thus, if we observe a tangential arc, we can essentially determine θ_E, and that immediately yields the mass enclosed by θ_E as

$$M_{2D}(D_l\theta_E) = \frac{c^2 D_l D_s \theta_E^2}{4GD_{ls}} . \tag{7.9}$$

By contrast, radial arcs are associated with the radial critical curve, which occurs at a location where $d\alpha/d\theta = 1$. That derivative involves $dM_{2D}/d\theta$, which is related to the projected density profile, $\Sigma(D_l\theta)$. The details are more complicated when the lens is not axisymmetric, but still the basic idea is that tangential arcs constrain the enclosed mass, while radial arcs constrain the density profile of the cluster.

Long arcs are visually spectacular but rare, because they require the source to lie on or near a caustic. Multiple images of small, faint galaxies are more common; for example, the cluster MACS J0416.1−2403 has more than 90 confirmed images of more than 30 different sources (see Fig. 7.3). The large collection of images can

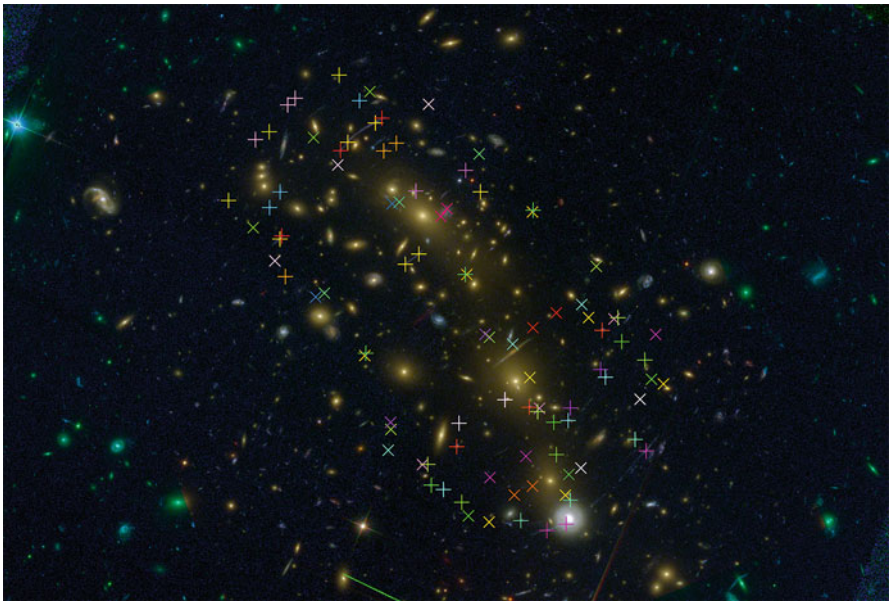


Fig. 7.3 *Hubble Space Telescope* image of the cluster MACS J0416.1−2403, taken as part of the *Hubble* Frontier Fields program (Lotz et al. 2017). The markers indicate confirmed lensed images, where the point type and color are the same for multiple images of the same source. (The processed *HST* image was provided by D. Coe for the *HFF* program. Data for the lensed images are given by Christensen et al. (2012), Jauzac et al. (2014), Schmidt et al. (2014), Diego et al. (2015b), Grillo et al. (2015), Treu et al. (2015), and Caminha et al. (2017))

provide a lot of information about the cluster mass distribution, especially if the images are spread out to cover most of the strong lensing region, but matching images requires some care. It may not be immediately obvious which small, faint galaxies are common images of a single background object. Information from color, morphology, and redshift can be used to determine which galaxies are likely to be related lensed images (see, e.g., Broadhurst et al. (2005) for more discussion). Redshifts are particularly important for two reasons. First, they provide a stringent test of the lensing hypothesis: if two galaxies do not have the same redshift, they cannot be lensed images of a single source. This test demands precise, well-measured spectroscopic redshifts. Second, having sources at different redshifts provides constraints at different distances from the cluster center (see Fig. 7.1), but the redshifts must be known in order to determine the factor D_{ls}/D_s for each image system. For this purpose it may be sufficient to use "photometric redshifts" estimated from color information, as long as a few of the image systems are anchored with precise spectroscopic redshifts (see Johnson and Sharon 2016).

Clusters contain structure on various scales that needs to be incorporated into lens models. The visible galaxies, which are tens of kiloparsecs across and hundreds of kiloparsecs apart, are surrounded by clouds of gas and dark matter that are of order one megaparsec in size. The gas can be detected directly because it is hot enough to glow at X-ray wavelengths, and it interacts with photons from the cosmic microwave background (see Sect. 9.1). The dark matter is detected through its gravitational influence on the galaxies and gas and through gravitational lensing. These large-scale components are often treated using either an NFW model or a nonsingular isothermal ellipsoid. It may be sufficient to use a single halo component if the cluster has reached dynamical equilibrium, or it may be necessary to use several such components if the system has had a more active history, such as a recent merger event, that has led to a more complicated mass distribution. Examples of such lens models are given in Sect. 7.3.

The galaxies can be observed with optical telescopes and inserted into lens models at their measured positions. They are often treated using truncated isothermal models, using Eq. (6.50) or its generalization to an elliptical mass distribution. With this model, each galaxy in the cluster is described by its position, velocity dispersion, core radius, and truncation radius.[2] In most cases there are too many galaxies to constrain all of their parameters using lensed images alone. It is therefore customary to estimate the parameters for galaxy i from its observed luminosity, L_i, using empirical scaling relations of the form (see Natarajan and Kneib 1997)

$$\sigma_i = M_0 \left(\frac{L_i}{L_0} \right)^{\eta_\sigma} \tag{7.10a}$$

[2]If the galaxies are allowed to be elliptical, each one has an ellipticity and position angle as well. Those are often fixed to match the shape of the visible galaxy.

$$s_i = s_0 \left(\frac{L_i}{L_0} \right)^{\eta_s} \tag{7.10b}$$

$$a_i = a_0 \left(\frac{L_i}{L_0} \right)^{\eta_a} . \tag{7.10c}$$

With this approach lens models need to fit just the three normalization parameters M_0, s_0, and a_0; it might be possible to fit for the power law exponents as well, although it is more common to take those from other observations. The simplest models let all the galaxies be singular or nearly so (meaning s_0 is fixed to be zero or small) and use typical values $(\eta_\sigma, \eta_a) = (0.25, 0.5)$, $(0.35, 0.5)$, or $(0.25, 0.4)$ (e.g., Natarajan and Kneib 1997; Grillo et al. 2015; Raney et al. 2018). The empirical scaling relations have intrinsic scatter, which has recently been incorporated into the error analysis for lens models (Raney et al. 2018).

All of the preceding discussion applies to models in which the mass distribution is built from components that have simple functional forms described by a modest number of parameters. While such "parametric" models can be made quite sophisticated, ultimately they are only as flexible as the functions out of which they are built. Some researchers have therefore introduced "nonparametric" methods to allow more freedom in the mass distribution. For example, one might divide the lens plane into a large number of pixels and then attempt to constrain the mass in each pixel. A different approach would be to pixelate the lens potential. In fact, the same idea can be applied to any choice of basis functions. A variety of implementations can be found in the literature (e.g., Abdelsalam et al. 1998; Bradač et al. 2005; Diego et al. 2005, 2007; Liesenborgs et al. 2006; Saha et al. 2006; Merten et al. 2015; Merten 2016). Some authors have developed hybrid methods that use parametric models for the galaxies and nonparametric methods for the large-scale mass distribution (e.g., Diego et al. 2015a). We will see examples of both nonparametric and hybrid lens models in Sect. 7.3.

7.2 Weak Lensing by Clusters

If a source lies outside the caustics, it will not be multiply imaged but it can still be distorted by lensing. With a single source, it is essentially impossible to determine how much of the galaxy's observed shape is caused by lensing shear and how much is due to its intrinsic ellipticity. However, with a large sample of sources, it becomes possible to detect shear through its *correlated* effects on the observed shapes. Measuring shear then provides valuable information about the mass distribution, characterized by the convergence. In this section we develop the basic theory of weak lensing by examining the relationship between shear and convergence.

7.2.1 General Relation Between Shear and Convergence

According to Eqs. (4.22) and (4.52), the convergence κ and the *complex shear* $\gamma \equiv \gamma_+ + i\gamma_\times$ are related to the lens potential by

$$\kappa = \frac{1}{2}(\partial_x^2 + \partial_y^2)\psi \tag{7.11a}$$

$$\gamma = \frac{1}{2}(\partial_x + i\partial_y)^2\psi\,. \tag{7.11b}$$

It then follows that

$$\begin{aligned}\nabla^2\kappa(x,y) &= \frac{1}{2}(\partial_x^2 + \partial_y^2)^2\psi(x,y)\\ &= \frac{1}{2}(\partial_x - i\partial_y)^2(\partial_x + i\partial_y)^2\psi(x,y)\\ &= (\partial_x - i\partial_y)^2\gamma(x,y)\,.\end{aligned} \tag{7.12}$$

For a given shear map $\gamma(x,y)$, it is straightforward to obtain the corresponding convergence field $\kappa(x,y)$. Based on our discussion of Green's functions in Sect. 4.1.1, we know that the solution of $\nabla^2 G(x,y) = \delta(x)\delta(y)$ is

$$G(x,y) = \frac{1}{2\pi}\ln\sqrt{x^2+y^2}\,. \tag{7.13}$$

Hence, the convergence in Eq. (7.12) is given by

$$\begin{aligned}\kappa(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-x',y-y')(\partial_x' - i\partial_y')^2\gamma(x',y')\mathrm{d}x'\mathrm{d}y' \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ln\sqrt{(x-x')^2+(y-y')^2}(\partial_x' - i\partial_y')^2\gamma(x',y')\,\mathrm{d}x'\mathrm{d}y'\,.\end{aligned} \tag{7.14}$$

Integration by parts allows us to rewrite this as (Kaiser and Squires 1993; Umetsu 2011)

$$\boxed{\kappa(x,y) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\gamma(x',y')}{[(x-x')+i(y-y')]^2}\,\mathrm{d}x'\mathrm{d}y'}\,. \tag{7.15}$$

Since shear is measured using the shapes of images of background galaxies (see Sect. 7.2.2), it is known only at certain locations. Thus the integral must be approximated by a sum, which means that the accuracy of the inferred convergence map depends on both the quantity and quality of shear measurements. It is also important to realize that the background galaxies have different redshifts and hence

different values of the D_{ls}/D_s factor, so the inferred convergence is effectively averaged over the distribution of source redshifts.

A similar Green's function analysis makes it possible to write the shear as an integral over the convergence (see Problem 7.1):

$$\gamma(x, y) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\kappa(x', y')}{[(x - x') - i(y - y')]^2} \, \mathrm{d}x' \mathrm{d}y' \,. \tag{7.16}$$

This expression will be useful in Sect. 7.2.3.

7.2.2 *Inferring Shear from Galaxy Shapes*

Since shear behaves as a quadrupole, we can estimate it from the quadrupole moment of the light distribution of a galaxy. The quadrupole-moment tensor is defined to be

$$Q_{ij} = \frac{\int_{\mathbb{R}^2} (\theta_i - \bar{\theta}_i)(\theta_j - \bar{\theta}_j) I(\boldsymbol{\theta}) \mathrm{d}^2\theta}{\int_{\mathbb{R}^2} I(\boldsymbol{\theta}) \mathrm{d}^2\theta} \,, \tag{7.17}$$

where $I(\boldsymbol{\theta})$ is the observed surface brightness of the galaxy and the image centroid is at

$$\bar{\boldsymbol{\theta}} = \frac{\int_{\mathbb{R}^2} \boldsymbol{\theta} I(\boldsymbol{\theta}) \mathrm{d}^2\theta}{\int_{\mathbb{R}^2} I(\boldsymbol{\theta}) \mathrm{d}^2\theta} \,. \tag{7.18}$$

An alternative perspective is offered in Problem 7.3. We define the complex ellipticity ϵ by

$$\epsilon = \frac{Q_{11} - Q_{22} + 2i\, Q_{12}}{Q_{11} + Q_{22}} \,. \tag{7.19}$$

A galaxy whose projected image is an ellipse with axis ratio $q < 1$ has ellipticity

$$\epsilon = \frac{1 - q^2}{1 + q^2} e^{2i\phi} \,, \tag{7.20}$$

where ϕ denotes the angle between the major axis and the horizontal. Note that ϵ is invariant under the rotation $\phi \to \phi + \pi$. The definition of ϵ ensures that this remains true even for non-elliptical images.

The complex ellipticity of the unlensed source,

$$\epsilon^{(s)} = \frac{Q_{11}^{(s)} - Q_{22}^{(s)} + 2i\, Q_{12}^{(s)}}{Q_{11}^{(s)} + Q_{22}^{(s)}} \,, \tag{7.21}$$

can be related to that of the lensed image through the amplification tensor A from Eq. (4.41). In terms of unlensed source coordinates $\boldsymbol{\beta}$, the quadrupole moment tensor is given by

$$Q_{ij}^{(s)} = \frac{\int_{\mathbb{R}^2}(\beta_i - \bar{\beta}_i)(\beta_j - \bar{\beta}_j) I^{(s)}(\boldsymbol{\beta}) \mathrm{d}^2\beta}{\int_{\mathbb{R}^2} I^{(s)}(\boldsymbol{\beta}) \mathrm{d}^2\beta}\,, \tag{7.22}$$

where $I^{(s)}(\boldsymbol{\beta})$ is the *intrinsic* surface brightness of the source, and the centroid position is

$$\bar{\boldsymbol{\beta}} = \frac{\int_{\mathbb{R}^2} \boldsymbol{\beta} I^{(s)}(\boldsymbol{\beta}) \mathrm{d}^2\beta}{\int_{\mathbb{R}^2} I^{(s)}(\boldsymbol{\beta}) \mathrm{d}^2\beta}\,. \tag{7.23}$$

If the source and image are small enough that we may regard A as constant, then the lens equation reduces to the linear form

$$\boldsymbol{\beta} - \bar{\boldsymbol{\beta}} = \mathsf{A}^{-1}(\bar{\boldsymbol{\theta}})[\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}]\,. \tag{7.24}$$

In this case, it can be shown that (see Problem 7.4)

$$\mathsf{Q}^{(s)} = \mathsf{A}^{-1}\mathsf{Q}\mathsf{A}^{-1}. \tag{7.25}$$

This expression allows us to write the complex ellipticity $\epsilon^{(s)}$ of the source (7.21) in terms of the image ellipticity ϵ and the **reduced shear** $g \equiv \gamma/(1-\kappa)$ as

$$\epsilon^{(s)} \;=\; \frac{\epsilon - 2g + g^2\epsilon^*}{1 + |g|^2 - 2\mathrm{Re}(g\epsilon^*)} \;\approx\; \epsilon - 2g \tag{7.26}$$

(e.g., Bartelmann and Schneider 2001), where we remind the reader that an asterisk denotes complex conjugation. In weak lensing, we may typically neglect terms of quadratic or higher order in g and ϵ. Note that $\epsilon^{(s)}$ depends on the convergence κ and complex shear γ only through the ratio g. If the unlensed sources have zero ellipticity on average, we can take the expected value of both sides of Eq. (7.26) to obtain

$$\langle\epsilon\rangle \;\approx\; 2\,\langle g\rangle \;\approx\; 2\,\langle\gamma\rangle\,, \tag{7.27}$$

where the second step follows if the convergence is small enough. Kaiser et al. (1995, commonly referred to as KSB) developed a method for inferring shear values from measured galaxy ellipticities. While the analytic formulation presented here captures the main astrophysical ideas, turning noisy measurements into a clean shear map is not straightforward. In addition, there may be physical effects that invalidate the assumption that source galaxies have uncorrelated ellipticities with $\langle\epsilon^{(s)}\rangle = 0$ (see the review by Troxel and Ishak 2015).

7.2.3 Convergence and Tangential Shear Profiles

Relative to position vector $\boldsymbol{\theta} = [\, R\cos\phi \;\; R\sin\phi \,]^T$, we define

$$\Gamma_+(R,\phi) = -\gamma_+(R,\phi)\cos 2\phi - \gamma_\times(R,\phi)\sin 2\phi \tag{7.28a}$$

$$\Gamma_\times(R,\phi) = \gamma_+(R,\phi)\sin 2\phi - \gamma_\times(R,\phi)\cos 2\phi\,. \tag{7.28b}$$

The Γ_+ component corresponds to tangential shear relative to $\boldsymbol{\theta}$ when it is positive and to radial shear when it is negative; the $\Gamma_\times$ component corresponds to shear at an angle of $\pm\pi/4$ relative to $\boldsymbol{\theta}$. Returning to Eq. (7.16) and extracting the real and imaginary components, we can write

$$\gamma_+(\boldsymbol{\theta}) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\kappa(x',y')\frac{(x-x')^2-(y-y')^2}{[(x-x')^2+(y-y')^2]^2}\mathrm{d}x'\mathrm{d}y' \tag{7.29a}$$

$$\gamma_\times(\boldsymbol{\theta}) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\kappa(x',y')\frac{2(x-x')(y-y')}{[(x-x')^2+(y-y')^2]^2}\mathrm{d}x'\mathrm{d}y'\,. \tag{7.29b}$$

Converting the integrals to polar coordinates yields

$$\gamma_+(R,\phi) = -\frac{1}{\pi}\int_0^{\infty}\int_0^{2\pi}R'\kappa(R',\phi') \times \frac{R^2\cos(2\phi)+R'^2\cos(2\phi')-2RR'\cos(\phi+\phi')}{[R^2+R'^2-2RR'\cos(\phi-\phi')]^2}\mathrm{d}\phi'\mathrm{d}R' \tag{7.30a}$$

$$\gamma_\times(R,\phi) = -\frac{1}{\pi}\int_0^{\infty}\int_0^{2\pi}R'\kappa(R',\phi') \times \frac{R^2\sin(2\phi)+R'^2\sin(2\phi')-2RR'\sin(\phi+\phi')}{[R^2+R'^2-2RR'\cos(\phi-\phi')]^2}\mathrm{d}\phi'\mathrm{d}R'\,. \tag{7.30b}$$

Equations (7.28) then become:

$$\Gamma_+(R,\phi) = \frac{1}{\pi}\int_0^{\infty}\int_0^{2\pi}R'\kappa(R',\phi') \times \frac{R^2+R'^2\cos[2(\phi-\phi')]-2RR'\cos(\phi-\phi')}{[R^2+R'^2-2RR'\cos(\phi-\phi')]^2}\mathrm{d}\phi'\mathrm{d}R' \tag{7.31a}$$

$$\Gamma_\times(R,\phi) = -\frac{1}{\pi}\int_0^\infty \int_0^{2\pi} R'\kappa(R',\phi') \times \frac{R'^2\sin[2(\phi-\phi')] - 2RR'\sin(\phi-\phi')}{[R^2+R'^2-2RR'\cos(\phi-\phi')]^2}\mathrm{d}\phi'\mathrm{d}R' \,. \tag{7.31b}$$

If the lens is axisymmetric, the convergence is independent of ϕ, i.e., $\kappa = \kappa(R)$. Equations (7.31) then take the form

$$\Gamma_+(R) = \frac{1}{\pi}\int_0^\infty \kappa(R')W(R,R')R'\mathrm{d}R' \tag{7.32a}$$

$$\Gamma_\times(R) = \frac{1}{\pi}\int_0^\infty \kappa(R')V(R,R')R'\mathrm{d}R' \,. \tag{7.32b}$$

The integration kernels are given by

$$\begin{aligned} W(R,R') &= \int_0^{2\pi} \frac{R^2+R'^2\cos[2(\phi-\phi')] - 2RR'\cos(\phi-\phi')}{[R^2+R'^2-2RR'\cos(\phi-\phi')]^2}\mathrm{d}\phi' \\ &= \int_0^{2\pi} \frac{R^2+R'^2\cos 2\varphi - 2RR'\cos\varphi}{[R^2+R'^2-2RR'\cos\varphi]^2}\mathrm{d}\varphi \end{aligned} \tag{7.33a}$$

$$\begin{aligned} V(R,R') &= -\int_0^{2\pi} \frac{R'^2\sin[2(\phi-\phi')] - 2RR'\sin(\phi-\phi')}{[R^2+R'^2-2RR'\cos(\phi-\phi')]^2}\mathrm{d}\phi' \\ &= -\int_{-\pi}^{\pi} \frac{R'^2\sin 2\varphi - 2RR'\sin\varphi}{[R^2+R'^2-2RR'\cos\varphi]^2}\mathrm{d}\varphi \\ &= 0 \,, \end{aligned} \tag{7.33b}$$

where $\varphi \equiv \phi - \phi'$. The final line of Eq. (7.33b) follows because the integrand is an odd function of φ. From Eqs. (7.32b) and (7.33b), we see that $\Gamma_\times(R)$ vanishes for an axisymmetric lens. If the convergence is constant, both components of the shear vanish, and so Eq. (7.32a) implies that

$$\int_0^\infty W(R,R')\,R'\,\mathrm{d}R' = 0 \,. \tag{7.34}$$

To derive an explicit expression for the tangential shear, we evaluate the integral in Eq. (7.33a). For $R' \neq R$,

$$W(R,R') = \begin{cases} 2\pi R^{-2} & \text{if } R' < R \\ 0 & \text{if } R' > R \,. \end{cases} \tag{7.35}$$

The integral diverges if $R' = R$, corresponding to a delta function. In other words,

$$W(R, R') = \frac{2\pi}{R^2}\Theta(R - R') + F(R)\delta(R - R')\,, \tag{7.36}$$

where Θ is the Heaviside step function. In order to satisfy Eq. (7.34), we must have $F(R) = -\pi/R$. Thus,

$$W(R, R') = 2\pi\left[\frac{1}{R^2}\Theta(R - R') - \frac{1}{2R}\delta(R - R')\right]. \tag{7.37}$$

Substituting this into Eq. (7.32a), we obtain

$$\Gamma_+(R) = \frac{2}{R^2}\int_0^R \kappa(R')R'\mathrm{d}R' - \kappa(R) \tag{7.38}$$

for the tangential shear of an axisymmetric lens. This expression agrees with the definition of shear in Eq. (2.87).

For a circular background source lensed by a foreground galaxy cluster, the tangential shear Γ_+ is simply the measured ellipticity of the image. If there are multiple noncircular sources, we can use Eq. (7.27) to relate the average ellipticity to the average shear. At an angular distance R from the center of the lens, the azimuthal average of the shear components in Eqs. (7.28) and (7.30) is given by

$$\begin{aligned}
\langle\Gamma_+\rangle_\phi &\equiv \frac{1}{2\pi}\int_0^{2\pi}\Gamma_+(R,\phi)\,\mathrm{d}\phi \\
&= \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^{\infty}\int_0^{2\pi}\kappa(R',\phi')R' \\
&\quad\times\frac{R^2 + R'^2\cos[2(\phi-\phi')] - 2RR'\cos(\phi-\phi')}{[R^2 + R'^2 - 2RR'\cos(\phi-\phi')]^2}\,\mathrm{d}\phi\,\mathrm{d}R'\,\mathrm{d}\phi' \\
&= \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^{\infty}\kappa(R',\phi')W(R,R')R'\,\mathrm{d}R'\,\mathrm{d}\phi' \qquad (7.39a) \\
\langle\Gamma_\times\rangle_\phi &\equiv \frac{1}{2\pi}\int_0^{2\pi}\Gamma_\times(R,\phi)\,\mathrm{d}\phi \\
&= \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^{\infty}\int_0^{2\pi}\kappa(R',\phi')R' \\
&\quad\times\frac{R'^2\sin[2(\phi-\phi')] - 2RR'\sin(\phi-\phi')}{[R^2 + R'^2 - 2RR'\cos(\phi-\phi')]^2}\,\mathrm{d}\phi\,\mathrm{d}R'\,\mathrm{d}\phi' \\
&= \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^{\infty}\kappa(R',\phi')V(R,R')R'\,\mathrm{d}R'\,\mathrm{d}\phi'\,, \qquad (7.39b)
\end{aligned}$$

where we have used the definitions of $W(R, R')$ and $V(R, R')$ from Eqs. (7.33a). With the help of Eqs. (7.33b) and (7.37), we obtain

$$\langle \Gamma_+ \rangle_\phi = \frac{1}{\pi R^2} \int_0^{2\pi} \int_0^R \kappa(R', \phi')\, R'\, \mathrm{d}R'\, \mathrm{d}\phi' - \frac{1}{2\pi} \int_0^{2\pi} \kappa(R, \phi')\, \mathrm{d}\phi' \quad (7.40a)$$

$$\langle \Gamma_\times \rangle_\phi = 0\,. \quad (7.40b)$$

The first term in Eq. (7.40a) represents the average convergence inside a disk of radius R, while the second term is the azimuthally averaged convergence evaluated at radius R. In words, given a circular aperture, the mean tangential shear along the circumference is given by the **density contrast** between the mean convergence within the aperture and the mean convergence along the edge.

Suppose there are N galaxies with ellipticities ϵ_i, for $i = 1, 2, \ldots, N$, within an annulus of inner radius R and thickness ΔR. The average ellipticity is then

$$\langle \epsilon \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \approx 2 \langle \gamma \rangle_\phi\,, \quad (7.41)$$

where the last step follows from Eq. (7.27). From Eqs. (7.28) we have

$$\Gamma \equiv \Gamma_+ + i\Gamma_\times = -e^{-2i\phi}\gamma = e^{i(\pi - 2\phi)}\gamma\,. \quad (7.42)$$

In terms of $\varepsilon \equiv e^{i(\pi-2\phi)}\epsilon$, Eq. (7.27) becomes $\langle \varepsilon \rangle \approx 2\langle \Gamma \rangle$. Since $\langle \Gamma_\times \rangle_\phi = 0$, we can write

$$\langle \mathrm{Re}\, \varepsilon \rangle = 2 \langle \Gamma_+ \rangle_\phi\,, \qquad \langle \mathrm{Im}\, \varepsilon \rangle = 0\,. \quad (7.43)$$

If the second of these relations fails to hold, some effect other than (single-plane) lensing is at work.

7.2.4 *E-modes and B-modes*

To interpret weak lensing quantities, it is useful to draw an analogy with electromagnetism and speak of Γ_+ as an E-mode and $\Gamma_\times$ as a B-mode. To understand this terminology, consider an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$. Gauss's law tells us that $\nabla \cdot \mathbf{B} = 0$. If the magnetic field is independent of time, Faraday's law (4.92) reduces to $\nabla \times \mathbf{E} = \mathbf{0}$. Thus, a vector field whose curl vanishes can be thought of as an E-mode, while a vector field whose divergence vanishes is called a B-mode. This nomenclature is meaningful because any well-behaved vector field can be expressed as the sum of curl-free and divergence-free terms (Helmholtz's

theorem).[3] We have seen that lensing produces only E-modes, at least when there is a single lens plane. Therefore any measurement of a non-zero B-mode indicates either a physical phenomenon other than single-plane lensing or systematic effects in the measurements and analysis.

7.3 Applications

One set of cluster lensing applications examines the physical properties of the clusters themselves (see Kneib and Natarajan (2011) and Hoekstra et al. (2013) for reviews). Many cluster lenses have been studied individually, but several recent projects have worked with samples of a few dozen clusters.

The Sloan Giant Arcs Survey (SGAS, Hennawi et al. 2008) searched for lensed arcs in ground-based imaging from the Sloan Digital Sky Survey (York et al. 2000) and then obtained follow-up observations with telescopes in Arizona and Hawaii. Oguri et al. (2012) performed a joint strong and weak lensing analysis of 28 clusters using parametric models with an NFW halo. The Cluster Lensing and Supernova survey with *Hubble* (CLASH, Postman et al. 2012) observed 25 clusters using the *Hubble Space Telescope* and then performed several lensing analyses. Zitrin et al. (2015) used both strong and weak lensing data to constrain parametric models including an NFW halo along with the cluster member galaxies, and they also considered a type of nonparametric model in which the NFW halo was replaced by a smoothed version of the observed light distribution. Merten et al. (2015) also used strong and weak lensing data to constrain nonparametric models in which the lens potential was reconstructed on an adaptive grid. Umetsu et al. (2014, 2016) added weak lensing data from wider-field observations with the *Subaru Telescope*. In both SGAS and CLASH, one goal was to measure cluster density profiles and constrain the relation between a cluster's "total" mass and its concentration. Since clusters do not have sharp edges, it is necessary to define a radius within which the "total" mass is measured. This is often taken to be the radius r_{200} within which the mean density is 200 times larger than the background critical density of the Universe, which is a reasonable estimate of the region in which the cluster has reached a state of dynamical equilibrium. Then the concentration of an NFW halo is defined as $c_{200} = r_{200}/r_s$ where r_s is the scale radius in Eq. (7.1). SGAS and CLASH found results that were compatible with one another and with theoretical predictions for massive clusters ($\sim 10^{15} M_\odot$), but SGAS found a steeper relation between M_{200} and c_{200} such that less massive clusters ($\sim 10^{14} M_\odot$) were more concentrated than expected.

The MAssive Cluster Survey (MACS, Ebeling et al. 2001; Repp and Ebeling 2018) provided *HST* data for a different set of clusters. A project called "Weighing

[3] Note that the electric field is not, strictly speaking, an E-mode when the magnetic field depends on time, since $\nabla \times \mathbf{E} = -(1/c)\partial \mathbf{B}/\partial t$.

the Giants" then performed a weak lensing analysis of 51 clusters from MACS and other surveys, which served as the basis for constraints on cosmological parameters (von der Linden et al. 2014; Applegate et al. 2014; Mantz et al. 2015). Clusters are good probes of cosmology because their formation is sensitive to interplay between gravitational collapse and cosmic expansion (see Allen et al. 2011), but comparisons between observations and theoretical predictions rely on having well-measured masses. Mantz et al. (2015) used masses from weak lensing to derive constraints from clusters that are complementary to those from observations of the cosmic microwave background (see Chap. 9). Mantz et al. (2016) then compared weak lensing masses with properties of the hot X-ray gas to probe astrophysical processes within clusters. (See Giodini et al. (2013) for a general discussion of scaling relations between masses and X-ray properties of clusters.)

One cluster deserves individual mention. In the system 1E 0657−558, a collision between two sub-components caused the X-ray gas to be separated from the galaxies; one of the gas clouds exhibits a bow shock similar to that produced by a bullet, leading to the moniker "Bullet Cluster" (Markevitch et al. 2002). A lensing analysis reveals that most of the matter in the system is concentrated around the galaxies, not the gas (Clowe et al. 2004, 2006). This is precisely what would be expected if dark matter is a substance that experiences little or no physical interactions other than gravity, and it is difficult to explain in theories that invoke modifications to gravity or dynamics in place of dark matter (but see Angus et al. (2007) for a dissenting view).

Another set of applications focuses on the galaxies behind clusters; in effect, the clusters are used as "cosmic telescopes" whose magnification makes distant galaxies bigger and brighter and hence easier to observe. In addition to CLASH, the *Hubble* Frontier Fields (HFF, Lotz et al. 2017) program provided an extraordinarily deep view of six clusters specially chosen to be good cosmic telescopes, while the Reionization Lensing Cluster Survey (RELICS, Salmon et al. 2017) offered a larger sample of 41 clusters but did not go as deep in each field. These programs have enabled discoveries of some of the most distant known galaxies, at redshifts $z \sim 9$–11 that correspond to times as little as ∼500 Myr after the Big Bang (Zheng et al. 2012; Coe et al. 2013; Bouwens et al. 2014; Zitrin et al. 2014; Infante et al. 2015; Salmon et al. 2018).

In this set of applications, lens modeling is still essential because we need to know how lensing distorts our view of the region behind a cluster. Lensing magnification changes the apparent size and brightness of an image, of course, and it also changes the effective area of a survey: a given angular area in the image plane, $\mathrm{d}\Omega_I$, corresponds to a different area in the source plane, $\mathrm{d}\Omega_S = |\mu|^{-1}\mathrm{d}\Omega_I$. Therefore it must be taken into account not only when studying individual sources but also when studying statistical properties such as the number density of galaxies.

Each cosmic telescope program has included lens modeling, but the HFF program went further and invited different teams to provide lens models for

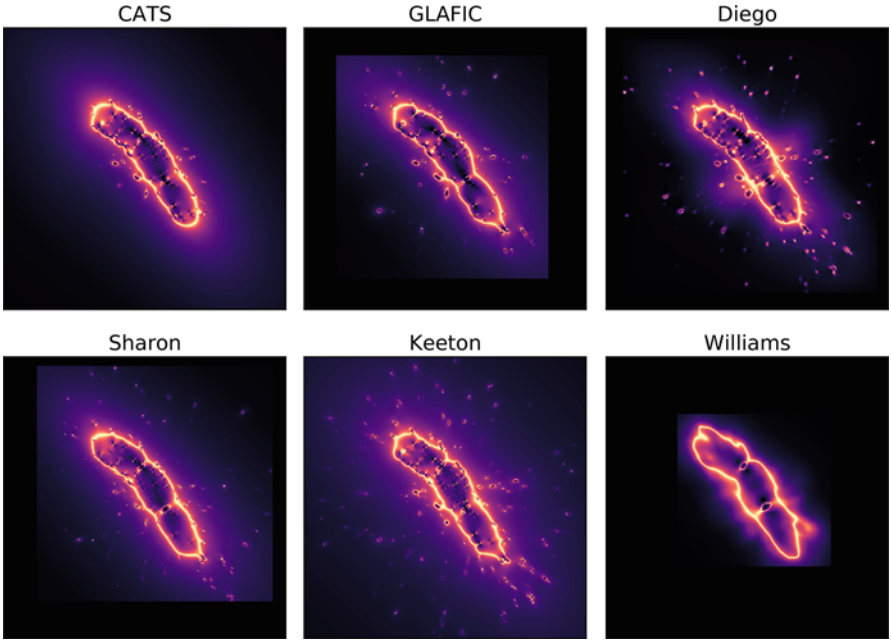


Fig. 7.4 Examples of magnification maps for a source at redshift $z = 9$ in the *Hubble* Frontier Field MACS J0416.1−2403 provided by different lens modeling teams. The field itself is shown in Fig. 7.3. The models in the left and middle columns are parametric. In the right column, the Williams model is nonparametric, while the Diego model is hybrid. The lens models are publicly available at https://archive.stsci.edu/prepds/frontier/lensmodels (In the GLAFIC, Sharon, and Williams panels it is apparent that the region used to compute the magnification is smaller than the region shown here)

the same fields, which are now publicly available.[4] The idea was to compare different approaches to lens modeling as a way to quantify systematic uncertainties. Figure 7.4 shows some of the resulting magnification maps for the cluster MACS J0416.1−2403. The models labeled CATS, Sharon, GLAFIC, and Keeton are all parametric; the CATS and Sharon teams use the same modeling software but with different input data and assumptions; while the GLAFIC and Keeton teams use their own software. The Williams model is fully nonparametric, while the Diego model is hybrid.

Whether the maps are considered to be similar or different is a matter of perspective. On the one hand, it seems impressive to find broad agreement about something we cannot observe directly (namely, the mass distribution and associated gravitational field), even between models that are quite diverse in their methods and assumptions (e.g., parametric vs. nonparametric). On the other hand, broad agree-

[4] https://archive.stsci.edu/prepds/frontier/lensmodels.

ment may not be good enough when it comes to placing quantitative constraints on the abundance of high-redshift galaxies; detailed studies suggest that the differences between teams are generally larger than the uncertainties quoted by the teams, which implies that lens modeling uncertainties are being underestimated (see Priewe et al. 2017; Raney et al. 2018). Even accounting for the modeling uncertainties, however, cluster lensing is already playing a significant role in studies of the earliest stages of galaxy formation in the Universe (e.g., Oesch et al. 2018; Atek et al. 2018; and references therein), and it will continue to do so: 2 of the 13 programs selected for the Director's Discretionary Early Release Science Program with the *James Webb Space Telescope* involve using clusters as cosmic telescopes.[5]

The cluster SDSS J1110+6459 (which was discovered as part of SGAS) has already provided a particularly high-resolution view of star formation in a galaxy at redshift $z = 2.481$. The galaxy is lensed into a long arc, which is shown by *HST* imaging to contain two dozen clumps of star formation. Johnson et al. (2017a) performed a strong lensing analysis using a parametric model along with nonparametric perturbations to improve the fit and then reconstructed what the source looked like before it was lensed. They found that the star-forming regions are as small as 30–50 pc, which is far smaller than the scales typically probed at such redshifts. This system illustrates how lensing has the potential to revolutionize our understanding of star formation in distant galaxies (see also Johnson et al. 2017b; Rigby et al. 2017).

Problems

7.1 Use a Green's function analysis like that in Sect. 7.2.1 to derive Eq. (7.16) for the shear as an integral over convergence.

7.2 Notice from Eq. (7.15) that the convergence, κ, is the convolution of the complex shear, γ, with the function

$$h(x, y) = -\frac{1}{\pi}\frac{1}{(x + iy)^2} . \tag{7.44}$$

(a) Starting from Eq. (7.11), show that the Fourier transforms of κ and γ, denoted by $\tilde{\kappa}$ and $\tilde{\gamma}$, are related by

$$\tilde{\kappa}(l_x, l_y) = f(l_x, l_y)\tilde{\gamma}(l_x, l_y) , \tag{7.45}$$

and find $f(l_x, l_y)$.

[5] https://jwst.stsci.edu/observing-programs/approved-ers-programs.

(b) Show that $f(l_x, l_y)$ is the Fourier transform of $h(x, y)$, and use the convolution theorem to derive Eq. (7.15).
(c) Use the fact that differentiation of a function corresponds to multiplication of its Fourier transform by an appropriate factor to derive Eq. (7.12).

7.3 An alternative definition of the quadrupole-moment tensor (7.17) of an image with surface brightness $I(\boldsymbol{\theta})$ is

$$Q_{ij} = \frac{\int_{\mathbb{R}^2}(\theta_i - \bar{\theta}_i)(\theta_j - \bar{\theta}_j) I(\boldsymbol{\theta}) W(\boldsymbol{\theta}) \mathrm{d}^2\theta}{\int_{\mathbb{R}^2} I(\boldsymbol{\theta}) W(\boldsymbol{\theta}) \mathrm{d}^2\theta}, \tag{7.46}$$

where we take the centroid position to be $\bar{\boldsymbol{\theta}} = 0$.

(a) Find an appropriate weighting function $W(\boldsymbol{\theta})$ for each of the following cases:

- Suppose that there is a minimum detectable surface brightness $I_{\min}$.
- Suppose that measuring the surface brightness is only possible inside of an ellipse with semi-major axis a and semi-minor axis b.
- Suppose that the surface brightness can be most easily measured within the ellipse mentioned above, but with a weaker signal farther out.

(b) Consider the case that $I(\boldsymbol{\theta})$ is constant within an ellipse with semi-axes a and b ($a > b$). Assuming that the denominator in Eq. (7.46) is fixed, compute Q_{ij} for each of the weighting functions in part (7.3). How sensitive is the complex ellipticity (7.19) to the choice of $W(\boldsymbol{\theta})$?

7.4 Use conservation of surface brightness and the properties of the amplification tensor to show that the quadrupole moment tensors of an unlensed source and its lensed image are related by Eq. (7.25) in the linear approximation (7.24).

7.5 Consider a massive, spherical NFW halo with total mass $M_{200} = 3 \times 10^{15}\, M_\odot$ and radius $r_{200} = 2.5$ Mpc. It has concentration $c_{200} = r_{200}/r_s = 3.5$. The lens has redshift $z_l = 0.38$ and angular diameter distance $D_l = 1066$ Mpc. Let's consider sources with $D_{ls}/D_s = 0.5$ (for weak lensing this reflects an average over source redshifts).

(a) What is the Einstein radius? (Solve for it graphically or numerically.)
(b) Derive and plot the shear profile.
(c) Suppose you want to measure the shear at a distance of 10 arcmin from the center of the cluster. We said in Sect. 7.2 that shear can be estimated from the average ellipticity of images. However, in practice there is "shape noise" because the number of sources is finite and their intrinsic shapes may not average to zero. Suppose the source galaxy shapes are drawn from a Gaussian distribution with mean zero and standard deviation $\sigma_\epsilon \approx 0.3$. How many source galaxies would you need in order to detect that γ at 10 arcmin is different from 0 at more than 3σ?

7.6 Suppose a spherical NFW lens has two sources directly behind it, producing two concentric Einstein rings. The lens has redshift $z = 0.38$ and distance $D_l = 1066$ Mpc. Source #1 has $z_1 = 1.0$, $D_1 = 1654$ Mpc, and $D_{l1} = 919$ Mpc. Source #2 has $z_2 = 2.0$, $D_2 = 1742$ Mpc, and $D_{l2} = 1251$ Mpc.

(a) Using the cluster from Problem 7.5, find the radii of the two rings.
(b) Suppose instead that you observe two concentric rings with radii $8.7''$ and $16.9''$. Find the masses enclosed by the two rings. If you have only ring #1, can you determine the parameters of the NFW lens (i.e., ρ_s and r_s)? What if you have both rings?

References

Abdelsalam, H. M., Saha, P., & Williams, L. L. R. (1998). Non-parametric reconstruction of cluster mass distribution from strong lensing - Modelling Abell 370. *Monthly Notices of the Royal Astronomical Society, 294*, 734.

Allen, S. W., Evrard, A. E., & Mantz, A. B. (2011). Cosmological parameters from observations of galaxy clusters. *Annual Review of Astronomy and Astrophysics, 49*, 409.

Angus, G. W., Shan, H. Y., Zhao, H. S., & Famaey, B. (2007). On the proof of dark matter, the law of gravity, and the mass of neutrinos. *The Astrophysical Journal Letters, 654*, L13.

Applegate, D. E., von der Linden, A., Kelly, P. L., Allen, M. T., Allen, S. W., Burchat, P. R., et al. (2014). Weighing the Giants - III. Methods and measurements of accurate galaxy cluster weak-lensing masses. *Monthly Notices of the Royal Astronomical Society, 439*, 48.

Atek, H., Richard, J., Kneib, J.-P., & Schaerer, D. (2018). The extreme faint end of the UV luminosity function at $z \sim 6$ through Gravitational Telescopes: A comprehensive assessment of strong lensing uncertainties. *Monthly Notices of the Royal Astronomical Society, 479*, 5184–5195.

Bartelmann, M. (1996). Arcs from a universal dark-matter halo profile. *Astronomy and Astrophysics, 313*, 697.

Bartelmann, M., & Schneider, P. (2001). Weak gravitational lensing. *Physics Reports, 340*, 291.

Bouwens, R. J., Bradley, L., Zitrin, A., Coe, D., Franx, M., Zheng, W., et al. (2014). A census of star-forming galaxies in the $z \sim 9$–10 universe based on HST+Spitzer Observations over 19 clash clusters: Three candidate $z \sim 9$–10 galaxies and improved constraints on the star formation rate density at $z \sim 9.2$. *The Astrophysical Journal, 795*, 126.

Bradač, M., Schneider, P., Lombardi, M., & Erben, T. (2005). Strong and weak lensing united. I. The combined strong and weak lensing cluster mass reconstruction method. *Astronomy and Astrophysics, 437*, 39.

Broadhurst, T., Benítez, N., Coe, D., Sharon, K., Zekser, K., White, R., et al. (2005). Strong-lensing analysis of A1689 from deep advanced camera images. *The Astrophysical Journal, 621*, 53.

Caminha, G. B., Grillo, C., Rosati, P., Balestra, I., Mercurio, A., Vanzella, E., et al. (2017). A refined mass distribution of the cluster MACS J0416.1−2403 from a new large set of spectroscopic multiply lensed sources. *Astronomy and Astrophysics, 600*, A90.

Christensen, L., Richard, J., Hjorth, J., Milvang-Jensen, B., Laursen, P., Limousin, M., et al. (2012). The low-mass end of the fundamental relation for gravitationally lensed star-forming galaxies at $1 < z < 6$. *Monthly Notices of the Royal Astronomical Society, 427*, 1953.

Clowe, D., Bradač, M., Gonzalez, A. H., Markevitch, M., Randall, S. W., Jones, C., et al. (2006). A direct empirical proof of the existence of dark matter. *The Astrophysical Journal Letters, 648*, L109.

Clowe, D., Gonzalez, A., & Markevitch, M. (2004). Weak-lensing mass reconstruction of the interacting cluster 1E 0657-558: Direct evidence for the existence of dark matter. *The Astrophysical Journal, 604*, 596.

Coe, D., Zitrin, A., Carrasco, M., Shu, X., Zheng, W., Postman, M., et al. (2013). CLASH: Three strongly lensed images of a candidate $z \sim 11$ galaxy. *The Astrophysical Journal, 762*, 32.

Diego, J. M., Broadhurst, T., Benitez, N., Umetsu, K., Coe, D., Sendra, I., et al. (2015a). A free-form lensing grid solution for A1689 with new multiple images. *Monthly Notices of the Royal Astronomical Society, 446*, 683.

Diego, J. M., Broadhurst, T., Molnar, S. M., Lam, D., & Lim, J. (2015b). Free-form lensing implications for the collision of dark matter and gas in the frontier fields cluster MACS J0416.1-2403. *Monthly Notices of the Royal Astronomical Society, 447*, 3130.

Diego, J. M., Protopapas, P., Sandvik, H. B., & Tegmark, M. (2005). Non-parametric inversion of strong lensing systems. *Monthly Notices of the Royal Astronomical Society, 360*, 477.

Diego, J. M., Tegmark, M., Protopapas, P., & Sandvik, H. B. (2007). Combined reconstruction of weak and strong lensing data with WSLAP. *Monthly Notices of the Royal Astronomical Society, 375*, 958.

Ebeling, H., Edge, A. C., & Henry, J. P. (2001). MACS: A quest for the most massive galaxy clusters in the universe. *The Astrophysical Journal, 553*, 668.

Giodini, S., Lovisari, L., Pointecouteau, E., Ettori, S., Reiprich, T. H., & Hoekstra, H. (2013). Scaling relations for galaxy clusters: Properties and evolution. *Space Science Reviews, 177*, 247.

Grillo, C., Suyu, S. H., Rosati, P., Mercurio, A., Balestra, I., Munari, E., et al. (2015). CLASH-VLT: Insights on the mass substructures in the frontier fields cluster MACS J0416.1-2403 through accurate strong lens modeling. *The Astrophysical Journal, 800*, 38.

Hennawi, J. F., Gladders, M. D., Oguri, M., Dalal, N., Koester, B., Natarajan, P., et al. (2008). A New Survey for giant arcs. *The Astronomical Journal, 135*, 664.

Hoekstra, H., Bartelmann, M., Dahle, H., Israel, H., Limousin, M., & Meneghetti, M. (2013). Masses of galaxy clusters from gravitational lensing. *Space Science Reviews, 177*, 75.

Infante, L., Zheng, W., Laporte, N., Troncoso Iribarren, P., Molino, A., Diego, J. M., et al. (2015). Young galaxy candidates in the Hubble Frontier Fields. II. MACS J0416-2403. *The Astrophysical Journal, 815*, 18.

Jauzac, M., Clément, B., Limousin, M., Richard, J., Jullo, E., Ebeling, H., et al. (2014). Hubble Frontier Fields: a high-precision strong-lensing analysis of galaxy cluster MACSJ0416.1-2403 using $\sim$200 multiple images. *Monthly Notices of the Royal Astronomical Society, 443*, 1549.

Johnson, T. L., Rigby, J. R., Sharon, K., Gladders, M. D., Florian, M., Bayliss, M. B., et al. (2017a). Star formation at $z = 2.481$ in the lensed galaxy SDSS J1110+6459: Star formation down to 30 pc scales. *The Astrophysical Journal Letters, 843*, L21.

Johnson, T. L., & Sharon, K. (2016). The systematics of strong lens modeling quantified: The effects of constraint selection and redshift information on magnification, mass, and multiple image predictability. *The Astrophysical Journal, 832*, 82.

Johnson, T. L., Sharon, K., Gladders, M. D., Rigby, J. R., Bayliss, M. B., Wuyts, E., et al. (2017b). Star formation at $z = 2.481$ in the lensed galaxy SDSS J1110+6459. I. Lens modeling and source reconstruction. *The Astrophysical Journal, 843*, 78.

Kaiser, N., & Squires, G. (1993). Mapping the dark matter with weak gravitational lensing. *The Astrophysical Journal, 404*, 441.

Kaiser, N., Squires, G., & Broadhurst, T. (1995). A method for weak lensing observations. *The Astrophysical Journal, 449*, 460.

Keeton, C. R. (2001). A catalog of mass models for gravitational lensing. astro-ph/0102341.

Kneib, J.-P., & Natarajan, P. (2011). Cluster lenses. *The Astronomy and Astrophysics Review, 19*, 47

Liesenborgs, J., De Rijcke, S., & Dejonghe, H. (2006). A genetic algorithm for the non-parametric inversion of strong lensing systems. *Monthly Notices of the Royal Astronomical Society, 367*, 1209.

Lotz, J. M., Koekemoer, A., Coe, D., Grogin, N., Capak, P., Mack, J., et al. (2017). The Frontier fields: Survey design and initial results. *The Astrophysical Journal, 837*, 97.

Mantz, A. B., Allen, S. W., Morris, R. G., von der Linden, A., Applegate, D. E., Kelly, P. L., et al. (2016). Weighing the giants- V. Galaxy cluster scaling relations. *Monthly Notices of the Royal Astronomical Society, 463*, 3582.

Mantz, A. B., von der Linden, A., Allen, S. W., Applegate, D. E., Kelly, P. L., Morris, R. G., et al. (2015). Weighing the giants - IV. Cosmology and neutrino mass. *Monthly Notices of the Royal Astronomical Society, 446*, 2205.

Markevitch, M., Gonzalez, A. H., David, L., Vikhlinin, A., Murray, S., Forman, W., et al. (2002). A textbook example of a bow shock in the merging galaxy cluster 1E 0657−56. *The Astrophysical Journal Letters, 567*, L27.

Merten, J. (2016). Mesh-free free-form lensing - I. Methodology and application to mass reconstruction. *Monthly Notices of the Royal Astronomical Society, 461*, 2328.

Merten, J., Meneghetti, M., Postman, M., Umetsu, K., Zitrin, A., Medezinski, E., et al. (2015). CLASH: The concentration-mass relation of galaxy clusters. *The Astrophysical Journal, 806*, 4.

Natarajan, P., & Kneib, J.-P. (1997). Lensing by galaxy haloes in clusters of galaxies. *Monthly Notices of the Royal Astronomical Society, 287*, 833.

Navarro, J. F., Frenk, C. S., & White, S. D. M. (1995). Simulations of X-ray clusters. *Monthly Notices of the Royal Astronomical Society, 275*, 720.

Navarro, J. F., Frenk, C. S., & White, S. D. M. (1996). The structure of cold dark matter halos. *The Astrophysical Journal, 462*, 563.

Navarro, J. F., Frenk, C. S., & White, S. D. M. (1997). A universal density profile from hierarchical clustering. *The Astrophysical Journal, 490*, 493.

Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I., & Stefanon, M. (2018). The dearth of $z \sim 10$ galaxies in all HST legacy fields—The rapid evolution of the galaxy population in the First 500 Myr. *The Astrophysical Journal, 855*, 105.

Oguri, M., Bayliss, M. B., Dahle, H., Sharon, K., Gladders, M. D., Natarajan, P., et al. (2012). Combined strong and weak lensing analysis of 28 clusters from the Sloan Giant Arcs Survey. *Monthly Notices of the Royal Astronomical Society, 420*, 3213.

Postman, M., Coe, D., Benítez, N., Bradley, L., Broadhurst, T., Donahue, M., et al. (2012). The cluster lensing and supernova survey with hubble: An overview. *Astrophysical Journal Supplements, 199*, 25.

Priewe, J., Williams, L. L. R., Liesenborgs, J., Coe, D., & Rodney, S. A. (2017). Lens models under the microscope: comparison of Hubble Frontier Field cluster magnification maps. *Monthly Notices of the Royal Astronomical Society, 465*, 1030.

Raney, C., Keeton, C. R., & Brennan, S. (2018, to be submitted). *Monthly Notices of the Royal Astronomical Society*.

Repp, A., & Ebeling, H. (2018). Science from a glimpse: Hubble SNAPshot observations of massive galaxy clusters. *Monthly Notices of the Royal Astronomical Society, 479*, 844–864.

Rigby, J. R., Johnson, T. L., Sharon, K., Whitaker, K., Gladders, M. D., Florian, M., et al. (2017). Star formation at $z = 2.481$ in the lensed galaxy SDSS J1110+6459. II. What is missed at the normal resolution of the Hubble Space Telescope?. *The Astrophysical Journal, 843*, 79.

Saha, P., Read, J. I., & Williams, L. L. R. (2006). Two strong-lensing clusters confront universal dark matter profiles. *The Astrophysical Journal Letters, 652*, L5.

Salmon, B., Coe, D., Bradley, L., Bouwens, R., Bradac, M., Huang, K.-H., et al. (2017). The Reionization Lensing Cluster Survey (RELICS) and the brightest high-z galaxies. arXiv e-prints, arXiv:1710.08930.

Salmon, B., Coe, D., Bradley, L., Bradač, M., Huang, K.-H., Strait, V., et al. (2018). A candidate $z \sim 10$ galaxy strongly lensed into a spatially resolved arc. *Astrophysical Journal Letters, 864*, L22.

Schmidt, K. B., Treu, T., Brammer, G. B., Bradač, M., Wang, X., Dijkstra, M., et al. (2014). Through the looking GLASS: HST spectroscopy of faint galaxies lensed by the frontier fields cluster MACSJ0717.5+3745. *The Astrophysical Journal Letters, 782*, L36.

Schramm, T. (1990). Realistic elliptical potential wells for gravitational lens models. *Astronomy and Astrophysics, 231*, 19.

Treu, T., Schmidt, K. B., Brammer, G. B., Vulcani, B., Wang, X., Bradač, M., et al. (2015). The Grism Lens-Amplified Survey from Space (GLASS). I. Survey overview and first data release. *The Astrophysical Journal, 812*, 114.

Troxel, M. A., & Ishak, M. (2015). The intrinsic alignment of galaxies and its impact on weak gravitational lensing in an era of precision cosmology. *Physics Reports, 558*, 1.

Umetsu, K. (2011). Cluster weak gravitational lensing. In A. Cavaliere & Y. Rephaeli (Eds.), *Astrophysics of galaxy clusters* (p. 269). Amsterdam: IOS Press.

Umetsu, K., Medezinski, E., Nonino, M., Merten, J., Postman, M., Meneghetti, M., et al. (2014). CLASH: Weak-lensing shear-and-magnification analysis of 20 galaxy clusters. *The Astrophysical Journal, 795*, 163.

Umetsu, K., Zitrin, A., Gruen, D., Merten, J., Donahue, M., & Postman, M. (2016). CLASH: Joint analysis of strong-lensing, weak-lensing shear, and magnification data for 20 galaxy clusters. *The Astrophysical Journal, 821*, 116.

von der Linden, A., Allen, M. T., Applegate, D. E., Kelly, P. L., Allen, S. W., Ebeling, H., et al. (2014). Weighing the giants - I. Weak-lensing masses for 51 massive galaxy clusters: Project overview, data analysis methods and cluster images. *Monthly Notices of the Royal Astronomical Society, 439*, 2.

York, D. G., Adelman, J., Anderson, J. E., Jr., Anderson, S. F., Annis, J., Bahcall, N. A., et al. (2000). The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal, 120*, 1579.

Zheng, W., Postman, M., Zitrin, A., Moustakas, J., Shu, X., Jouvel, S., et al. (2012). A magnified young galaxy from about 500 million years after the Big Bang. *Nature, 489*, 406.

Zitrin, A., Fabris, A., Merten, J., Melchior, P., Meneghetti, M., Koekemoer, A., et al. (2015). Hubble Space Telescope combined strong and weak lensing analysis of the CLASH sample: Mass and magnification models and systematic uncertainties. *The Astrophysical Journal, 801*, 44.

Zitrin, A., Zheng, W., Broadhurst, T., Moustakas, J., Lam, D., Shu, X., et al. (2014). A geometrically supported $z \sim 10$ candidate multiply imaged by the Hubble Frontier Fields cluster A2744. *The Astrophysical Journal Letters, 793*, L12.

Chapter 8
Weak Lensing by Large-Scale Structure

In the remaining chapters, we apply lensing to cases where no single object, such as a galaxy or cluster of galaxies, is responsible for the deflection of light from a background source. Instead, all matter along the line of sight to the source acts collectively as a lens. Comparing large samples of observed galaxy shapes to what we would expect for particular cosmological models allows us to constrain the properties of the Universe itself.

The first step is to describe the statistical properties of density fluctuations in three dimensions, from which we can infer the statistics of the two-dimensional convergence field relevant for lensing. Linear perturbation theory, in which density fluctuations are assumed to be small, provides a framework for lensing by large-scale structure (Sect. 8.1.1), even though a nonlinear treatment is needed for a quantitative comparison between theory and observation. Such considerations cannot be ignored by the researcher, but they do not substantially affect the basic approach to the subject. (See Sect. 8.1.2 for a flavor of what is involved.) For a given model, whether linear or nonlinear, the correlation function and power spectrum encapsulate much of the statistics of structure formation (Sects. 8.2 and 8.3). The lensed versions of these quantities can be inferred from galaxy shapes, which are distorted by the "cosmic shear" from matter in the foreground (Sect. 8.4). Understanding this phenomenon sets the stage for the applications presented in the final section (Sect. 8.5).

8.1 Structure Formation

According to the cosmological principle, the universe is homogeneous, with some specified constant density. However, we know that the universe is only approximately homogeneous and that inflation (or some other mechanism) produced small fluctuations that eventually grew into the large-scale structure observed today. Let

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_8

the average density of the universe at time t be $\bar{\rho}(t)$, and the density at a particular point be $\rho(\mathbf{x}, t)$. It is convenient to define the **density contrast** as

$$\delta(\mathbf{x}, t) \equiv \frac{\rho(\mathbf{x}, t) - \bar{\rho}(t)}{\bar{\rho}(t)} \,. \tag{8.1}$$

8.1.1 Linear Perturbation Theory

The utility of the density contrast becomes clear in the context of **linear perturbation theory**. The idea is to write the density and other relevant quantities as sums of two terms, one of which is much smaller than the other. This allows us to obtain solutions of the governing equations that are accurate to first order in the perturbation terms.

In the present discussion, we consider a universe described by a pressureless fluid, as we expect dark matter to be, subject only to gravity. We assume this fluid to be homogeneous on large scales but with inhomogeneities on small scales that lead to gravitational instabilities. At some point, these initially small fluctuations will become large enough to be outside the linear regime. In that case, our only recourse is to obtain numerical solutions to the equations of general relativistic hydrodynamics. Until then, though, it is valid to describe the growth of structure in terms of a fluid subject to Newtonian gravity. Clusters of galaxies are the densest objects for which this approach may be justified.

An inviscid fluid (i.e., one without viscosity) is described by its density, velocity, and gravitational potential. We write these quantities as

$$\rho(\mathbf{x}, t) = \bar{\rho}(t)[1 + \delta(\mathbf{x}, t)] \tag{8.2a}$$

$$\mathbf{u}(\mathbf{x}, t) = H(t)\mathbf{x} + \mathbf{v}(\mathbf{x}, t) \tag{8.2b}$$

$$\Phi(\mathbf{x}, t) = \Phi_0(\mathbf{x}, t) + \Phi_1(\mathbf{x}, t) \,, \tag{8.2c}$$

where the second term in each equation is assumed to be much smaller than the first in absolute value. Note that the unperturbed density depends only on time, as it must for a homogeneous, isotropic universe. We also remark that the Hubble parameter, $H(t)$, can only be regarded as constant at small redshift.

The perturbed density, velocity, and gravitational potential are determined by the **continuity equation** (conservation of mass), the **Euler equation** (Newton's second law applied to inviscid fluids), and Poisson's equation:

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} + \rho \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \tag{8.3a}$$

$$\frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} = -\nabla\Phi \tag{8.3b}$$

$$\nabla^2\Phi = 4\pi G\rho\,, \tag{8.3c}$$

where the **convective derivative**[1] is defined by

$$\frac{\mathrm{D}}{\mathrm{D}t} \equiv \frac{\partial}{\partial t} + (\mathbf{u}\cdot\nabla) \tag{8.4}$$

(see, e.g., Thompson 2006, p. 2). We neglect the pressure term in Euler equation (8.3b), since it does not apply to dark matter.

In the framework of linear perturbation theory, we require that both the perturbed and unperturbed quantities satisfy the governing equations (8.3). In particular, the unperturbed density, velocity, and gravitational potential satisfy

$$\dot{\bar{\rho}} + 3H\bar{\rho} = 0 \tag{8.5a}$$

$$(\dot{H} + H^2)\mathbf{x} = -\nabla\Phi_0 \tag{8.5b}$$

$$\nabla^2\Phi_0 = 4\pi G\bar{\rho}\,, \tag{8.5c}$$

where an overdot indicates differentiation with respect to time. We can simplify Eq. (8.5b) by writing the Hubble parameter in terms of the cosmic scale factor $H = \dot{a}/a$. We then have

$$\begin{aligned}\dot{H} &= \frac{\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} \\ &= \frac{\ddot{a}}{a} - H^2\,.\end{aligned} \tag{8.6}$$

The unperturbed Euler equation (8.5b) thus reduces to

$$\nabla\Phi_0 = -\frac{\ddot{a}}{a}\mathbf{x}\,. \tag{8.7}$$

Taking the divergence of both sides yields

$$\nabla^2\Phi_0 = -\frac{3\ddot{a}}{a}\,. \tag{8.8}$$

[1] This operator is often called the material derivative, among various other names.

Combining this with Poisson's equation (8.5c) gives

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\bar{\rho}\,, \tag{8.9}$$

which is known as Friedmann's second equation for a pressureless fluid.[2]

The next step is to write down the perturbed fluid equations. We consider the continuity, Euler, and Poisson's equations in turn. We write the continuity equation as

$$\bar{\rho}(1+\delta)\nabla\cdot(H\mathbf{x}+\mathbf{v}) = -\frac{\partial}{\partial t}[\bar{\rho}(1+\delta)] - \bar{\rho}(H\mathbf{x}+\mathbf{v})\cdot\nabla\delta\,. \tag{8.10}$$

Neglecting quadratic terms in the perturbation, we find

$$3H\bar{\rho}(1+\delta) + \bar{\rho}\nabla\cdot\mathbf{v} = -\dot{\bar{\rho}}(1+\delta) - \bar{\rho}\frac{\partial\delta}{\partial t} - H\bar{\rho}\mathbf{x}\cdot\nabla\delta\,. \tag{8.11}$$

Simplifying this with the aid of the unperturbed continuity equation (8.5a), we obtain

$$\boxed{\frac{\partial\delta}{\partial t} + H\mathbf{x}\cdot\nabla\delta + \nabla\cdot\mathbf{v} = 0}\,. \tag{8.12}$$

The perturbed Euler equation can be written as

$$\begin{aligned}\nabla(\Phi_0+\Phi_1) &= -\left[\frac{\partial}{\partial t} + (H\mathbf{x}+\mathbf{v})\cdot\nabla\right](H\mathbf{x}+\mathbf{v}) \\ &= -\left[\dot{H}\mathbf{x} + H^2(\mathbf{x}\cdot\nabla)\mathbf{x} + H(\mathbf{v}\cdot\nabla)\mathbf{x} + \frac{\partial\mathbf{v}}{\partial t} + H(\mathbf{x}\cdot\nabla)\mathbf{v}\right] \\ &= -\left[\dot{H}\mathbf{x} + H^2\mathbf{x} + H\mathbf{v} + \frac{\partial\mathbf{v}}{\partial t} + H(\mathbf{x}\cdot\nabla)\mathbf{v}\right]\,. \end{aligned} \tag{8.13}$$

From the unperturbed Euler equation, this reduces to

$$\boxed{\frac{\partial\mathbf{v}}{\partial t} + H(\mathbf{x}\cdot\nabla)\mathbf{v} + H\mathbf{v} = -\nabla\Phi_1}\,. \tag{8.14}$$

For Poisson's equation we have

$$\nabla^2(\Phi_0+\Phi_1) = 4\pi G\bar{\rho}(1+\delta)\,. \tag{8.15}$$

[2] In the case of non-zero pressure p, $\bar{\rho} \to \bar{\rho} + 3p/c^2$.

This and Eq. (8.5c) imply that the perturbation Φ_1 satisfies

$$\boxed{\nabla^2 \Phi_1 = 4\pi G \bar{\rho} \delta}\,. \tag{8.16}$$

We now transform this equation to a form commonly used in cosmology. Since matter density is inversely proportional to volume, the average density $\bar{\rho}$ can be expressed as $\bar{\rho} = \rho_0 a^{-3}$ in terms of a constant density ρ_0 and the scale factor $a(t)$, where $a = 1$ at present. We define the comoving position vector $\mathbf{r}$ by the relation $\mathbf{x} \equiv a(t)\mathbf{r}$, where $\mathbf{x}$ is the proper position vector. Thus, the comoving gradient $\hat{\boldsymbol{\nabla}}$ satisfies $\hat{\boldsymbol{\nabla}} = a(t)\boldsymbol{\nabla}$. This allows us to rewrite Poisson's equation (8.16) in the form

$$\hat{\nabla}^2 \Phi_1 = 4\pi G \rho_0 a^{-1} \delta\,. \tag{8.17}$$

We now have a set of coupled, linear partial differential equations for δ, $\mathbf{v}$, and Φ_1. To study the time evolution of δ, we begin by taking the divergence of the perturbed Euler equation (8.14):

$$\left(H + \frac{\partial}{\partial t}\right) \boldsymbol{\nabla} \cdot \mathbf{v} + H \boldsymbol{\nabla} \cdot [(\mathbf{x} \cdot \boldsymbol{\nabla})\mathbf{v}] = -\nabla^2 \Phi_1\,. \tag{8.18}$$

Combining the identity

$$\boldsymbol{\nabla} \cdot [(\mathbf{x} \cdot \boldsymbol{\nabla})\mathbf{v}] = \boldsymbol{\nabla} \cdot \mathbf{v} + (\mathbf{x} \cdot \boldsymbol{\nabla}) \boldsymbol{\nabla} \cdot \mathbf{v} \tag{8.19}$$

with the perturbed Poisson's equation gives

$$\left(2H + \frac{\partial}{\partial t} + H\mathbf{x} \cdot \boldsymbol{\nabla}\right) \boldsymbol{\nabla} \cdot \mathbf{v} = -4\pi G \bar{\rho} \delta\,. \tag{8.20}$$

Using the perturbed continuity equation (8.12) to eliminate $\boldsymbol{\nabla} \cdot \mathbf{v}$, we obtain

$$\left(2H + \frac{\partial}{\partial t} + H\mathbf{x} \cdot \boldsymbol{\nabla}\right) \left(\frac{\partial}{\partial t} + H\mathbf{x} \cdot \boldsymbol{\nabla}\right) \delta = 4\pi G \bar{\rho} \delta\,. \tag{8.21}$$

In terms of the convective derivative, this becomes

$$\left(2H + \frac{\mathrm{D}}{\mathrm{D}t} - \mathbf{v} \cdot \boldsymbol{\nabla}\right) \left(\frac{\mathrm{D}}{\mathrm{D}t} - \mathbf{v} \cdot \boldsymbol{\nabla}\right) \delta = 4\pi G \bar{\rho} \delta\,, \tag{8.22}$$

where we have used Eqs. (8.2c) and (8.4). To first order in $\mathbf{v}$ and δ, we find

$$\boxed{\left(\frac{\mathrm{D}^2}{\mathrm{D}t^2} + 2H \frac{\mathrm{D}}{\mathrm{D}t} - 4\pi G \bar{\rho}\right) \delta = 0}\,. \tag{8.23}$$

for a pressureless fluid.

If δ is independent of position over distances much smaller than the Hubble length cH_0^{-1}, the convective derivative reduces to the ordinary time derivative, and Eq. (8.23) takes the form

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{4\pi G\rho_0}{a^3}\delta\,, \tag{8.24}$$

assuming a pressureless, matter-dominated universe. Here we have expressed the Hubble parameter and the mean cosmic density in terms of the scale factor a. According to Eq. (3.127), the scale factor for a matter-dominated universe is

$$a(t) = (6\pi G\rho_0)^{1/3}\, t^{2/3}\,. \tag{8.25}$$

Substituting this expression into Eq. (8.24) yields

$$\ddot{\delta} + \frac{4}{3t}\dot{\delta} = \frac{2}{3t^2}\delta\,. \tag{8.26}$$

We attempt a solution of the form $\delta \propto t^\alpha$, which holds for all t, provided that

$$\alpha^2 + \frac{1}{3}\alpha - \frac{2}{3} = 0\,. \tag{8.27}$$

The solutions are $\alpha = 2/3$ and $\alpha = -1$, so that $\delta(t) = At^{2/3} + Bt^{-1}$, for some constants A and B. We keep only the first term, which grows with time, and discard the second, which decays. In terms of the scale factor we see that $\delta \propto a$.

The density contrast for a flat, *radiation*-dominated universe will be needed in Sect. 8.3, where we derive the matter power spectrum using linear perturbation theory. Relativistic matter, including radiation, obeys the equation of state $p = \rho c^2/3$, for pressure p and density ρ. Incorporating pressure into the fluid equations leads to a modified version of Eq. (8.24), namely,

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{32\pi G\rho_0}{3a^4}\delta \tag{8.28}$$

(cf. Peacock 1999, p. 465), where the scale factor is now

$$a(t) = \left(\frac{32}{3}\pi G\rho_0\right)^{1/4} t^{1/2}\,. \tag{8.29}$$

We follow the same procedure as for the matter-dominated universe and conclude that $\delta(t) = Ct + Dt^{-1}$, where C and D are constants. We again discard the decaying solution, so that $\delta \propto t \propto a^2$.

8.1.2 Beyond the Linear Regime

In the case $\delta(\mathbf{x}, t) \gtrsim 1$, the fluid equations become nonlinear. In general, the solutions can only be obtained numerically. However, it is possible to describe the formation of a single dark-matter halo within an FRW model. Peacock (1999) uses Newtonian gravity to consider spherical collapse in a flat, matter-dominated (Einstein-de Sitter) universe. An overdense region behaves like a closed universe, initially increasing in size due to cosmic expansion, then contracting due to gravitation. When an overdense sphere reaches its maximum radius, linear theory predicts $\delta \sim 1$, while the actual density contrast is $\delta \sim 5$.

A sphere composed entirely of dark matter would collapse to a singularity, since there would be no outward pressure to counteract the pull of gravity. In practice, halos contain ordinary matter as well, thus allowing the kinetic energy of collapse to become random motion. Thermal equilibrium is achieved when the virial theorem, which relates the kinetic and potential energy of a system, applies. In that case, linear theory gives $\delta \sim 1.5$, while the spherical collapse model predicts a value one hundred times greater. For convenience, the radius of a dark matter halo is often defined such that the density there is two hundred times the background, cosmic density. The model we have described here offers only a rough approximation to reality, so quantities of interest, such as the matter power spectrum discussed below, are obtained from numerical simulations.

8.2 Statistics of Density Fluctuations

In principle, one can calculate the density contrast $\delta(\mathbf{x}, t)$ in the linear regime by solving Eq. (8.23) for appropriate boundary and initial conditions. In the absence of this information, we study the behavior of δ statistically. To do this, we identify a volume V large enough so that the average density is a well-defined quantity. We then consider the fluctuations around the average value in terms of the random field δ.

8.2.1 Correlation Function and Power Spectrum

To determine whether the density contrast at $\mathbf{x}$ is correlated to that at $\mathbf{x}'$, we compute the covariance

$$\begin{aligned} C(\mathbf{x}, \mathbf{x}') &\equiv \langle \left[\delta(\mathbf{x}) - \langle\delta(\mathbf{x})\rangle\right] \left[\delta(\mathbf{x}') - \langle\delta(\mathbf{x}')\rangle\right] \rangle \\ &= \langle \delta(\mathbf{x})\delta(\mathbf{x}')\rangle \\ &\equiv \xi_\delta(|\mathbf{x} - \mathbf{x}'|)\,, \end{aligned} \tag{8.30}$$

where the angle brackets denote the average over an ensemble of regions, each with volume V. The second line follows since $\langle\delta\rangle = 0$ by construction. The cosmological principle implies that the correlation function ξ_δ is positive and depends only on $|\mathbf{x} - \mathbf{x}'| \equiv r$. As r increases, $\xi_\delta(r)$ decreases, approaching 0 asymptotically. The size of a typical dark matter halo (overdense region) or void (underdense region) is given by the **correlation length**[3] r_c, where

$$r_c = \frac{\int_0^\infty r^3 \xi_\delta(r) \mathrm{d}r}{\int_0^\infty r^2 \xi_\delta(r) \mathrm{d}r} \,. \tag{8.31}$$

It is often convenient to work in Fourier space. Let $\tilde{\delta}(\mathbf{k})$ be the Fourier transform of $\delta(\mathbf{x})$. The correlation function of the transformed variables is given by

$$\begin{aligned}
\langle \tilde{\delta}(\mathbf{k}) \tilde{\delta}^*(\mathbf{k}') \rangle &= \left\langle \int_{\mathbb{R}^3} \delta(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}} \mathrm{d}^3x \int_{\mathbb{R}^3} \delta(\mathbf{x}') e^{-i\mathbf{k}'\cdot\mathbf{x}'} \mathrm{d}^3x' \right\rangle \\
&= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} e^{i\mathbf{k}\cdot\mathbf{x}} e^{-i\mathbf{k}'\cdot\mathbf{x}'} \langle \delta(\mathbf{x}) \delta(\mathbf{x}') \rangle \, \mathrm{d}^3x \, \mathrm{d}^3x' \\
&= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} e^{i\mathbf{k}\cdot\mathbf{x}} e^{-i\mathbf{k}'\cdot\mathbf{x}'} \xi_\delta(|\mathbf{x} - \mathbf{x}'|) \, \mathrm{d}^3x \, \mathrm{d}^3x' \,.
\end{aligned} \tag{8.32}$$

Here the asterisk denotes complex conjugation, and the correlation function is assumed to fall off rapidly enough to justify integration over the infinite volume $\mathbb{R}^3$. On changing the integration variables from $\mathbf{x}$, $\mathbf{x}'$ to $\mathbf{r} \equiv \mathbf{x} - \mathbf{x}'$, $\mathbf{x}'$, this becomes

$$\begin{aligned}
\langle \tilde{\delta}(\mathbf{k}) \tilde{\delta}^*(\mathbf{k}') \rangle &= \int_{\mathbb{R}^3} e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}'} \mathrm{d}^3x' \int_{\mathbb{R}^3} \xi_\delta(r) e^{i\mathbf{k}'\cdot\mathbf{r}} \mathrm{d}^3r \\
&= (2\pi)^3 \delta^{(3)}(\mathbf{k} - \mathbf{k}') P_\delta(k) \,,
\end{aligned} \tag{8.33}$$

where $r \equiv |\mathbf{r}|$, $k \equiv |\mathbf{k}|$, and $\delta^{(3)}$ denotes the Dirac delta function in three dimensions. The **power spectrum**

$$P_\delta(k) = \int_{\mathbb{R}^3} \xi_\delta(r) e^{i\mathbf{k}\cdot\mathbf{r}} \mathrm{d}^3r \tag{8.34}$$

is the Fourier transform of the correlation function. We refer to it as the *matter* power spectrum in order to distinguish it from other power spectra used in cosmology.

That the power spectrum only depends on the magnitude of $\mathbf{k}$ follows from introducing spherical coordinates (r, θ, ϕ), with $\mathbf{k} \cdot \mathbf{r} = kr\cos\theta$. This leads to

[3] The definition used here is one of many found in the literature, which all predict similar values for a given correlation function.

$$P_\delta(k) = \int_0^{2\pi}\int_0^{\infty}\int_0^{\pi} \xi_\delta(r) e^{ikr\cos\theta} r^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}r \, \mathrm{d}\phi$$
$$= 2\pi \int_0^{\infty} r^2 \xi_\delta(r) \left(\frac{e^{ikr} - e^{-ikr}}{ikr} \right) \mathrm{d}r \,. \tag{8.35}$$

In terms of the spherical Bessel function $j_0(kr) = \sin(kr)/(kr)$, this becomes

$$\boxed{P_\delta(k) = \int_0^{\infty} 4\pi r^2 \xi_\delta(r) j_0(kr) \mathrm{d}r} \,. \tag{8.36}$$

We may also encounter random fields defined in the plane. The convergence, $\kappa(\boldsymbol{\theta})$, is one example. The convergence power spectrum can be worked out in analogy with the matter power spectrum, except that the position and wave vectors are two-dimensional. The convergence power spectrum has the same form as the matter power spectrum (8.33), so we may immediately write

$$\langle \tilde{\kappa}(\mathbf{l}) \tilde{\kappa}^*(\mathbf{l}') \rangle = (2\pi)^2 \delta^{(2)}(\mathbf{l} - \mathbf{l}') P_\kappa(l) \,, \tag{8.37}$$

where $l \equiv |\mathbf{l}|$. Besides replacing δ by κ, the remaining differences between this expression and Eq. (8.33) arise from the two-dimensional context in which we are now working. Introducing plane polar coordinates (R, ϕ), we have for the convergence power spectrum,

$$P_\kappa(l) = \int_0^{\infty}\int_0^{2\pi} \xi_\kappa(R) e^{ilR\cos\phi} R \, \mathrm{d}\phi \, \mathrm{d}R$$
$$= \int_0^{\infty} R\xi_\kappa(R) \int_0^{2\pi} e^{ilR\cos\phi} \, \mathrm{d}\phi \, \mathrm{d}R \,. \tag{8.38}$$

Now we use the identity

$$J_0(lR) \equiv \frac{1}{2\pi} \int_0^{2\pi} e^{\pm ilR\cos(\phi)} \, \mathrm{d}\phi \tag{8.39}$$

to write the convergence power spectrum as

$$\boxed{P_\kappa(l) = \int_0^{\infty} 2\pi R \xi_\kappa(R) J_0(lR) \, \mathrm{d}R} \,. \tag{8.40}$$

Both Eqs. (8.36) and (8.40) may be inverted to express the correlation function as an integral over the power spectrum.

Since the correlation function and the power spectrum are related by Fourier transformation, they contain the same information. In analyzing observations, one

or the other may be more convenient. The correlation function $\langle\delta(\mathbf{x})\delta(\mathbf{x}')\rangle$ depends on $\mathbf{x}$ and $\mathbf{x}'$ through the single variable $r = |\mathbf{x} - \mathbf{x}'|$. Since $\langle\tilde{\delta}(\mathbf{k})\tilde{\delta}^*(\mathbf{k}')\rangle$ vanishes unless $\mathbf{k} = \mathbf{k}'$, the Fourier components $\{\tilde{\delta}(\mathbf{k})\}$ are statistically independent. If we further assume that $\delta(\mathbf{x})$ is normally distributed, we say that $\delta(\mathbf{x})$ is a **Gaussian random field**.

8.2.2 *Limber's Equation*

Since angular positions in the sky are easier to measure than displacements in three dimensions, one often encounters projections of the density contrast δ. Here we show how to express the correlation function or power spectrum of a projected quantity in terms of its corresponding unprojected version. Consider the random fields $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$, which are weighted integrals of δ along the line of sight. For specified weighting functions $Q_1(\boldsymbol{\theta})$ and $Q_2(\boldsymbol{\theta})$, we write

$$g_i(\boldsymbol{\theta}) = \int_0^{\chi_h} Q_i(\boldsymbol{\theta})\delta(f_k(\chi)\boldsymbol{\theta}, \chi)\,\mathrm{d}\chi\,, \tag{8.41}$$

where $i = 1, 2$. The covariance of $g_1(\boldsymbol{\vartheta})$ and $g_2(\boldsymbol{\vartheta} + \boldsymbol{\theta})$ is given by

$$\begin{aligned}\xi_{12}(\theta) &\equiv \langle g_1(\boldsymbol{\vartheta})g_2(\boldsymbol{\vartheta} + \boldsymbol{\theta})\rangle \\ &= \int_0^{\chi_h} \mathrm{d}\chi \int_0^{\chi_h} \mathrm{d}\chi'\, Q_1(\chi)Q_2(\chi') \\ &\quad \times\langle\delta(f_k(\chi)\boldsymbol{\vartheta}, \chi)\delta(f_k(\chi')(\boldsymbol{\vartheta} + \boldsymbol{\theta}), \chi')\rangle\,,\end{aligned} \tag{8.42}$$

where $\theta = |\boldsymbol{\theta}|$. Since the correlation function ξ_δ of the density contrast depends only on distance, so too does ξ_{12}. If $g_2 = g_1$, then ξ_{11} is said to be the autocorrelation of g_1. Otherwise, we call ξ_{12} the cross-correlation of g_1 and g_2.

In the Limber approximation, we assume that the correlation function ξ_{12} vanishes when $|\Delta\chi| > \chi_c$, where $\Delta\chi \equiv \chi' - \chi$, and χ_c is the comoving correlation length. Since $\chi_c \ll \chi_h$ we can set $f_k(\chi + \Delta\chi) \approx f_k(\chi)$ and $Q_i(\chi + \Delta\chi) \approx Q_i(\chi)$. Equation (8.42) then becomes

$$\boxed{\xi_{12}(\theta) \approx \int_0^{\chi_h} \mathrm{d}\chi \int_{-\infty}^{\infty} \mathrm{d}\Delta\chi\; Q_1(\chi)Q_2(\chi)\xi_\delta\left(\sqrt{[f_k(\chi)]^2\theta^2 + (\Delta\chi)^2}\right)}\,, \tag{8.43}$$

which is one form of **Limber's equation** (Limber 1953).

To relate the projected and unprojected power spectra, we first express the correlation function in Eq. (8.42) in terms of Fourier transforms as

$$\xi_\delta = \langle\delta(f_k(\chi)\boldsymbol{\vartheta}, \chi)\delta(f_k(\chi + \Delta\chi)(\boldsymbol{\vartheta} + \boldsymbol{\theta}), \chi + \Delta\chi)\rangle$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^3} \frac{\mathrm{d}^3k}{(2\pi)^3} \\
&\quad \int_{\mathbb{R}^3} \frac{\mathrm{d}^3k'}{(2\pi)^3}\, e^{-if_k(\chi)\mathbf{k}_\perp\cdot\boldsymbol{\vartheta}} e^{if_k(\chi+\Delta\chi)\mathbf{k}'_\perp\cdot(\boldsymbol{\vartheta}+\boldsymbol{\theta})} e^{-ik_\parallel\chi} e^{ik'_\parallel(\chi+\Delta\chi)} \langle\tilde{\delta}(\mathbf{k})\tilde{\delta}^*(\mathbf{k}')\rangle \\
&\approx \int_{\mathbb{R}^3} \frac{\mathrm{d}^3k}{(2\pi)^3}\, e^{if_k(\chi)\mathbf{k}_\perp\cdot\boldsymbol{\theta}} e^{ik_\parallel\Delta\chi}\, P_\delta(k)\,, \qquad (8.44)
\end{aligned}
$$

where we have used the definition of the power spectrum (8.33) and the Limber approximation. We have also written the wave vectors in terms of components parallel and perpendicular to the line of sight. Substituting Eq. (8.44) into Eq. (8.43) leads to

$$
\begin{aligned}
\xi_{12}(\theta) &\approx \int_0^{\chi_h} \mathrm{d}\chi \int_{\mathbb{R}^2} \frac{\mathrm{d}^2k_\perp}{(2\pi)^2} \int_{-\infty}^{\infty} \mathrm{d}k_\parallel\, Q_1(\chi)Q_2(\chi) e^{if_k(\chi)\mathbf{k}_\perp\cdot\boldsymbol{\theta}} \delta^{(1)}(k_\parallel) P_\delta(k) \\
&= \int_0^{\chi_h} \mathrm{d}\chi \int_{\mathbb{R}^2} \frac{\mathrm{d}^2k_\perp}{(2\pi)^2}\, Q_1(\chi)Q_2(\chi) e^{if_k(\chi)\mathbf{k}_\perp\cdot\boldsymbol{\theta}} P_\delta(k_\perp)\,, \qquad (8.45)
\end{aligned}
$$

where $k_\perp \equiv |\mathbf{k}_\perp|$, and the δ-function in the first line follows from the identity

$$
\int_{-\infty}^{\infty} e^{ik_\parallel\Delta\chi}\mathrm{d}\Delta\chi = 2\pi\delta^{(1)}(k_\parallel)\,. \qquad (8.46)
$$

We can reduce the number of integrals in Eq. (8.45) by using Eq. (8.39) to write

$$
\xi_{12}(\theta) \approx \int_0^{\chi_h} \mathrm{d}\chi \int_0^{\infty} \frac{k_\perp \mathrm{d}k_\perp}{2\pi}\, Q_1(\chi)Q_2(\chi) J_0(k_\perp\theta) P_\delta(k_\perp)\,. \qquad (8.47)
$$

To relate the projected power spectrum P_{12} to the correlation function ξ_{12}, we multiply both sides of Eq. (8.45) by $e^{-i\mathbf{l}\cdot\boldsymbol{\theta}}$ and integrate over $\boldsymbol{\theta}$. This gives

$$
\begin{aligned}
&\int_{\mathbb{R}^2} \mathrm{d}^2\theta\, \xi_{12}(\theta) e^{-i\mathbf{l}\cdot\boldsymbol{\theta}} \\
&\quad = \int_0^{\chi_h} \mathrm{d}\chi \int_{\mathbb{R}^2} \frac{\mathrm{d}^2k_\perp}{(2\pi)^2} \int_{\mathbb{R}^2} \mathrm{d}^2\theta\, Q_1(\chi)Q_2(\chi) e^{i[f_k(\chi)\mathbf{k}_\perp-\mathbf{l}]\cdot\boldsymbol{\theta}} P_\delta(k_\perp)\,.
\end{aligned}
$$

The left-hand side is just the power spectrum, and the θ-integral on the right-hand side yields the factor

$$
\delta^{(2)}(f_k(\chi)\mathbf{k}_\perp - \mathbf{l}) = f_k(\chi)^{-2}\delta^{(2)}(\mathbf{k}_\perp - f_k(\chi)^{-1}\mathbf{l})\,. \qquad (8.48)
$$

Thus, we arrive at

$$\boxed{P_{12}(l) \approx \int_0^{\chi_h} \mathrm{d}\chi \, \frac{Q_1(\chi) Q_2(\chi)}{f_k^2(\chi)} \, P_\delta\left(\frac{l}{f_k(\chi)}\right)}, \tag{8.49}$$

which is another version of Limber's equation.

8.3 Modeling and Measuring the Matter Power Spectrum

It has been argued (Harrison 1970; Peebles and Yu 1970; Zeldovich 1972) that the variance of fluctuations in the primordial gravitational potential were scale free, with all wave numbers contributing equally. One way to justify this assertion is to consider a homogeneous, isotropic universe with local perturbations described by a small gravitational potential Φ_1. The simplest metric for such a model is obtained by combining the weak-field metric (3.71) and the Robertson-Walker (RW) metric (3.123), namely,

$$\mathrm{d}s^2 = \left(1 + \frac{2\Phi_1}{c^2}\right) c^2 \mathrm{d}t^2 - a^2(t) \left(1 - \frac{2\Phi_1}{c^2}\right) \left[\frac{\mathrm{d}r^2}{1 - kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\phi^2)\right] . \tag{8.50}$$

Quantum fluctuations in the early universe, which translate into leading-order corrections to the RW metric, are then proportional to the gravitational potential.

The variance of the perturbation to the gravitational potential Φ_1 is related to the potential correlation function by $\sigma_\Phi^2 = \xi_\Phi(0)$. By inverting Eq. (8.34) and setting $\mathbf{r} = \mathbf{0}$, we find

$$\sigma_\Phi^2 = \int_{\mathbb{R}^3} \frac{\mathrm{d}^3 k}{(2\pi)^3} P_\Phi(k) = \frac{1}{2\pi^2} \int_0^\infty \mathrm{d}k \, k^2 P_\Phi(k) \, , \tag{8.51}$$

where $P_\Phi(k)$ is the power spectrum of Φ_1. Now consider the perturbed Poisson's equation (8.16) in Fourier space,

$$-k^2 \tilde{\Phi}_1 = 4\pi G \bar{\rho} \tilde{\delta} \, , \tag{8.52}$$

where $\tilde{\Phi}_1$ and $\tilde{\delta}$ are the Fourier transforms of Φ_1 and δ. The covariance of $\tilde{\Phi}_1(\mathbf{k})$ and $\tilde{\Phi}_1^*(\mathbf{k}')$ is then

$$\langle \tilde{\Phi}_1(\mathbf{k}) \tilde{\Phi}_1^*(\mathbf{k}') \rangle = \left(\frac{4\pi G \bar{\rho}}{k k'}\right)^2 \langle \tilde{\delta}(\mathbf{k}) \tilde{\delta}^*(\mathbf{k}') \rangle \, . \tag{8.53}$$

Thus, from the definition of the power spectrum (8.33),

$$P_\Phi(k) = (4\pi G \bar{\rho})^2 \frac{P_\delta(k)}{k^4} \, . \tag{8.54}$$

Assuming the *primordial* matter power spectrum has the form $P_\delta^0 \propto k$, known as the Harrison-Zel'dovich power spectrum, the variance of Φ_1 has the scale-free form

$$\sigma_\Phi^2 \propto \int \frac{\mathrm{d}k}{k}\,, \tag{8.55}$$

which is invariant under the transformation $k \to \alpha k$, for any constant α. The theory of inflation also predicts a power law $P_\delta^0 \propto k^n$ but with n slightly below unity.

The primordial matter power spectrum was modified by various physical processes. In the linear regime, the matter power spectrum during a particular epoch can be written as $P_\delta(k) = T(k)P_\delta^0(k)$. The **transfer function**, $T(k)$, is governed by the evolution of the density contrast δ. The dependence of T on the wavenumber k arises from that of δ on the scale factor a, but the details do not concern us here. For a universe dominated by matter or radiation, we found in Sect. 8.1.1 that $\delta \propto a$ (matter-dominated case) and $\delta \propto a^2$ (radiation-dominated case). Since $P_\delta \propto |\tilde{\delta}|^2$ and $a \propto k^{-1}$ (roughly speaking, at least), the transfer function for a fixed P_δ^0 has the form $T(k) \propto k^{-2}$ for a matter-dominated universe and $T(k) \propto k^{-4}$ for a radiation-dominated universe (e.g., Bartelmann and Schneider 2001).

If the correlation function were known for every pair of angular positions, one could infer the power spectrum for any given redshift. However, that would require knowing the density contrast at all points. Instead of the fine-grained density contrast $\delta(\mathbf{x})$, which is difficult to measure precisely, one considers the coarse-grained quantity $\delta_b(\mathbf{x})$, smoothed over a sphere of radius b according to

$$\delta_b(\mathbf{x}) = \left(\frac{4}{3}\pi b^3\right)^{-1} \int_{\mathbb{R}^3} \delta(\mathbf{y})\Theta(b - |\mathbf{x} - \mathbf{y}|)\,\mathrm{d}^3 y\,, \tag{8.56}$$

where Θ is the Heaviside step function. In terms of the *filter function*

$$W_b(r) \equiv \left(\frac{4}{3}\pi b^3\right)^{-1} \Theta(b - r)\,, \tag{8.57}$$

we can write

$$\delta_b(\mathbf{x}) = \int_{\mathbb{R}^3} \delta(\mathbf{y}) W_b(|\mathbf{x} - \mathbf{y}|)\,\mathrm{d}^3 y = [\delta * W_b]\,(\mathbf{x})\,. \tag{8.58}$$

In words, the smoothed (coarse-grained) density contrast is the convolution of the actual (fine-grained) density contrast with a filter function. This particular choice of W_b, known as the **top-hat filter**,[4] is the simplest of a variety of filter functions that work equally well. One example, the Gaussian, has the convenient property that its Fourier transform is also Gaussian.

[4]This name comes from the two-dimensional analog of W_b, whose graph resembles a top hat.

To obtain the coarse-grained power spectrum $P_\delta^b(k)$, we start by taking the Fourier transform of Eq. (8.58), namely

$$\tilde{\delta}_b(\mathbf{k}) = \tilde{\delta}(\mathbf{k})\tilde{W}_b(k)\,, \tag{8.59}$$

where for the top-hat filter

$$\tilde{W}_b(k) = 3\,\frac{\sin(kb) - kb\cos(kb)}{(kb)^3} \tag{8.60}$$

(e.g., Bracewell 1965). According to Eq. (8.33),

$$\begin{aligned}\langle\tilde{\delta}_b(\mathbf{k})\tilde{\delta}_b^*(\mathbf{k}')\rangle &= \tilde{W}_b(k)\tilde{W}_b^*(k')\langle\tilde{\delta}(\mathbf{k})\tilde{\delta}^*(\mathbf{k}')\rangle \\ &= (2\pi)^3\delta^{(3)}(\mathbf{k}-\mathbf{k}')|\tilde{W}_b(k)|^2 P_\delta(k)\,.\end{aligned} \tag{8.61}$$

Thus, the coarse-grained power spectrum is given by

$$P_\delta^b(k) = |\tilde{W}_b(k)|^2 P_\delta(k)\,. \tag{8.62}$$

In terms of the power spectrum, the variance of δ_b is

$$\sigma_b^2 = \int_{\mathbb{R}^3}\frac{\mathrm{d}^3k}{(2\pi)^3}P_\delta^b(k) = \int_{\mathbb{R}^3}\frac{\mathrm{d}^3k}{(2\pi)^3}|\tilde{W}_b(k)|^2 P_\delta(k)\,. \tag{8.63}$$

Thus, the variance of the density contrast is said to *normalize*[5] the power spectrum P_δ^b. A convenient length scale b for coarse-graining is the radius $b \sim 8h^{-1}$ Mpc of a typical galaxy cluster, where h is the Hubble constant in units of 100 km s^{-1} Mpc^{-1}. We can estimate the corresponding variance σ_8^2 from galaxy counts and obtain the matter power spectrum P_δ using Eqs. (8.60) and (8.63).

8.4 Cosmic Shear

In addition to the possibility that a distant source is lensed by one or more mass concentrations along the line of sight, there is a weaker yet measurable effect of cosmological significance. Because of the mass sheet degeneracy and the unknown intrinsic size of the source, the shear due to matter at all distances along the line of sight is the quantity of interest. This **cosmic shear** allows us to constrain cosmological parameters by measuring the shear of sources at a large number of

[5]For a power spectrum of the form $P(k) = Ak^n$, any method for determining the constant A is referred to as "normalization," even if it does not involve integrating $P(k)$.

different redshifts. The key mathematical difference between cosmic shear and lensing by discrete objects is that the thin-lens approximation no longer holds. However, the analysis is somewhat simplified since the total deflection is quite small.

Consider a source at comoving distance χ from the observer. The reduced deflection originating in a slab of thickness $\mathrm{d}\chi'$ at distance $\chi' < \chi$ from the observer is given by

$$\mathrm{d}\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{d_{ls}}{d_s}\mathrm{d}\hat{\boldsymbol{\alpha}}(d_l\boldsymbol{\theta}) = \frac{2}{c^2}\frac{d_{ls}}{d_s}\boldsymbol{\nabla}_\perp\Phi_1(d_l\boldsymbol{\theta}, \chi')\,\mathrm{d}\chi'\,, \tag{8.64}$$

where $\boldsymbol{\nabla}_\perp$ is the portion of the comoving gradient perpendicular to the χ'-axis. Using Eq. (3.122), we express the comoving angular-diameter distances[6] as $d_l = f_k(\chi')$, $d_s = f_k(\chi)$, and $d_{ls} = f_k(\chi - \chi')$. The gravitational potential Φ_1 satisfies the comoving form (8.17) of Poisson's equation.

Integrating Eq. (8.64) over χ', we obtain

$$\boldsymbol{\alpha}(\boldsymbol{\theta}|\chi) = \frac{2}{c^2}\int_0^\chi \frac{f_k(\chi - \chi')}{f_k(\chi)}\boldsymbol{\nabla}_\perp\Phi_1(f_k(\chi')\boldsymbol{\theta}, \chi')\,\mathrm{d}\chi' \tag{8.65}$$

for the total reduced deflection. Here we use the Born approximation (cf. Sect. 2.1) and integrate along the straight line between the observer and source, rather than along the actual path of the light ray. In the special case of a single lens at distance d_l, Eq. (8.65) becomes

$$\boldsymbol{\alpha}(\boldsymbol{\theta}|\chi) = \frac{2}{c^2}\frac{d_{ls}}{d_s}\int_0^\chi \boldsymbol{\nabla}_\perp\Phi_1(d_l\boldsymbol{\theta}, \chi')\,\mathrm{d}\chi'\,. \tag{8.66}$$

In the limit $\chi \to \infty$, which is appropriate for a spatially concentrated lens far from both observer and source, we have

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) \equiv \lim_{\chi\to\infty}\boldsymbol{\alpha}(\boldsymbol{\theta}|\chi) = \frac{d_{ls}}{d_s}\hat{\boldsymbol{\alpha}}(d_l\boldsymbol{\theta})\,, \tag{8.67}$$

where $\hat{\boldsymbol{\alpha}}$ is given by Eq. (2.6).

In terms of the angular gradient $\boldsymbol{\nabla}_{\boldsymbol{\theta}} = f_k(\chi')\boldsymbol{\nabla}_\perp$, the reduced deflection takes the form

$$\boldsymbol{\alpha}(\boldsymbol{\theta}|\chi) = \frac{2}{c^2}\int_0^\chi \frac{f_k(\chi - \chi')}{f_k(\chi)f_k(\chi')}\boldsymbol{\nabla}_{\boldsymbol{\theta}}\Phi_1(f_k(\chi')\boldsymbol{\theta}, \chi')\,\mathrm{d}\chi'\,. \tag{8.68}$$

[6]The *comoving* angular-diameter distance is the *physical* angular-diameter distance modulo the scale factor.

Since $\nabla_{\boldsymbol{\theta}}$ may be taken outside the integral, we can write this as

$$\boldsymbol{\alpha}(\boldsymbol{\theta}|\chi) = \nabla_{\boldsymbol{\theta}}\psi_{\rm eff}(\boldsymbol{\theta}|\chi)\,, \tag{8.69}$$

with the effective lens potential

$$\psi_{\rm eff}(\boldsymbol{\theta}|\chi) = \frac{2}{c^2}\int_0^\chi \frac{f_k(\chi-\chi')}{f_k(\chi)f_k(\chi')}\Phi_1(f_k(\chi')\boldsymbol{\theta},\chi')\,{\rm d}\chi'\,. \tag{8.70}$$

8.4.1 Effective Convergence

Associated with $\psi_{\rm eff}$ is an effective convergence $\kappa_{\rm eff}$ defined by $\nabla^2_{\boldsymbol{\theta}}\psi_{\rm eff} = 2\kappa_{\rm eff}$. We can compute $\kappa_{\rm eff}$ in two ways. Using Eq. (8.68), it is given by

$$\begin{aligned}\kappa_{\rm eff}(\boldsymbol{\theta}|\chi) &= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\cdot\boldsymbol{\alpha}(\boldsymbol{\theta}|\chi)\\ &= \frac{1}{c^2}\int_0^\chi \frac{f_k(\chi-\chi')}{f_k(\chi)f_k(\chi')}\nabla^2_{\boldsymbol{\theta}}\Phi_1(f_k(\chi')\boldsymbol{\theta},\chi')\,{\rm d}\chi'\\ &= \frac{1}{c^2}\int_0^\chi \frac{f_k(\chi-\chi')f_k(\chi')}{f_k(\chi)}\nabla^2_\perp\Phi_1(f_k(\chi')\boldsymbol{\theta},\chi'){\rm d}\chi'\,.\end{aligned} \tag{8.71}$$

Alternatively, we can express $\kappa_{\rm eff}$ in terms of the density contrast δ. The convergence due to a thin slab perpendicular to the line of sight with proper width ${\rm d}D_{\rm prop}$ is given by

$${\rm d}\kappa = \frac{{\rm d}\Sigma}{\Sigma_{\rm crit}} = \frac{4\pi G}{c^2}\frac{D_l D_{ls}}{D_s}(\rho-\bar\rho)\,{\rm d}D_{\rm prop}\,, \tag{8.72}$$

where $\Sigma_{\rm crit}$ is defined in Eq. (2.68). Since a light ray propagating through the unperturbed homogeneous, isotropic universe with density $\bar\rho$ is unlensed by definition, the convergence arises from the difference $\rho-\bar\rho = \rho_0 a^{-3}\delta$. For a light ray traveling along the straight line from source to observer, the RW metric (3.123b) gives $c{\rm d}t = a(t)\,{\rm d}\chi'$, where $\chi' < \chi$. This allows us to write the scale factor as $\hat a(\chi') \equiv [a\circ t](\chi')$. We can also replace the angular-diameter distances by their comoving counterparts and likewise for the differential proper distance ${\rm d}D_{\rm prop} = \hat a(\chi')\,{\rm d}\chi'$. Putting the pieces together, and integrating along the line of sight, we obtain

$$\kappa_{\rm eff}(\boldsymbol{\theta}|\chi) = \frac{4\pi G\rho_0}{c^2}\int_0^\chi \frac{f_k(\chi-\chi')f_k(\chi')}{f_k(\chi)\hat a(\chi')}\delta(f_k(\chi')\boldsymbol{\theta},\chi')\,{\rm d}\chi'\,. \tag{8.73}$$

White and Hu (2000) showed through numerical simulations that Eqs. (8.71) and (8.73) make similar predictions for $\kappa_{\rm eff}$, and lead to nearly the same statistical

conclusions, e.g., for the convergence power spectrum. If the two-dimensional Laplacian $\nabla_\perp^2 \Phi_1$ could be replaced by the three-dimensional Laplacian $\hat{\nabla}^2 \Phi_1$ of Eq. (8.17), then Eqs. (8.71) and (8.73) would be identical. Some authors claim that $\partial^2 \Phi_1 / \partial \chi'^2 = 0$, at least when averaged along the line of sight (e.g., Bartelmann and Schneider 2001; Kilbinger 2015), but a rigorous proof has yet to emerge.

In order to compute the convergence power spectrum, which can be inferred from gravitational lensing, we must average $\kappa_{\rm eff}$ over all source distances $\chi < \chi_h$, where the comoving horizon distance χ_h is defined by $\hat{a}(\chi_h) = 0$, corresponding to infinite redshift. Let the probability that there is a source between χ and $\chi + \mathrm{d}\chi$ be $g(\chi)\,\mathrm{d}\chi$. Then

$$\begin{aligned}
\bar{\kappa}_{\rm eff}(\boldsymbol{\theta}) &\equiv \int_0^{\chi_h} \kappa_{\rm eff}(\boldsymbol{\theta}|\chi) g(\chi)\,\mathrm{d}\chi \\
&= \frac{4\pi G\rho_0}{c^2} \int_0^{\chi_h} \int_0^{\chi} \frac{f_k(\chi-\chi') f_k(\chi')}{f_k(\chi)\hat{a}(\chi')} \delta(f_k(\chi')\boldsymbol{\theta}, \chi') g(\chi)\,\mathrm{d}\chi'\,\mathrm{d}\chi \\
&= \frac{4\pi G\rho_0}{c^2} \int_0^{\chi_h} \int_0^{\chi_h} \frac{f_k(\chi-\chi') f_k(\chi')}{f_k(\chi)\hat{a}(\chi')} \\
&\quad \times \delta(f_k(\chi')\boldsymbol{\theta}, \chi') g(\chi) \Theta(\chi-\chi')\,\mathrm{d}\chi'\,\mathrm{d}\chi\,, \qquad (8.74)
\end{aligned}$$

where the step function allows us to extend the χ' integral out to the horizon distance. Since both integrals have the same limits, we can exchange the dummy variables χ and χ' to write

$$\begin{aligned}
\bar{\kappa}_{\rm eff}(\boldsymbol{\theta}) &= \frac{4\pi G\rho_0}{c^2} \int_0^{\chi_h} \int_0^{\chi_h} \frac{f_k(\chi'-\chi) f_k(\chi)}{f_k(\chi')\hat{a}(\chi)} \\
&\quad \times \delta(f_k(\chi)\boldsymbol{\theta}, \chi) g(\chi') \Theta(\chi'-\chi)\,\mathrm{d}\chi'\,\mathrm{d}\chi \\
&= \frac{4\pi G\rho_0}{c^2} \int_0^{\chi_h} \frac{f_k(\chi)}{\hat{a}(\chi)} W(\chi) \delta(f_k(\chi)\boldsymbol{\theta}, \chi)\,\mathrm{d}\chi\,, \qquad (8.75)
\end{aligned}$$

where

$$W(\chi) \equiv \int_\chi^{\chi_h} \frac{f_k(\chi'-\chi)}{f_k(\chi')} g(\chi')\,\mathrm{d}\chi'\,. \qquad (8.76)$$

To give the weighting function $W(\chi)$ a physical interpretation, it is helpful to think of χ as the comoving distance from observer to lens and χ' as that from observer to source. Thus, W is a weighted average of the ratio of lens-source distance to observer-source distance. Bearing this in mind, we can integrate Eq. (8.72) out to some redshift z. The effective convergence averaged over redshift is then

$$\bar{\kappa}_{\rm eff} = \frac{4\pi G\rho_0}{c^2} \int_0^{\infty} D(0,z) V(z) (\rho - \bar{\rho})\,\mathrm{d}z\,, \qquad (8.77)$$

where $D(z_1, z_2)$ is the angular diameter distance of z_2 from z_1. If $p(z)\mathrm{d}z$ is the probability that there is a source between z and $z + \mathrm{d}z$, the weighting function is given by

$$V(z) \equiv \int_z^\infty \frac{D(z, z')}{D(0, z')} p(z')\, \mathrm{d}z' . \tag{8.78}$$

Note that redshift is directly observable, while comoving distance is not.

8.4.2 *Correlation Functions and Power Spectra of Shear and Convergence*

Since the effective convergence is a two-dimensional projection of the three-dimensional density contrast, the power spectra of these fields are related by Limber's equation (8.49). For the effective convergence (8.75), the weighting functions in Eq. (8.41) satisfy $Q_1(\chi) = Q_2(\chi) = Q(\chi)$, with

$$Q(\chi) \equiv \frac{4\pi G\rho_0}{c^2} \frac{f_k(\chi) W(\chi)}{\hat{a}(\chi)} . \tag{8.79}$$

We then conclude from Eq. (8.49) that

$$P_\kappa(l) \approx \frac{9\Omega_0^2 H_0^4}{4c^4} \int_0^{\chi_h} \mathrm{d}\chi \left[\frac{W(\chi)}{\hat{a}(\chi)}\right]^2 P_\delta\left(\frac{l}{f_k(\chi)}, \chi\right) , \tag{8.80}$$

where we have written the density ρ_0 in terms of the parameters Ω_0 and H_0. The second argument of P_δ reminds us that the matter power spectrum evolves with cosmic epoch, for which the comoving distance is a proxy.

Equation (8.80) shows that the effective-convergence power spectrum depends directly on the statistics of density fluctuations in the universe. While P_κ is not directly observable, it is closely related to the statistics of shear, which can be measured. Problem 7.2 establishes that the Fourier transforms of the convergence and shear fields are related by $\tilde{\gamma}(\mathbf{l}) = f(\mathbf{l})\tilde{\kappa}(\mathbf{l})$, and it turns out that $|f(\mathbf{l})| = 1$. Therefore the shear and convergence power spectra are identical: $P_\gamma(l) = P_\kappa(l)$. In practice, measurements usually focus on correlation functions rather than power spectra. Since there are two components of shear, there are three possible correlation functions that can be written schematically as $\langle\gamma_+\gamma_+\rangle$, $\langle\gamma_\times\gamma_\times\rangle$, and $\langle\gamma_+\gamma_\times\rangle$. The cross term vanishes in the ensemble average, and the remaining two terms are combined into

$$\xi_\pm(\vartheta) = \langle\gamma_+(\boldsymbol{\theta})\gamma_+(\boldsymbol{\theta}')\rangle \pm \langle\gamma_\times(\boldsymbol{\theta})\gamma_\times(\boldsymbol{\theta}')\rangle , \tag{8.81}$$

which depend only on the distance $\vartheta = |\boldsymbol{\theta} - \boldsymbol{\theta}'|$ between the points. These correlation functions can be written in terms of the complex shear $\gamma = \gamma_+ + i\gamma_\times$ as

$$\xi_+(\vartheta) = \langle \gamma(\boldsymbol{\theta})\gamma^*(\boldsymbol{\theta}')\rangle \quad \text{and} \quad \xi_-(\vartheta) = \mathrm{Re}\langle \gamma(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}')\rangle\,. \tag{8.82}$$

Converting to Fourier space and again relating $\tilde{\gamma}$ and $\tilde{\kappa}$ allow us to write the correlation functions as integrals over the convergence power spectrum,

$$\xi_+(\vartheta) = \frac{1}{2\pi}\int P_\kappa(l)\, J_0(l\vartheta)\, l\, \mathrm{d}l \tag{8.83a}$$

$$\xi_-(\vartheta) = \frac{1}{2\pi}\int P_\kappa(l)\, J_4(l\vartheta)\, l\, \mathrm{d}l\,. \tag{8.83b}$$

The angular integrals can be evaluated because P_κ depends only on the length of $\mathbf{l}$ (not on its direction), and they yield Bessel functions J_0 and J_4.

8.5 Applications

8.5.1 Cosmic Shear

Since its first detections (Bacon et al. 2000; Van Waerbeke et al. 2000; Wittman et al. 2000), cosmic shear has drawn a lot of attention as a probe of large-scale structure in the universe and, by extension, cosmological parameters. The measurements are challenging, but concerted effort with both observations and analysis has allowed current surveys, such as the Dark Energy Survey (DES, Abbott et al. 2018), Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP, Aihara et al. 2018), and Kilo-Degree Survey (KiDS, de Jong et al. 2017), to measure cosmic shear over many hundreds of square degrees. We highlight key aspects of the measurements here and refer readers to the literature for technical details.

Once galaxies have been observed, the first step is to measure their shapes. As discussed in Sect. 7.2.2, galaxy shapes can in principle be measured directly from quadrupole moments of the light distribution. The task is complicated, however, by distortions from Earth's atmosphere and telescope optics, not to mention noise in the data (from photon counts and detector read noise). The community has developed and tested sophisticated methods to deal with distortions and noise precisely enough to uncover the subtle effects of cosmic shear (e.g., Mandelbaum et al. 2015; Huff and Mandelbaum 2017; Sheldon and Huff 2017; Fenech Conti et al. 2017). Briefly, the methods rely on models for the distortions that can be calibrated from simulations and/or the data directly. Current surveys provide shape measurements for tens of millions of galaxies (Zuntz et al. 2018; Hildebrandt et al. 2017; Mandelbaum et al. 2018).

The second step is to determine how the source galaxies are distributed in redshift. This is necessary for computing the overall effective convergence (note the factor $g(\chi)$ that appears in Eq. (8.75)). It also offers the opportunity, given enough data, to separate the sources in redshift bins and perform "tomography" to study how the convergence varies with source redshift,[7] which offers information about the growth of structure. Measuring spectroscopic redshifts is not feasible for the vast number of faint galaxies needed to measure cosmic shear, but "photometric redshifts" can be estimated by analyzing the brightnesses of galaxies in several broadband filters. Methods for computing photometric redshifts can be calibrated on subsamples of galaxies that have spectroscopic data, and they have achieved sufficient accuracy and precision to be used for cosmic shear analyses (see Hildebrandt et al. (2010) and references therein). Redshift distributions for current surveys are reported by Hoyle et al. (2018), Tanaka et al. (2018), and Hildebrandt et al. (2017). Strategies are being developed to meet the challenge of attaining the accuracy and precision that will be needed for future surveys (see Newman et al. 2015; Mandelbaum 2018).

The measured shapes can be used to compute the shear correlation functions $\xi_{\pm}(\vartheta)$ defined in Eq. (8.83). The ensemble average is approximated by a weighted sum of all galaxies with a separation of ϑ (where the weighting can account for variations in data quality). The shear correlation functions have been measured using 26 million galaxies in DES (Troxel et al. 2018), more than 12 million galaxies in HSC-SSP (Mandelbaum et al. 2018), and nearly 15 million galaxies in KiDS (Hildebrandt et al. 2017). The measurements can be compared against theoretical predictions by combining Eqs. (8.80) and (8.83) to write $\xi_{\pm}$ in terms of integrals over the matter power spectrum P_{δ}. Because linear perturbation theory breaks down on some of the relevant scales, predictions for P_{δ} must be derived from numerical simulations, but that process is well understood. Cosmic shear predictions are most sensitive to two cosmological parameters: the overall amount of matter in the universe, characterized by the density parameter Ω_M, and the amplitude of density fluctuations, characterized by σ_8 (see Sect. 8.3). They depend mainly on the combination $S_8 \equiv \sigma_8(\Omega_M/0.3)^{0.5}$. Among current surveys, DES yields $S_8 = 0.782 \pm 0.027$ (Troxel et al. 2018), while KiDS yields $S_8 = 0.745 \pm 0.039$ (Hildebrandt et al. 2017), and these results are generally consistent with other probes of large-scale structure. Combining cosmic shear with other probes improves constraints on cosmological parameters (see, e.g., Abbott et al. 2018), and as surveys expand still further cosmic shear is expected to play a significant role in probing dark energy (see Albrecht et al. 2006).

[7] Comparing shapes of sources at redshift z to those at $z + \Delta z$ allows one to infer the matter content between z and $z + \Delta z$. It is similar in spirit to medical tomography, in that information from n dimensions is used to reconstruct an $(n + 1)$-dimensional picture.

8.5.2 Cosmic Magnification

This chapter has focused on changes in shape caused by shear, but there are also effects from magnification, namely, changes in the size and brightness of lensed images and in the effective area of a survey. To quantify the two effects, let $n_{\rm src}(S_{\rm src}) = {\rm d}^2 N_{\rm src}(S_{\rm src})/{\rm d}S_{\rm src}\,{\rm d}\Omega_{\rm src}$ be the number of sources with fluxes between $S_{\rm src}$ and $S_{\rm src} + {\rm d}S_{\rm src}$ in a small angular area ${\rm d}\Omega_{\rm src}$. With lensing magnification, the corresponding flux and area in the image plane are $S_{\rm img} = \mu\, S_{\rm src}$ and ${\rm d}\Omega_{\rm img} = \mu\,{\rm d}\Omega_{\rm src}$. Therefore the number of images with fluxes between $S_{\rm img}$ and $S_{\rm img} + {\rm d}S_{\rm img}$ in a small angular area ${\rm d}\Omega_{\rm img}$ is

$$n_{\rm img}(S_{\rm img}) = \frac{1}{\mu^2}\, n_{\rm src}\left(\frac{S_{\rm img}}{\mu}\right) . \tag{8.84}$$

Suppose the number of sources can be approximated as a power law $n_{\rm src}(S_{\rm src}) = A S_{\rm src}^{-\eta}$ where A is a constant and we expect $\eta > 0$ since bright sources are less abundant than faint ones. Then the number of images is

$$n_{\rm img}(S_{\rm img}) = \mu^{\eta-2}\, A\, S_{\rm img}^{-\eta} . \tag{8.85}$$

Lensing modifies the number of galaxies by the factor $\mu^{\eta-2}$. The factor of μ^{η} arises because magnification makes images brighter and therefore easier to detect. The factor of μ^{-2} arises because magnification reduces the area in the source plane that corresponds to a given portion of the image plane. The combined effect can either increase or decrease the number of lensed images depending on whether the increase in brightness or loss of area is the stronger effect.

The predicted dependence on the power-law slope η of the source luminosity function can be tested with observations that probe a range of fluxes. This is a valuable way to distinguish lensing from possible contaminants that would not depend on the source luminosity function in the same way.

As with other effects in lensing, cosmic magnification varies with the projected separation between lenses and sources. Therefore one way to detect it is to measure cross-correlations between background source populations and foreground lens populations (Scranton et al. 2005; Hildebrandt et al. 2009; Garcia-Fernandez et al. 2018). Another way is through changes in the distribution of galaxy sizes and brightnesses (Schmidt et al. 2012; Huff and Graves 2014). To date, cosmic magnification has been measured, but the statistical significance is not as high as it is for cosmic shear. However, cosmic magnification is still a useful complement to cosmic shear as the systematic uncertainties impacting each method are often distinct, allowing for cross validation of results.

Problems

8.1 The assumption of a pressureless fluid ($p = 0$) holds for dark matter, but not for ordinary (baryonic) matter, let alone radiation or dark energy. As mentioned in the footnote to Friedmann's second equation (8.9), in the case of non-zero pressure, the equation becomes

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\bar{\rho} + \frac{3p}{c^2}\right) .$$

(a) Show that this result follows from differentiating Friedmann's first equation (3.125) with respect to time, while imposing energy conservation in the form $\mathrm{d}\left(\bar{\rho}c^2 V\right) = -p\,\mathrm{d}V$, with $V = \frac{4}{3}\pi a^3$.

(b) Now assume an equation of state $p = w\bar{\rho}c^2$, where w is a constant. Consider the cases of radiation domination ($w = 1/3$) and dark energy domination ($w = -1$). What qualitative differences in the solution to Friedmann's second equation are expected in these two cases and in the pressureless case?

8.2 We know that the correlation function decreases monotonically, but this can happen in many ways. Here, we consider two simplified but instructive examples.

(a) For the correlation function $\xi_\delta(r) = Ae^{-ar}$, where A and a are positive constants, evaluate the correlation length r_c and the power spectrum $P_\delta(k)$ defined in Eqs. (8.31) and (8.36), respectively.

(b) Now consider another correlation function which is also finite for $0 \le r < \infty$, but which falls off as r^{-s} for $r \to \infty$. For what value(s) of s are the correlation length and the power spectrum well-defined (finite) quantities?

8.3 Cosmic shear measures an integral over the three-dimensional density field of the universe. Here we examine the redshift range that has significant weight in the integral. The effective convergence averaged over the population of sources can be written in the form

$$\bar{\kappa}_{\mathrm{eff}} = \int_0^\infty \int_{z_l}^\infty F(z_l, z_s)\,[\rho(z_l) - \bar{\rho}(z_l)]\,\mathrm{d}z_s\,\mathrm{d}z_l\,, \tag{8.86}$$

where z_l and z_s are lens and source redshifts, respectively. Recall that the weighting function $F(z_l, z_s)$ depends on the source redshift distribution $p(z)$. The exact form of $F(z_l, z_s)$ can be inferred from Sect. 8.4.1.

Using a model derived from the COSMOS field, Ilbert et al. (2009) found

$$p(z) = A\frac{z^a + z^{ab}}{z^b + c}\,, \tag{8.87}$$

where A must be chosen to satisfy the normalization condition $\int_0^\infty p(z)\,\mathrm{d}z = 1$. They examined a range of magnitudes, but we consider only the bright and faint ends of their sample: the bright sample has $(a, b, c) = (0.497, 12.643, 0.381)$, while the faint sample has $(a, b, c) = (0.126, 4.146, 5.925)$.

For both samples, plot $F(z_l, z_s)$ as a function of both the lens and source redshifts, and then plot $\int_{z_l}^\infty F(z_l, z_s)\,\mathrm{d}z_s$. How much does the redshift range depend on the depth to which source galaxies can be measured?

References

Abbott, T. M. C., Abdalla, F. B., Allam, S., Amara, A., Annis, J., Asorey, J., et al. (2018). The Dark Energy Survey data release 1. arXiv e-prints, arXiv:1801.03181.

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., Aleksić, J., Allam, S., et al. (2018). Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D, 98*, 043526.

Aihara, H., Arimoto, N., Armstrong, R., Arnouts, S., Bahcall, N. A., Bickerton, S., et al. (2018). The Hyper Suprime-Cam SSP Survey: Overview and survey design. *Publications of the Astronomical Society of Japan , 70*, S4.

Albrecht, A., Bernstein, G., Cahn, R., Freedman, W. L., Hewitt, J., Hu, W., et al. (2006). Report of the Dark Energy Task Force. arXiv Astrophysics e-prints, astro-ph/0609591.

Bacon, D. J., Refregier, A. R., & Ellis, R. S. (2000). Detection of weak gravitational lensing by large-scale structure. *Monthly Notices of the Royal Astronomical Society, 318*, 625.

Bartelmann, M., & Schneider, P. (2001). Weak gravitational lensing. *Physics Reports, 340*, 291.

Bracewell, R. (1965). *The Fourier transform and its applications*. New York: McGraw-Hill.

de Jong, J. T. A., Verdois Kleijn, G. A., Erben, T., Hildebrandt, H., Kuijken, K., Sikkema, G., et al. (2017). The third data release of the Kilo-Degree Survey and associated data products. *Astronomy and Astrophysics, 604*, A134.

Fenech Conti, I., Herbonnet, R., Hoekstra, H., Merten, J., Miller, L., & Viola, M. (2017). Calibration of weak-lensing shear in the Kilo-Degree Survey. *Monthly Notices of the Royal Astronomical Society, 467*, 1627.

Garcia-Fernandez, M., Sanchez, E., Sevilla-Noarbe, I., Suchyta, E., Huff, E. M., Gaztanaga, E., et al. (2018). Weak lensing magnification in the Dark Energy Survey Science Verification data. *Monthly Notices of the Royal Astronomical Society, 476*, 1071.

Harrison, E. R. (1970). Fluctuations at the threshold of classical cosmology. *Physical Review D, 1*, 2726.

Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L. A., Wolf, C., Abdalla, F. B., et al. (2010). PHAT: PHoto-z accuracy testing. *Astronomy and Astrophysics, 523*, A31.

Hildebrandt, H., van Waerbeke, L., & Erben, T. (2009). CARS: The CFHTLS-Archive-Research Survey. III. First detection of cosmic magnification in samples of normal high-z galaxies. *Astronomy and Astrophysics, 507*, 683.

Hildebrandt, H., Viola, M., Heymans, C., Joudaki, S., Kuijken, K., Blake, C., et al. (2017). KiDS-450: Cosmological parameter constraints from tomographic weak gravitational lensing. *Monthly Notices of the Royal Astronomical Society, 465*, 1454.

Hoyle, B., Gruen, D., Bernstein, G. M., Rau, M. M., De Vicente, J., Hartley, W. G., et al. (2018). Dark Energy Survey year 1 results: Redshift distributions of the weak-lensing source galaxies. *Monthly Notices of the Royal Astronomical Society, 478*, 592.

Huff, E., & Mandelbaum, R. (2017). Metacalibration: Direct self-calibration of biases in shear measurement. arXiv e-prints, arXiv:1702.02600.

Huff, E. M., & Graves, G. J. (2014). Magnificent magnification: Exploiting the other half of the lensing signal. *The Astrophysical Journal Letters, 780*, L16.

Ilbert, O., Capak, P., Salvato, M., Aussel, H., McCracken, H. J., Sanders, D. B., et al. (2009). Cosmos photometric redshifts with 30-bands for 2-deg^2. *The Astrophysical Journal, 690*, 1236.

Kilbinger, M. (2015). Cosmology with cosmic shear observations: A review. *Reports on Progress in Physics, 78*, 086901.

Limber, D. N. (1953). The analysis of counts of the extragalactic nebulae in terms of a fluctuating density field. *The Astrophysical Journal, 117*, 134.

Mandelbaum, R. (2018). Weak lensing for precision cosmology. *Annual Review of Astronomy and Astrophysics, 56*, 393–433.

Mandelbaum, R., Miyatake, H., Hamana, T., Oguri, M., Simet, M., Armstrong, R., et al. (2018). The first-year shear catalog of the Subaru Hyper Suprime-Cam Subaru Strategic Program Survey. *Publications of the Astronomical Society of Japan, 70*, S25.

Mandelbaum, R., Rowe, B., Armstrong, R., Bard, D., Bertin, E., Bosch, J., et al. (2015). GREAT3 results - I. Systematic errors in shear estimation and the impact of real galaxy morphology. *Monthly Notices of the Royal Astronomical Society, 450*, 2963.

Newman, J. A., Abate, A., Abdalla, F. B., Allam, S., Allen, S. W., Ansari, R., et al. (2015). Spectroscopic needs for imaging dark energy experiments. *Astroparticle Physics, 63*, 81.

Peacock, J. A. (1999). *Cosmological physics*. Cambridge: Cambridge University Press.

Peebles, P. J. E., & Yu, J. T. (1970). Primeval adiabatic perturbation in an expanding universe. *The Astrophysical Journal, 162*, 815.

Schmidt, F., Leauthaud, A., Massey, R., Rhodes, J., George, M. R., Koekemoer, A. M., et al. (2012). A detection of weak-lensing magnification using galaxy sizes and magnitudes. *The Astrophysical Journal Letters, 744*, L22.

Scranton, R., Ménard, B., Richards, G. T., Nichol, R. C., Myers, A. D., Jain, B., et al. (2005). Detection of cosmic magnification with the Sloan Digital Sky Survey. *The Astrophysical Journal, 633*, 589.

Sheldon, E. S., & Huff, E. M. (2017). Practical weak-lensing shear measurement with metacalibration. *The Astrophysical Journal, 841*, 24.

Tanaka, M., Coupon, J., Hsieh, B.-C., Mineo, S., Nishizawa, A. J., Speagle, J., et al. (2018). Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1. *Publications of the Astronomical Society of Japan, 70*, S9.

Thompson, M. J. (2006). *An introduction to astrophysical fluid dynamics*. London: Imperial College Press.

Troxel, M. A., MacCrann, N., Zuntz, J., Eifler, T. F., Krause, E., Dodelson, S., et al. (2018). Dark Energy Survey year 1 results: Cosmological constraints from cosmic shear. *Physical Review D, 98*, 043528.

Van Waerbeke, L., Mellier, Y., Erben, T., Cuillandre, J. C., Bernardeau, F., Maoli, R., et al. (2000). Detection of correlated galaxy ellipticities from CFHT data: First evidence for gravitational lensing by large-scale structures. *Astronomy and Astrophysics, 358*, 30.

White, M., & Hu, W. (2000). A new algorithm for computing statistics of weak lensing by large-scale structure. *The Astrophysical Journal, 537*, 1.

Wittman, D. M., Tyson, J. A., Kirkman, D., Dell'Antonio, I., & Bernstein, G. (2000). Detection of weak gravitational lensing distortions of distant galaxies by cosmic dark matter at large scales. *Nature, 405*, 143.

Zeldovich, Y. B. (1972). A hypothesis, unifying the structure and the entropy of the Universe. *Monthly Notices of the Royal Astronomical Society, 160*, 1P.

Zuntz, J., Sheldon, E., Samuroff, S., Troxel, M. A., Jarvis, M., MacCrann, N., et al. (2018). Dark Energy Survey year 1 results: Weak lensing shape catalogues. *Monthly Notices of the Royal Astronomical Society, 481*, 1149.

Chapter 9
Lensing of the Cosmic Microwave Background

The most distant light source that can be observed is the cosmic microwave background. Its cosmological significance is outlined in Sect. 9.1. As in the case of cosmic shear, the lensed power spectrum is the key quantity of interest. However, shear is ill-defined when the background source is the entire sky. Nevertheless, lensing "rearranges" the distribution of photons in a way that alters the observed pattern of temperature fluctuations that characterize cosmic background radiation. The spherical nature of the sky results in a power spectrum that depends on a discrete index, rather than a continuous wavenumber (Sect. 9.2). On the angular scales affected by lensing, the sky can be treated as "flat" in the vicinity of a given point. We use this approximation to derive expressions for the lensed power spectrum (Sect. 9.3.1) and the lens potential (Sect. 9.3.2). Finally, we mention applications in Sect. 9.4.

9.1 Prediction, Detection, and Significance

We noted in Sect. 3.5 that the discovery of the **cosmic microwave background** (CMB) settled the debate between the steady-state and Big Bang models, with the Big Bang emerging triumphant. The conventional narrative is that Penzias and Wilson (1965) serendipitously detected the CMB while repurposing a disused antenna for radio astronomy and that Dicke et al. (1965) simultaneously predicted the existence of the CMB on theoretical grounds. This is the truth, but not the whole truth. Alpher et al. (1948) considered the problem of nucleosynthesis at the high temperatures that would follow a Big Bang, but it was not until the work of Gamow (1948) that the context of an expanding universe was explicitly included. Alpher and Herman (1948) predicted that radiation emitted when neutral atoms "recombined" from the primordial plasma should have a blackbody spectrum with a temperature

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1_9

of ∼5 K today. Thus, the intensity of this isotropic background radiation would have its peak in the microwave band. No such signal would arise in a steady-state universe.

There were at least two reasons for the delay between prediction and detection. The Hubble constant was thought to be significantly larger than the currently accepted value, implying that the Earth was older than the universe itself. In addition, the Big Bang could only account for the abundances of the few lightest elements. Ironically, Hoyle, who championed the steady-state view, helped to unburden the Big Bang from having to explain the origin of heavy elements. He and coauthors (Burbidge et al. 1957) showed that stellar nucleosynthesis filled in the gap.

Penzias and Wilson (1965) indeed found that the CMB is isotropic, but the single frequency (4.08 GHz) to which their antenna was sensitive was insufficient to confirm a blackbody spectrum. Measurements over the following quarter century, culminating with the Cosmic Background Explorer (COBE) satellite, showed that the CMB has the expected blackbody form, with peak intensity at 160 GHz. This corresponds to a temperature of $T = 2.725$ K, a few degrees lower than the originally predicted value. For his work on the Far Infrared Absolute Spectrophotometer, the instrument on COBE that measured the CMB spectrum, John Mather shared the Nobel Prize in Physics in 2006. The nearly exact agreement between the data and the theoretical curve, shown in Fig. 9.1, establishes the CMB as an almost perfect blackbody. The COBE results were refined by the WMAP and Planck satellites.

The differential microwave radiometer onboard COBE, the development of which earned George Smoot a share of the 2006 Nobel Prize in Physics, found fractional temperature variations of $\sim 10^{-5}$ between points separated by $\geq 7^\circ$ on the

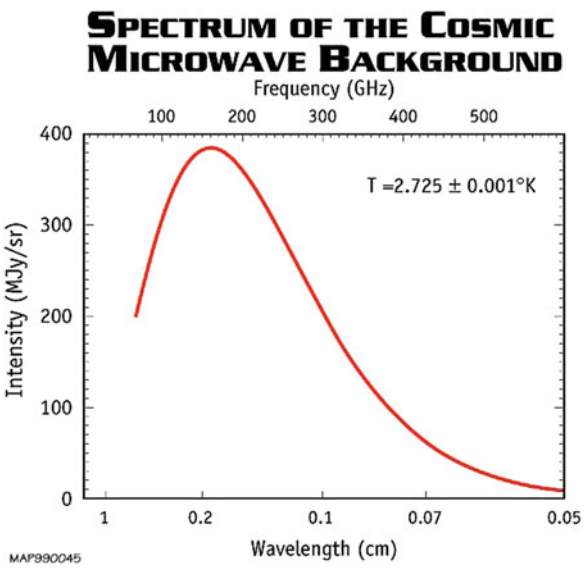


Fig. 9.1 Spectrum for a blackbody with temperature of 2.725 K. Data from the Far Infrared Absolute Spectrophotometer onboard COBE agree with the theoretical curve so well that they and their error bars are too small to see in the figure. Intensity in MJy/sr ($1\,\mathrm{Jy} \equiv 10^{-26}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$) versus frequency (top axis) or, equivalently, wavelength (bottom axis) (Credit: NASA/WMAP Science Team)

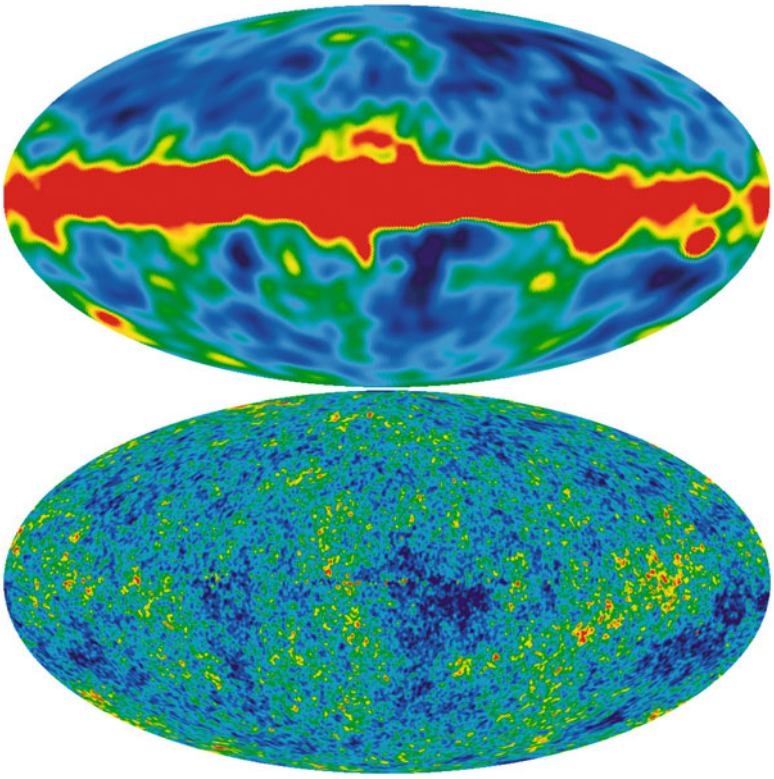


Fig. 9.2 Maps of observed CMB temperature fluctuations (after a dipole term due to the solar system's motion relative to the CMB has been removed). The celestial sphere is projected onto an ellipse, with the Milky Way lying along the major axis. The coolest and warmest points on the sky are shown in blue and red, respectively. The top panel shows a map from COBE with a temperature range of ± 100 μK; the red band in the middle shows foreground emission from the Milky Way. The bottom panel shows a higher-resolution map from WMAP with a temperature range of $\pm$ 200 μK; foreground emission from the Milky Way has been removed (Credit: NASA/WMAP Science Team)

sky (see Fig. 9.2). The successor WMAP (e.g., Bennett et al. 2013) and Planck (e.g., Planck Collaboration et al. 2018b) satellites mapped the temperature fluctuations on angular scales of $14'$ and $5'$, respectively. Partial CMB sky maps with still greater angular resolution have been obtained by the ground-based Atacama Cosmology Telescope (Das et al. 2011a) and South Pole Telescope (Keisler et al. 2011).

Anisotropy can arise from nonuniformity in the CMB itself or from matter along a given line of sight. These are known as primary and secondary anisotropies, respectively. It is conceptually convenient to picture secondary anisotropies as modifications to an intrinsically uniform CMB. Note, however, that all forms of anisotropy are ultimately due to density fluctuations in the early universe, to which

observed temperature fluctuations are proportional. We can nevertheless regard primary and secondary anisotropies as statistically independent for all practical purposes.

Primary anisotropies can be understood by considering light emitted when it decoupled from matter, close to the time of recombination. While cosmic expansion increases the wavelength of all photons by the same amount, photons of a given initial energy undergo a further redshift that depends on the depth of the potential well from which they were emitted at recombination. Thus, a lower (higher) temperature on the sky corresponds to a region of higher (lower) density on the *surface of last scattering*. Sachs and Wolfe (1967) argued on general relativistic grounds that this effect induces temperature fluctuations at the 1% level. This is counteracted to some extent by virtue of regions of higher density yielding photons of greater energy, but the **Sachs-Wolfe effect** remains detectable and important.

We now turn to secondary anisotropies. The motion of the solar system relative to the CMB reference frame induces a significant dipole term, which can be corrected for. Of cosmological significance is the amplification of inhomogeneities in the early universe, which were the seeds of objects that merged to form galaxies and clusters of galaxies. The distribution of this matter close to a particular line of sight causes the observed temperature to differ from its intrinsic value. Thus, secondary anisotropy arises from both large-scale structure and individual objects.

Large-scale structure and cosmic expansion give rise to the *integrated* Sachs-Wolfe (ISW) effect, which accounts for the cumulative change in energy of a photon as it traverses regions of varying density in an expanding universe. Suppose that a photon with initial energy E_i enters a region of size d. If the potential in the region evolves on time scales $t \lesssim d/c$, the final energy of the photon will be $E_f \neq E_i$. The dominant contribution is from the expansion of the universe rather than intrinsically time-dependent gravitational fields. The power law expansion of a matter-dominated or radiation-dominated universe is insufficient to yield a detectable ISW signal. For example, Rees and Sciama (1968) showed that the ISW temperature fluctuations are quadratic in the density contrast[1] in an Einstein-de Sitter model and thus can be ignored in comparison with first-order effects. Moreover, it is only relevant at late times, when the universe tends toward exponential expansion due to dark energy.

The most significant source of secondary anisotropy caused by discrete matter concentrations is the effect described by Sunyaev and Zeldovich (1970), in which CMB photons gain energy by interacting with hot electrons. Such "inverse" Compton scattering occurs mostly in galaxy clusters, which typically have large amounts of ionized gas. The *kinetic* and *thermal* **Sunyaev-Zeldovich effects** refer, respectively, to the increase in energy of a CMB photon by the motion of the cluster relative to the CMB rest frame and the random motion of hot electrons within the cluster. The latter contribution increases the apparent temperature of the CMB blackbody spectrum, so that the intensity of photons with energy lower (higher) than a threshold value is lower (higher) than it would otherwise be.

[1] This quantity is assumed to be small, which is valid on the large scales of interest here.

It stands to reason that density fluctuations along the line of sight to the surface of last scattering deflect photons from the CMB. Since gravitational lensing is achromatic, a CMB with a constant temperature would be unaffected by light deflection. In the presence of temperature anisotropy, however, lensing introduces additional corrections. This is because the temperature of a photon emitted from a particular direction on the sky differs from its true position. Therefore, we need to introduce the formalism for describing CMB temperature fluctuations before we can describe how the statistics of temperature fluctuations are modified by lensing.

9.2 Temperature Fluctuations

The cosmic microwave background offers an indirect probe of the matter power spectrum at the time of matter-radiation decoupling. What can actually be measured is the angular power spectrum of temperature fluctuations. The approach mirrors that of Sect. 8.2.1, except that we work on the surface of a sphere instead of in an unbounded Euclidian space. As in the case of the density contrast, we characterize the temperature fluctuations in the CMB by the fractional difference in temperature relative to the average CMB temperature, $\bar{T} = 2.725\,\mathrm{K}$. Since any scalar function on the sphere can be expressed as an infinite linear combination of spherical harmonics, we write the temperature fluctuations field as

$$\tau(\theta, \phi) \equiv \frac{T(\theta, \phi) - \bar{T}}{\bar{T}} = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta, \phi)\,. \tag{9.1}$$

The spherical harmonics $Y_{lm}(\theta, \phi)$ are defined in the Appendix C.3.1 and form a complete set on the sphere (see Appendix C.1). The temperature autocorrelation function is defined by

$$\langle \tau(\theta, \phi)\tau(\theta', \phi') \rangle \equiv \xi_\tau(\vartheta)\,, \tag{9.2}$$

where ϑ is the angle between the points (θ, ϕ) and (θ', ϕ'), with

$$\cos\vartheta = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi')\,. \tag{9.3}$$

The function ξ_τ depends only on ϑ according to the usual assumptions of isotropy and homogeneity.

As in our earlier discussion of the density contrast, we now wish to compute the power spectrum of temperature fluctuations. The qualitative difference between the present and earlier cases is that the former considered *spatial* fluctuations, whereas the latter deals with *angular* fluctuations. Because the sphere is finite, the power spectrum is described by a discrete index rather than a continuous variable. In close analogy with our earlier discussion of density fluctuations, we obtain the temperature power spectrum by computing

$$\begin{aligned}\langle a_{lm} a^*_{l'm'} \rangle &= \int_{\mathbb{S}^2} \mathrm{d}\Omega \int_{\mathbb{S}^2} \mathrm{d}\Omega' \, Y^*_{lm}(\theta, \phi) \, Y_{l'm'}(\theta', \phi') \langle \tau(\theta, \phi) \tau(\theta', \phi') \rangle \\ &= \int_{\mathbb{S}^2} \mathrm{d}\Omega \int_{\mathbb{S}^2} \mathrm{d}\Omega' \, Y^*_{lm}(\theta, \phi) \, Y_{l'm'}(\theta', \phi') \xi_\tau(\vartheta) \,, \end{aligned} \tag{9.4}$$

where we have used the orthonormality of spherical harmonics (C.46) to invert Eq. (9.1) to solve for the a_{lm} and $\mathrm{d}\Omega = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$ is an element of solid angle on the two-sphere $\mathbb{S}^2$. We can write the correlation function as the Legendre series

$$\xi_\tau(\vartheta) = \sum_{\lambda=0}^{\infty} C_\lambda P_\lambda(\cos\vartheta) = \sum_{\lambda=0}^{\infty} \sum_{\mu=-\lambda}^{\lambda} \frac{4\pi C_\lambda}{2\lambda + 1} Y_{\lambda\mu}(\theta, \phi) Y^*_{\lambda\mu}(\theta', \phi') \,, \tag{9.5}$$

where we have used the addition theorem for spherical harmonics (C.47). Substituting this expression into Eq. (9.4) gives

$$\begin{aligned}\langle a_{lm} a^*_{l'm'} \rangle &= \sum_{\lambda=0}^{\infty} \sum_{\mu=-\lambda}^{\lambda} \frac{4\pi C_\lambda}{2\lambda + 1} \int_{\mathbb{S}^2} \mathrm{d}\Omega \, Y_{\lambda\mu}(\theta, \phi) Y^*_{lm}(\theta, \phi) \\ &\quad \times \int_{\mathbb{S}^2} \mathrm{d}\Omega' \, Y^*_{\lambda\mu}(\theta', \phi') Y_{l'm'}(\theta', \phi') \\ &= \sum_{\lambda=0}^{\infty} \sum_{\mu=-\lambda}^{\lambda} \frac{4\pi C_\lambda}{2\lambda + 1} \delta_{\lambda l} \delta_{\mu m} \, \delta_{\lambda l'} \delta_{\mu m'} \\ &= \frac{4\pi C_l}{2l + 1} \delta_{ll'} \delta_{mm'} \,, \end{aligned} \tag{9.6}$$

where the power spectrum is given by

$$C_l = \frac{2}{2l + 1} \int_0^\pi \xi_\tau(\theta) P_l(\cos\theta) \sin\theta \, \mathrm{d}\theta \,. \tag{9.7}$$

Observed values of a quantity proportional to C_l are plotted against both l and the corresponding angular scale in Fig. 9.3. The most prominent feature in the power spectrum is the sharp peak at $l \sim 200$, corresponding to an angular scale of $\sim 1°$. This arises from what are essentially sound waves in the early universe. Since dark matter has no pressure, such waves are referred to as *baryon acoustic oscillations*. The higher harmonics, which occur at multiples of the fundamental mode ($l \sim 200$), contribute smaller peaks. The anisotropies described in Sect. 9.1 are also revealed by the power spectrum but in a less pronounced way.

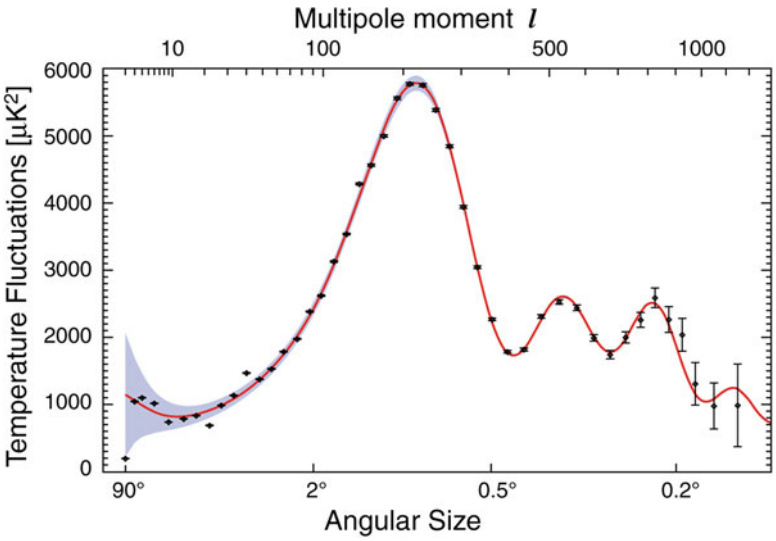


Fig. 9.3 Power spectrum of temperature fluctuations shown in the bottom panel of Fig. 9.2. The quantity on the vertical axis is proportional to C_l defined by Eq. (9.7). The horizontal axes give the multipole order l (top axis) and the corresponding angular scale (bottom axis). WMAP data with error bars are indicated along with the best fit ΛCDM model (solid line) (Credit: NASA/WMAP Science Team)

9.3 Lensed CMB

Because the CMB is lensed, the temperature contrast seen at $\boldsymbol{\theta}$ actually reflects the contrast at $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})$. Thus, the lensed temperature fluctuations $\mathscr{T}(\boldsymbol{\theta})$ can be written in terms of the intrinsic fluctuations as

$$\mathscr{T}(\boldsymbol{\theta}) = \tau(\boldsymbol{\theta} - \boldsymbol{\alpha}). \tag{9.8}$$

In this section we examine how lensing affects the statistics of temperature fluctuations (Sect. 9.3.1) and then consider how to use observed temperature fluctuations to reconstruct the lensing potential (Sect. 9.3.2). To simplify the presentation, we work in the flat sky approximation, so we can work with Fourier modes rather than spherical harmonics. This approach is reasonable on small scales and captures key conceptual results. Our presentation follows Hu and Okamoto (2002) and Lewis and Challinor (2006). Okamoto and Hu (2003) give the generalization to the full, curved sky (also see Problem 9.1).

The analysis begins with the Fourier transform of the lensed temperature fluctuations:

$$\tilde{\mathscr{T}}(\mathbf{l}) = \int_{\mathbb{R}^2} \mathrm{d}^2\theta \; \tau(\boldsymbol{\theta} - \boldsymbol{\alpha}) e^{-i\mathbf{l}\cdot\boldsymbol{\theta}}$$

$$= \int_{\mathbb{R}^2} \mathrm{d}^2\theta \left[\int_{\mathbb{R}^2} \frac{\mathrm{d}^2\ell}{(2\pi)^2} \tilde{\tau}(\boldsymbol{\ell}) e^{i\boldsymbol{\ell}\cdot(\boldsymbol{\theta}-\boldsymbol{\alpha})} \right] e^{-i\mathbf{l}\cdot\boldsymbol{\theta}}$$

$$= \int_{\mathbb{R}^2} \frac{\mathrm{d}^2\ell}{(2\pi)^2} \tilde{\tau}(\boldsymbol{\ell}) \int_{\mathbb{R}^2} \mathrm{d}^2\theta \; e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}} e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\alpha}} \,. \tag{9.9}$$

Since large-scale lensing effects are expected to be small in amplitude, it is instructive to expand $e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\alpha}}$ in a Taylor series. Working to second order in $\boldsymbol{\alpha}$ yields

$$\tilde{\mathscr{T}}(\mathbf{l}) \approx \int_{\mathbb{R}^2} \frac{\mathrm{d}^2\ell}{(2\pi)^2} \tilde{\tau}(\boldsymbol{\ell}) \int_{\mathbb{R}^2} \mathrm{d}^2\theta \; e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}} \left[1 - i\boldsymbol{\ell}\cdot\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})^2 \right] . \tag{9.10}$$

Now we express $\boldsymbol{\alpha}(\boldsymbol{\theta})$ in terms of its Fourier transform and simplify, evaluating the $\boldsymbol{\theta}$ integral using $\int_{\mathbb{R}^2} e^{-i\mathbf{l}\cdot\boldsymbol{\theta}} \mathrm{d}^2\theta = (2\pi)^2 \delta^{(2)}(\mathbf{l})$:

$$\tilde{\mathscr{T}}(\mathbf{l}) \approx \tilde{\tau}(\mathbf{l}) - i \int_{\mathbb{R}^2} \frac{\mathrm{d}^2\ell}{(2\pi)^2} \tilde{\tau}(\boldsymbol{\ell}) \left[\boldsymbol{\ell}\cdot\tilde{\boldsymbol{\alpha}}(\mathbf{l}-\boldsymbol{\ell}) \right] \tag{9.11}$$

$$- \frac{1}{2} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \frac{\mathrm{d}^2\ell_1}{(2\pi)^2} \frac{\mathrm{d}^2\ell_2}{(2\pi)^2} \tilde{\tau}(\boldsymbol{\ell}_1) \left[\boldsymbol{\ell}_1\cdot\tilde{\boldsymbol{\alpha}}(\boldsymbol{\ell}_2) \right] \left[\boldsymbol{\ell}_1\cdot\tilde{\boldsymbol{\alpha}}(\mathbf{l}-\boldsymbol{\ell}_1-\boldsymbol{\ell}_2) \right] .$$

The key conceptual result is that lensing mixes modes: mode $\mathbf{l}$ in the lensed map depends not only on mode $\mathbf{l}$ in the unlensed maps but also on some range of modes controlled by the width of $\tilde{\boldsymbol{\alpha}}$ in Fourier space.

9.3.1 Lensed Power Spectrum

We first study the power spectrum of the lensed CMB. Using Eq. (9.9), the correlation between modes in the lensed map can be written as

$$\left\langle \tilde{\mathscr{T}}(\mathbf{l}) \tilde{\mathscr{T}}^*(\mathbf{l}') \right\rangle = \left\langle \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \frac{\mathrm{d}^2\ell}{(2\pi)^2} \frac{\mathrm{d}^2\ell'}{(2\pi)^2} \tilde{\tau}(\boldsymbol{\ell}) \tilde{\tau}^*(\boldsymbol{\ell}') \right. \tag{9.12}$$

$$\left. \times \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \mathrm{d}^2\theta \, \mathrm{d}^2\theta' \; e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}} e^{i(\mathbf{l}'-\boldsymbol{\ell}')\cdot\boldsymbol{\theta}'} e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\alpha}} e^{i\boldsymbol{\ell}'\cdot\boldsymbol{\alpha}'} \right\rangle .$$

Here we take ensemble averages over both the temperature fluctuations and the deflection field, and we assume they are statistically independent because the temperature fluctuations are generated in the early universe, while the deflection field arises from a combination of density fluctuations at late times. The ensemble average over intrinsic temperature fluctuations yields the power spectrum of temperature fluctuations according to $\langle \tilde{\tau}(\boldsymbol{\ell}) \tilde{\tau}^*(\boldsymbol{\ell}') \rangle = (2\pi)^2 \delta^{(2)}(\boldsymbol{\ell} - \boldsymbol{\ell}') P_\tau(\ell)$. Thus we can write

$$\left\langle \tilde{\mathscr{T}}(\mathbf{l})\tilde{\mathscr{T}}^*(\mathbf{l}')\right\rangle = \int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}P_\tau(\ell)\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'}\left\langle e^{i\boldsymbol{\ell}\cdot(\boldsymbol{\alpha}'-\boldsymbol{\alpha})}\right\rangle . \tag{9.13}$$

As before, we make a Taylor series expansion to second order in $\boldsymbol{\alpha}$:

$$\begin{aligned}\left\langle \tilde{\mathscr{T}}(\mathbf{l})\tilde{\mathscr{T}}^*(\mathbf{l}')\right\rangle \approx \int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}P_\tau(\ell)\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'} \\ \times\Big\langle 1 - i\boldsymbol{\ell}\cdot\boldsymbol{\alpha} + i\boldsymbol{\ell}\cdot\boldsymbol{\alpha}' \\ -\frac{1}{2}\Big[(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})^2 + (\boldsymbol{\ell}\cdot\boldsymbol{\alpha}')^2 - 2(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})(\boldsymbol{\ell}\cdot\boldsymbol{\alpha}')\Big]\Big\rangle .\end{aligned} \tag{9.14}$$

Because the deflection field has mean zero, the linear terms vanish in the ensemble average. For the quadratic terms, the analysis proceeds as follows. We express each $\boldsymbol{\alpha}$ in terms of its Fourier transform and then convert to the Fourier transform of the lensing potential using $\tilde{\boldsymbol{\alpha}}(\boldsymbol{\ell}) = i\boldsymbol{\ell}\tilde{\psi}(\boldsymbol{\ell})$ (which follows from $\boldsymbol{\alpha} = \nabla\psi$). Consider the term with $\langle(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})^2\rangle$:

$$\begin{aligned}&\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'}\left\langle(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})^2\right\rangle \\ &\quad= -\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'} \\ &\qquad\times\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell_1}{(2\pi)^2}\frac{\mathrm{d}^2\ell_2}{(2\pi)^2}(\boldsymbol{\ell}\cdot\boldsymbol{\ell}_1)(\boldsymbol{\ell}\cdot\boldsymbol{\ell}_2)\left\langle\tilde{\psi}(\boldsymbol{\ell}_1)\tilde{\psi}(\boldsymbol{\ell}_2)\right\rangle e^{i\boldsymbol{\ell}_1\cdot\boldsymbol{\theta}}e^{i\boldsymbol{\ell}_2\cdot\boldsymbol{\theta}} \\ &\quad= (2\pi)^2\delta^{(2)}(\mathbf{l}-\boldsymbol{\ell})\delta^{(2)}(\mathbf{l}'-\boldsymbol{\ell})\int_{\mathbb{R}^2}\mathrm{d}^2\ell_1\,(\boldsymbol{\ell}\cdot\boldsymbol{\ell}_1)^2P_\psi(\boldsymbol{\ell}_1)\,.\end{aligned} \tag{9.15}$$

In the last step, we introduce the power spectrum of the potential fluctuations, P_ψ, using $\langle\tilde{\psi}(\boldsymbol{\ell}_1)\tilde{\psi}(\boldsymbol{\ell}_2)\rangle = (2\pi)^2\delta^{(2)}(\boldsymbol{\ell}_1+\boldsymbol{\ell}_2)P_\psi(\boldsymbol{\ell}_1)$. The analysis is essentially identical for the term with $\langle(\boldsymbol{\ell}\cdot\boldsymbol{\alpha}')^2\rangle$. That leaves the term with $\langle(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})(\boldsymbol{\ell}\cdot\boldsymbol{\alpha}')\rangle$:

$$\begin{aligned}&\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'}\left\langle(\boldsymbol{\ell}\cdot\boldsymbol{\alpha})(\boldsymbol{\ell}\cdot\boldsymbol{\alpha}')\right\rangle \\ &\quad= -\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'} \\ &\qquad\times\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell_1}{(2\pi)^2}\frac{\mathrm{d}^2\ell_2}{(2\pi)^2}(\boldsymbol{\ell}\cdot\boldsymbol{\ell}_1)(\boldsymbol{\ell}\cdot\boldsymbol{\ell}_2)\left\langle\tilde{\psi}(\boldsymbol{\ell}_1)\tilde{\psi}(\boldsymbol{\ell}_2)\right\rangle e^{i\boldsymbol{\ell}_1\cdot\boldsymbol{\theta}}e^{i\boldsymbol{\ell}_2\cdot\boldsymbol{\theta}'} \\ &\quad= (2\pi)^2\int_{\mathbb{R}^2}\mathrm{d}^2\ell_1\,\delta^{(2)}(\mathbf{l}-\boldsymbol{\ell}-\boldsymbol{\ell}_1)\delta^{(2)}(\mathbf{l}'-\boldsymbol{\ell}-\boldsymbol{\ell}_1)(\boldsymbol{\ell}\cdot\boldsymbol{\ell}_1)^2P_\psi(\boldsymbol{\ell}_1)\,.\end{aligned} \tag{9.16}$$

Putting the pieces together back in Eq. (9.14), we obtain

$$\left\langle \tilde{\mathscr{T}}(\mathbf{l})\tilde{\mathscr{T}}^*(\mathbf{l}')\right\rangle \approx (2\pi)^2\delta^{(2)}(\mathbf{l}-\mathbf{l}')P_\tau(l)\left[1-\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}(\mathbf{l}\cdot\boldsymbol{\ell})^2P_\psi(\boldsymbol{\ell})\right] + (2\pi)^2\delta^{(2)}(\mathbf{l}-\mathbf{l}')\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}[\boldsymbol{\ell}\cdot(\mathbf{l}-\boldsymbol{\ell})]^2P_\tau(\boldsymbol{\ell})P_\psi(\mathbf{l}-\boldsymbol{\ell})\,. \tag{9.17}$$

The correlation is still diagonal (i.e., it vanishes if $\mathbf{l}\neq\mathbf{l}'$), which is a consequence of the fact that both the temperature and density fluctuations are statistically homogeneous and isotropic. Therefore, we can write $\langle\tilde{\mathscr{T}}(\mathbf{l})\tilde{\mathscr{T}}^*(\mathbf{l}')\rangle = (2\pi)^2\delta^{(2)}(\mathbf{l}-\mathbf{l}')P_{\mathscr{T}}(l)$ where the lensed power spectrum involves integrals over the unlensed power spectrum:

$$\begin{aligned} P_{\mathscr{T}}(l) &\approx P_\tau(l)\left[1-\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}(\mathbf{l}\cdot\boldsymbol{\ell})^2P_\psi(\boldsymbol{\ell})\right] \\ &\quad+\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}[\boldsymbol{\ell}\cdot(\mathbf{l}-\boldsymbol{\ell})]^2P_\tau(\boldsymbol{\ell})P_\psi(\mathbf{l}-\boldsymbol{\ell}) \\ &\approx P_\tau(l)\left[1-l^2\int_0^\infty\frac{\mathrm{d}\ell}{4\pi}\,\ell^3P_\psi(\ell)\right] \\ &\quad+\int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}[\boldsymbol{\ell}\cdot(\mathbf{l}-\boldsymbol{\ell})]^2P_\tau(\boldsymbol{\ell})P_\psi(\mathbf{l}-\boldsymbol{\ell})\,. \end{aligned} \tag{9.18}$$

In the first term, we have used the fact that P_ψ is isotropic to carry out the angular integration. We see that the lensed power spectrum has one term that is the unlensed power spectrum with a correction term that scales as l^2, along with a second term in which the unlensed power spectrum is convolved with the power spectrum of the lensing potential.

9.3.2 Reconstructing the Lensing Potential

Equation (9.18) shows how lensing affects the power spectrum of temperature fluctuations, but it is not very practical if we want to measure the lensing potential because P_ψ is buried inside integrals. For an alternate approach, we consider that our universe actually has one realization of the lensing potential, so we can omit the ensemble average over the lensing terms in Eq. (9.14). Then the terms that are linear in $\boldsymbol{\alpha}$ no longer vanish, so to lowest order we have

$$\left\langle \tilde{\mathcal{T}}(\mathbf{l})\tilde{\mathcal{T}}^*(\mathbf{l}')\right\rangle \approx \int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}P_\tau(\ell) \times \int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathrm{d}^2\theta\,\mathrm{d}^2\theta'\,e^{-i(\mathbf{l}-\boldsymbol{\ell})\cdot\boldsymbol{\theta}}e^{i(\mathbf{l}'-\boldsymbol{\ell})\cdot\boldsymbol{\theta}'}\left[1-i\boldsymbol{\ell}\cdot\boldsymbol{\alpha}+i\boldsymbol{\ell}\cdot\boldsymbol{\alpha}'\right]. \tag{9.19}$$

Each integral over $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ yields a Fourier transform if it contains a factor of $\boldsymbol{\alpha}$ or $\boldsymbol{\alpha}'$ (respectively) or a δ function otherwise:

$$\left\langle \tilde{\mathcal{T}}(\mathbf{l})\tilde{\mathcal{T}}^*(\mathbf{l}')\right\rangle \approx \int_{\mathbb{R}^2}\frac{\mathrm{d}^2\ell}{(2\pi)^2}P_\tau(\ell)\Big[(2\pi)^4\delta^{(2)}(\mathbf{l}-\boldsymbol{\ell})\delta^{(2)}(\mathbf{l}'-\boldsymbol{\ell}) -i\boldsymbol{\ell}\cdot\tilde{\boldsymbol{\alpha}}(\mathbf{l}-\boldsymbol{\ell})(2\pi)^2\delta^{(2)}(\mathbf{l}'-\boldsymbol{\ell})+i\boldsymbol{\ell}\cdot\tilde{\boldsymbol{\alpha}}(\boldsymbol{\ell}-\mathbf{l}')(2\pi)^2\delta^{(2)}(\mathbf{l}-\boldsymbol{\ell})\Big]. \tag{9.20}$$

Now we use the δ functions to evaluate the $\boldsymbol{\ell}$ integral, and we relate the Fourier transform of the deflection field to the Fourier transform of the lensing potential by $\tilde{\boldsymbol{\alpha}}(\boldsymbol{\ell}) = i\boldsymbol{\ell}\tilde{\psi}(\boldsymbol{\ell})$. Simplifying yields

$$\left\langle \tilde{\mathcal{T}}(\mathbf{l})\tilde{\mathcal{T}}(\mathbf{l}')^*\right\rangle \approx (2\pi)^2\delta^{(2)}(\mathbf{l}-\mathbf{l}')P_\tau(l)+(\mathbf{l}-\mathbf{l}')\cdot\left[-\mathbf{l}P_\tau(l)+\mathbf{l}'P_\tau(l')\right]\tilde{\psi}(\mathbf{l}-\mathbf{l}'). \tag{9.21}$$

There are two important conceptual results here. First, without lensing the correlation is diagonal (the first term has $\delta^{(2)}(\mathbf{l}-\mathbf{l}')$), but lensing creates nonzero off-diagonal terms. Second, if we know the unlensed power spectrum $P_\tau(l)$ (from measurements on scales where lensing has little effect, combined with cosmological models), then we can invert Eq. (9.21) to infer $\tilde{\psi}(\mathbf{l})$ and hence measure the lensing power spectrum $P_\psi(l)$.

9.4 Applications

For practical applications, the theory needs to be extended in two ways. First, a naive inversion of Eq. (9.21) would be sensitive to noise in the measurements. Hu and Okamoto (2002) construct statistical estimators for $\tilde{\psi}$ that are well behaved in the presence of noise, and they show how to use information from polarization of the CMB to supplement the information from temperature fluctuations and improve the measurement. Second, the analysis needs to account for the fact that the sky is not flat. Okamoto and Hu (2003) present a comprehensive analysis using spherical harmonics; the details are complicated, but the conceptual result is similar to what we learned above: the correlation matrix of modes in the lensed CMB has off-diagonal terms that are proportional to the lensing potential. Measurements of the off-diagonal terms can therefore be used to reconstruct the lensing potential, and that reconstruction can then be used to compute the lensing power spectrum.

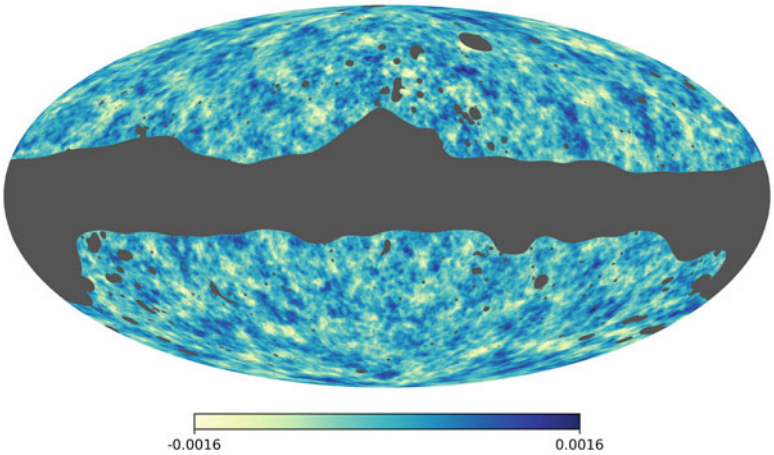


Fig. 9.4 Full-sky map of the lensing reconstruction from *Planck* measurements of the CMB. The color scale shows the amplitude of the deflection or the gradient of the reconstructed potential field. Gray regions are masked out due to foreground contamination. Figure courtesy J. Carron (see Planck Collaboration et al. 2018a)

To date, measurements of CMB lensing have been reported by five different experiments: the Atacama Cosmology Telescope (Das et al. 2011b; Sherwin et al. 2017), the South Pole Telescope (van Engelen et al. 2012; Story et al. 2015), POLARBEAR (Ade et al. 2014), BICEP2/Keck Array (BICEP2 Collaboration et al. 2016), and the *Planck* satellite (Planck Collaboration et al. 2014, 2016, 2018a). Figure 9.4 shows the *Planck* reconstruction of the lensing potential over the whole sky except for regions contaminated by light from the Milky Way galaxy. The dominant features are generally larger in scale than the features in the map of temperature fluctuations (compare Fig. 9.2). As a result, Fig. 9.5 shows that the power spectrum of the lensing potential peaks at lower multipole orders than the power spectrum of temperature fluctuations (compare Fig. 9.3). The measurements agree well with predictions for a ΛCDM cosmology, and they strengthen to some extent constraints on cosmological parameters. In the future, as measurements improve, CMB lensing is expected to play an important role in probing dark matter, dark energy, and fundamental physics such as neutrino masses by essentially mapping all of the matter in the universe (see Abazajian et al. 2016).

Problems

9.1 This problem explores power spectra for the curved sky and the flat sky approximation. Consider a Gaussian function defined for the curved sky. Without loss of generality, we can choose spherical coordinates such that the north pole is located at the center of the Gaussian and then write

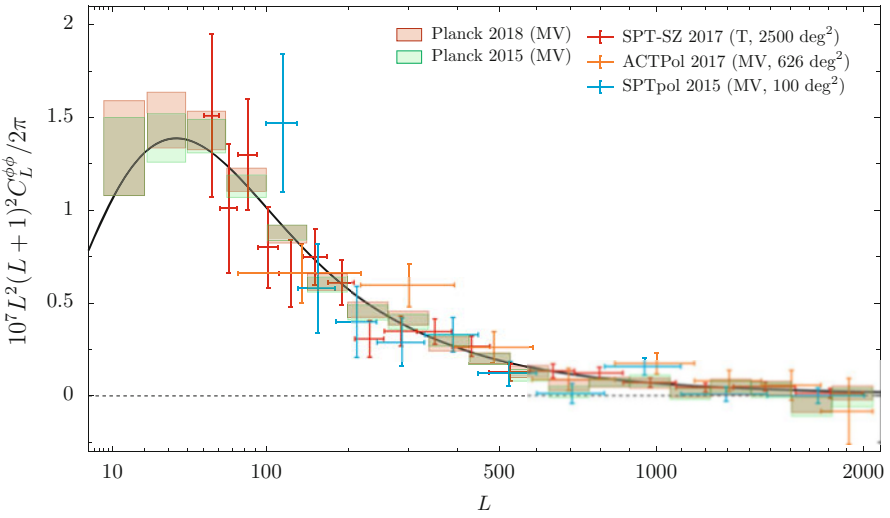


Fig. 9.5 Constraints on the power spectrum of the lensing potential. Green and red boxes show the *Planck* 2015 and 2018 results, respectively (Planck Collaboration et al. 2016, 2018a). Results are also shown for ground-based observations with the Atacama Cosmology Telescope (ACT, orange; Sherwin et al. 2017) and the South Pole Telescope (SPT, blue and red; Story et al. 2015; Simard et al. 2018). The solid black line shows the prediction for a standard ΛCDM cosmology. Figure courtesy J. Carron (see Planck Collaboration et al. 2018a)

$$f_{\text{curved}}(\theta, \phi) = \exp\left(-\frac{\theta^2}{2\sigma^2}\right), \tag{9.22}$$

where σ represents the width of the Gaussian. For the flat sky approximation, consider the tangent plane, or gnomonic, projection: each point on the sphere is projected along a radial line until it reaches the plane tangent to the north pole. Then coordinates in the tangent plane are given by $x = \tan\theta \cos\phi$ and $y = \tan\theta \sin\phi$, so the function on the flat sky is

$$f_{\text{flat}}(x, y) = \exp\left[-\frac{(\tan^{-1}\sqrt{x^2+y^2})^2}{2\sigma^2}\right]. \tag{9.23}$$

The goal now is to compute the power spectra using a spherical harmonic decomposition for f_{curved} and a Fourier decomposition for f_{flat}.

(a) Conceptually, do you expect the curved and flat sky power spectra to be more similar when σ is small or large?

(b) For different values of σ, compute the spherical harmonic decomposition, $f_{\text{curved}}(\theta, \phi) = \sum_{lm} a_{lm} Y_{lm}(\theta, \phi)$. What can you say about a_{lm} for $m \neq 0$? Plot the power spectrum, $P_{\text{curved}}(l) = \sum_{m=-l}^{l} |a_{lm}|^2$ for the different values of σ.

(c) Now compute the Fourier transform, $\tilde{f}_{\rm flat}(\mathbf{l})$, of $f_{\rm flat}(x, y)$. What can you say about the symmetry properties of $\tilde{f}_{\rm flat}(\mathbf{l})$? Plot the radial power spectrum,[2] $P_{\rm flat}(l) = l|\tilde{f}_{\rm flat}(l)|^2$, for the different values of σ. Note: the Fourier transform integral over $\mathbb{R}^2$ formally diverges, so limit the radial integral to some finite radius $r_{\rm max}$. How much do the results depend on $r_{\rm max}$?

(d) How do the curved and flat sky power spectra compare for the different values of σ?

Hint: it may be helpful to use symbolic or numerical mathematics software.

9.2 While the body of this chapter focused on lensing of the CMB by large-scale structure, there can also be features associated with individual galaxy clusters. Take the cluster lens model from Problem 7.5 and rescale it so the source is the CMB. On the angular scales that are relevant, we can approximate the CMB using a Taylor series expansion in position. If we work at zeroth order so the CMB is locally uniform, what would it look like when lensed? How about if we work at first order so the gradient is uniform? How does the direction of the gradient affect what we see? Make appropriate plots to illustrate the effects.

References

Abazajian, K. N., Adshead, P., Ahmed, Z., Allen, S. W., Alonso, D., Arnold, K. S., et al. (2016). *CMB-S4 science book* (1st ed.). arXiv e-prints, arXiv:1610.02743.

Ade, P. A. R., Akiba, Y., Anthony, A. E., Arnold, K., Atlas, M., Barron, D., et al. (2014). Measurement of the cosmic microwave background polarization lensing power spectrum with the POLARBEAR experiment. *Physical Review Letters, 113*, 021301.

Alpher, R. A., Bethe, H., & Gamow, G. (1948). The origin of chemical elements. *Physical Review, 73*, 803.

Alpher, R. A., & Herman, R. (1948). Evolution of the universe. *Nature, 162*, 774.

Bennett, C. L., Larson, D., Weiland, J. L., Jarosik, N., Hinshaw, G., Odegard, N., et al. (2013). Nine-year Wilkinson microwave anisotropy probe (WMAP) observations: Final maps and results. *The Astrophysical Journal Supplement Series, 208*, 20.

BICEP2 Collaboration, Keck Array Collaboration, Ade, P. A. R., Ahmed, Z., Aikin, R. W., Alexander, K. D., et al. (2016). BICEP2/Keck Array VIII: Measurement of gravitational lensing from large-scale B-mode polarization. *The Astrophysical Journal, 833*, 228.

Burbidge, E. M., Burbidge, G. R., Fowler, W. A., & Hoyle, F. (1957). Synthesis of the elements in stars. *Reviews of Modern Physics, 29*, 547.

Das, S., Marriage, T. A., Ade, P. A. R., Aguirre, P., Amiri, M., Appel, J. W., et al. (2011a). The Atacama Cosmology Telescope: A measurement of the cosmic microwave background power spectrum at 148 and 218 GHz from the 2008 Southern Survey. *The Astrophysical Journal, 729*, 62.

Das, S., Sherwin, B. D., Aguirre, P., Appel, J. W., Bond, J. R., Carvalho, C. S., et al. (2011b). Detection of the power spectrum of cosmic microwave background lensing by the Atacama Cosmology Telescope. *Physical Review Letters, 107*, 021301.

[2] This is defined such that the total power can be written as $P_{\rm flat} = \int_0^\infty P_{\rm flat}(l)\,{\rm d}l$.

Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. (1965). Cosmic black-body radiation. *The Astrophysical Journal, 142*, 414.

Gamow, G. (1948). The evolution of the universe. *Nature, 162*, 680.

Hu, W., & Okamoto, T. (2002). Mass reconstruction with cosmic microwave background polarization. *The Astrophysical Journal, 574*, 566.

Keisler, R., Reichardt, C. L., Aird, K. A., Benson, B. A., Bleem, L. E., Carlstrom, J. E., et al. (2011). A measurement of the damping tail of the cosmic microwave background power spectrum with the South Pole Telescope. *The Astrophysical Journal, 743*, 28.

Lewis, A., & Challinor, A. (2006). Weak gravitational lensing of the CMB. *Physics Reports, 429*, 1.

Okamoto, T., & Hu, W. (2003). Cosmic microwave background lensing reconstruction on the full sky. *Physical Review D, 67*, 083002

Penzias, A. A., & Wilson, R. W. (1965). A measurement of excess antenna temperature at 4080 Mc/s. *The Astrophysical Journal, 142*, 419.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., et al. (2014). Planck 2013 results. XVII. Gravitational lensing by large-scale structure. *Astronomy & Astrophysics, 571*, A17.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., et al. (2016). Planck 2015 results. XV. Gravitational lensing. *Astronomy & Astrophysics, 594*, A15.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., et al. (2018a). Planck 2018 results. VIII. Gravitational lensing. arXiv e-prints, arXiv:1807.06210.

Planck Collaboration, Akrami, Y., Arroja, F., Ashdown, M., Aumont, J., Baccigalupi, C., et al. (2018b). Planck 2018 results. I. Overview and the cosmological legacy of Planck. arXiv e-prints, arXiv:1807.06205.

Rees, M. J., & Sciama, D. W. (1968). Large-scale density inhomogeneities in the universe. *Nature, 217*, 511.

Sachs, R. K., & Wolfe, A. M. (1967). Perturbations of a cosmological model and angular variations of the microwave background. *The Astrophysical Journal, 147*, 73.

Sherwin, B. D., van Engelen, A., Sehgal, N., Madhavacheril, M., Addison, G. E., Aiola, S., et al. (2017). Two-season Atacama Cosmology Telescope polarimeter lensing power spectrum. *Physical Review D, 95*, 123529.

Simard, G., Omori, Y., Aylor, K., Baxter, E. J., Benson, B. A., Bleem, L. E., et al. (2018). Constraints on cosmological parameters from the angular power spectrum of a combined 2500 deg^2 SPT-SZ and Planck Gravitational Lensing Map. *The Astrophysical Journal, 860*, 137.

Story, K. T., Hanson, D., Ade, P. A. R., Aird, K. A., Austermann, J. E., Beall, J. A., et al. (2015). A measurement of the cosmic microwave background gravitational lensing potential from 100 square degrees of SPTpol data. *The Astrophysical Journal, 810*, 50.

Sunyaev, R. A., & Zeldovich, Y. B. (1970). Small-scale fluctuations of relic radiation. *Astrophysics and Space Science, 7*, 3.

van Engelen, A., Keisler, R., Zahn, O., Aird, K. A., Benson, B. A., Bleem, L. E., et al. (2012). A measurement of gravitational lensing of the microwave background using South Pole Telescope Data. *The Astrophysical Journal, 756*, 142.

Appendix A
Calculus of Variations

In many physical problems, the solution corresponds to the extremum of a function of one or more variables. Our derivation of the lens equation in Sect. 4.1.2 from Fermat's principle is an example. In the calculus of variations, one considers the more involved problem of finding the particular curve that yields an extremum of a functional or integral of a curve. Lagrangian mechanics is a reformulation of Newtonian mechanics as a problem in the calculus of variations. In this book we use methods of calculus of variations in deriving Eq. (3.60) for geodesics, which correspond to extrema of the arc length $\int \mathrm{d}s$.

In Sect. A.1, we derive the Euler-Lagrange differential equation, which determines the extremal curve for a given functional. In Sect. A.2, the equation is used to solve the brachistochrone problem, which is instructive and of historical interest. Finally, in Sect. A.3, we present the Einstein-Hilbert action and show that its extrema correspond to solutions of Einstein's field equations in a vacuum.

A.1 Euler-Lagrange Equations

Consider the functional

$$S[q] = \int_{\lambda_1}^{\lambda_2} L(q, \dot{q}; \lambda)\, \mathrm{d}\lambda\,, \tag{A.1}$$

where $q(\lambda)$ is a curve between fixed endpoints and $\dot{q} \equiv \mathrm{d}q/\mathrm{d}\lambda$. We call S the **action** and L the **Lagrangian**. Let $q^{(0)}(\lambda)$ be a path between the fixed endpoints that yields an extreme value of the functional. Now let $q = q^{(0)} + \epsilon q^{(1)}$, where $q^{(1)}$ is a path with endpoints $q^{(1)}(\lambda_1) = q^{(1)}(\lambda_2) = 0$, and ϵ is an adjustable parameter. For $q^{(0)}$ to yield a local extremum of S, $\mathrm{d}S[q]/\mathrm{d}\epsilon$ must vanish for $\epsilon = 0$. The derivative is given by

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}S[q] = \int_{\lambda_1}^{\lambda_2}\left(\frac{\partial L}{\partial q}q^{(1)} + \frac{\partial L}{\partial \dot{q}}\dot{q}^{(1)}\right)\mathrm{d}\lambda\,. \tag{A.2}$$

On integration by parts, this becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon}S[q] &= \int_{\lambda_1}^{\lambda_2}\left(\frac{\partial L}{\partial q}q^{(1)} - \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\partial L}{\partial \dot{q}}q^{(1)}\right)\mathrm{d}\lambda + \left.\frac{\partial L}{\partial \dot{q}}q^{(1)}\right|_{\lambda_1}^{\lambda_2} \\ &= \int_{\lambda_1}^{\lambda_2}\left(\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\partial L}{\partial \dot{q}}\right)q^{(1)}\,\mathrm{d}\lambda \end{aligned} \tag{A.3}$$

since $q^{(1)}(\lambda_1) = q^{(1)}(\lambda_2) = 0$. In order for $\mathrm{d}S[q^{(0)}]/\mathrm{d}\epsilon$ to vanish for *arbitrary* functions $q^{(1)}$, the quantity in parentheses in Eq. (A.3) must vanish, implying the Euler-Lagrange equation

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\partial L}{\partial \dot{q}}(q^{(0)}, \dot{q}^{(0)}; \lambda) = \frac{\partial L}{\partial q}(q^{(0)}, \dot{q}^{(0)}; \lambda)} \tag{A.4}$$

for the extremal path.

Suppose that the parameter λ is the time t and that the Lagrangian

$$L(q, \dot{q}; t) = \frac{1}{2}m\dot{q}^2 - V(q) \tag{A.5}$$

is the difference between the kinetic energy and the potential energy of a particle. For this Lagrangian, the Euler-Lagrange equation,

$$m\ddot{q} = -\frac{\mathrm{d}V}{\mathrm{d}q}\,, \tag{A.6}$$

is Newton's second law in one dimension. In three dimensions, each of the three coordinates q_i satisfies an Euler-Lagrange equation.

A.2 Brachistochrone

As an application of the Euler-Lagrange equation, consider the following problem: let A and B be two points, with B below (but not directly below) A, connected by a wire. A bead slides without friction along the wire from A to B solely under the influence of gravity, starting from rest. What shape of the wire minimizes the time to get from A to B? The optimal shape is called the **brachistochrone** or shortest time curve (see Fig. A.1).

In solving the problem, it is convenient to set $A = (0, 0)$ and choose down to be the positive direction for the vertical coordinate y, so that $B = (a, b)$ with $a > 0$ and $b > 0$. We represent the wire by a function $x(y)$, with element of arc length $\mathrm{d}s = \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2} = \sqrt{1 + x'^2}\,\mathrm{d}y$, where $x' \equiv \mathrm{d}x/\mathrm{d}y$. The speed of the particle at vertical displacement y is given by $v = \sqrt{2gy}$, as follows from conservation of kinetic plus potential energy. Thus, the time to get from A to B is

$$T[x] = \int \frac{\mathrm{d}s}{v} = \int_0^b \sqrt{\frac{1 + x'^2}{2gy}}\,\mathrm{d}y\,. \tag{A.7}$$

Substituting $L(x, x'; y) = \sqrt{(1 + x'^2)/2gy}$ in the Euler-Lagrange equation (A.4) leads to

$$\frac{\mathrm{d}}{\mathrm{d}y}\frac{\partial L}{\partial x'} - \frac{\partial L}{\partial x} = \frac{1}{\sqrt{2g}}\frac{\mathrm{d}}{\mathrm{d}y}\left[\frac{x'}{\sqrt{y(1 + x'^2)}}\right] = 0\,. \tag{A.8}$$

From Eq. (A.8), it is clear that the quantity in square brackets is a constant. Equating it to $\ell^{-1/2}$, where ℓ is a constant with the dimensions of length, and solving for x' lead to the first-order differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \pm\sqrt{\frac{y}{\ell - y}}\,. \tag{A.9}$$

It is solved by choosing the upper sign, consistent with a particle traveling downward and to the right, as in Fig. A.1, separating variables in the form

$$\int \mathrm{d}x = \int \sqrt{\frac{y}{\ell - y}}\,\mathrm{d}y\,, \tag{A.10}$$

and performing the integration with the help of the substitution $y = \ell \sin^2(\phi/2) = \frac{1}{2}\ell(1 - \cos\phi)$. Choosing the constant of integration so that the curve passes through the point $(0, 0)$, we obtain the familiar parametric equations for a cycloid,

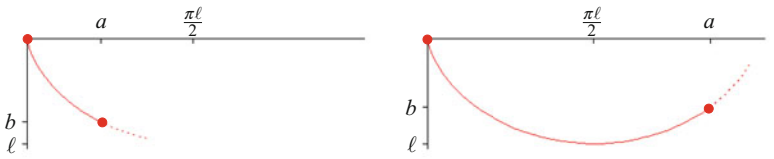


Fig. A.1 The brachistochrone (A.11) for $a/b < \pi/2$ (left) and $a/b > \pi/2$ (right). For $a/b > \pi/2$, the particle actually descends below $y = b$, before arriving at B

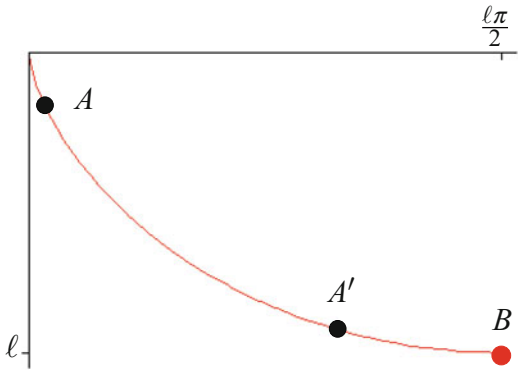


Fig. A.2 On the tautochrone, particles released from rest at different points A and A' reach B at the same time

$$\boxed{\begin{aligned} y &= \frac{1}{2}\ell(1-\cos\phi) \\ x &= \frac{1}{2}\ell(\phi-\sin\phi) \end{aligned}}\,. \tag{A.11}$$

The value of ℓ is determined by the requirement that the curve also pass through the point (a, b), which implies $a/b = (\phi - \sin\phi)/(1 - \cos\phi)$. This equation is transcendental but can be solved for ϕ numerically and has a unique solution $0 < \phi_0 < 2\pi$ for any positive b/a. The numerical value of ℓ is then $2b/(1-\cos\phi_0) > 0$.

The cycloid, also encountered in Sect. 3.5.3, has other interesting properties. For example, it has the **tautochrone** property that the time for a particle to reach the lowest point, starting from rest, is independent of its starting point (see Fig. A.2). Thus, the period of a particle sliding back and forth along a cycloidal path is independent of the amplitude of the motion. In the familiar case of motion along a circular arc, as in a simple pendulum, the period is only independent of amplitude for small oscillations.

A.3 Hilbert Action

In Sect. 3.2.3, the geodesic equation (3.60) was derived from a variational principle involving the proper time. That equation determines the trajectory $x^\lambda(\tau)$ of a particle in a gravitational field with a given metric tensor g. Here, we turn to the problem of determining the metric tensor itself and discuss the action from which g can be deduced in the special case of a vacuum. To do this, we first generalize Eq. (A.4) to the case of a *tensor* field.

For the purpose of illustration, we start with the simpler case of a *scalar* field $f(x_1, \ldots, x_n)$ and an action of the form

$$S[f] = \int \mathscr{L}(f, \{\partial_k f\})\, \mathrm{d}^n x\,, \tag{A.12}$$

where the **Lagrangian density** $\mathscr{L}$ depends on f and its partial derivatives $\partial_k f \equiv \partial f/\partial x^k$. The action is stationary when the Euler-Lagrange equation

$$\partial_k \left(\frac{\partial \mathscr{L}}{\partial(\partial_k f)} \right) = \frac{\partial \mathscr{L}}{\partial f} \tag{A.13}$$

is satisfied, where summation over k is implied.

We now turn to the action $S[\mathrm{g}]$ from which the metric tensor g in a vacuum can be derived. We expect the associated Lagrangian density to depend on the spacetime curvature, which is described by the **Riemann curvature tensor**

$$R^{\lambda}{}_{\mu\nu\rho} = \partial_\nu \Gamma^{\lambda}_{\mu\rho} - \partial_\rho \Gamma^{\lambda}_{\mu\nu} + \Gamma^{\lambda}_{\sigma\nu}\Gamma^{\sigma}_{\rho\mu} - \Gamma^{\lambda}_{\sigma\rho}\Gamma^{\sigma}_{\nu\mu}\,, \tag{A.14}$$

where the Christoffel symbols $\Gamma^{\kappa}_{\nu\rho}$ are defined by Eq. (3.61). Note that the Riemann tensor depends on g as well as its first and second partial derivatives. If all components of the Riemann tensor vanish, the spacetime described by g is flat. It is standard to define a *scalar* curvature, from which the Lagrangian density follows directly. The first step is to define the **Ricci tensor** by

$$R_{\mu\nu} = R^{\lambda}{}_{\mu\nu\lambda}\,. \tag{A.15}$$

The **Ricci scalar** is then defined by

$$R = g^{\mu\nu} R_{\mu\nu}\,. \tag{A.16}$$

The simplest action involving R is $S = \int R\,\mathrm{d}V$.

The volume element in a locally flat neighborhood of a point P is $\mathrm{d}^4x \equiv \mathrm{d}x^0\,\mathrm{d}x^1\,\mathrm{d}x^2\,\mathrm{d}x^3$. With respect to some other point P', the volume element is

$$\mathrm{d}V = \det\left(\frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\right)\mathrm{d}^4x = \lambda\,\mathrm{d}^4x\,, \tag{A.17}$$

where $\partial \mathbf{x}'/\partial \mathbf{x} \equiv \Lambda$ is a Lorentz transformation and $\lambda \equiv \det(\Lambda)$. In order for

$$\mathrm{d}s^2 = \mathrm{d}x^\mu \mathrm{d}x_\mu = \mathrm{d}x'^\mu \mathrm{d}x'_\mu = \mathrm{d}s'^2\,, \tag{A.18}$$

where $\{x^\mu\}$ and $\{x'^\mu\}$ are measured relative to P and P', we must have

$$\mathrm{g}(P') = \Lambda(P')^T \mathrm{g}(P)\Lambda(P') = \Lambda(P')^T \eta \Lambda(P')\,, \tag{A.19}$$

where η is the Minkowski metric tensor. Since $\det(\eta) = -1$, we conclude that $\lambda = \sqrt{-g}$, where $g \equiv \det(\mathrm{g})$. Thus, the volume element near any point P' is $\mathrm{d}V = \sqrt{-g}\,\mathrm{d}^4x$. Finally, the Hilbert action[1] is

$$S[\mathrm{g}] = \int R\sqrt{-g}\,\mathrm{d}^4x\,. \tag{A.20}$$

The associated Lagrangian density is then

$$\mathscr{L}(\mathrm{g}, \{\partial_\lambda \mathrm{g}\}, \{\partial_\lambda \partial_\mu \mathrm{g}\}) = R\sqrt{-g}\,. \tag{A.21}$$

Since g is a *tensor* field and $\mathscr{L}$ depends on the derivatives of g up to second order, the Euler-Lagrange equation in this case is more complicated than that for a scalar field (A.13). The details do not concern us here, so we merely quote the result:

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 0\,, \tag{A.22}$$

where G is called the **Einstein tensor**. Equations (A.22) are the **Einstein field equations** in a vacuum. Since the metric and Ricci tensors are symmetric, only ten of these sixteen equations are independent. If we multiply both sides of Eqs. (A.22) by $g^{\mu\nu}$ and sum over μ and ν, we see that $R = 0$.

We can allow for a nonzero energy density for the vacuum by letting $G_{\mu\nu} \to G_{\mu\nu} + \Lambda g_{\mu\nu}$, where Λ is called the **cosmological constant**. The Ricci scalar then becomes $R = 4\Lambda$. The associated action is obtained by letting $R \to R - 2\Lambda$ in Eq. (A.20). The metric of a flat universe with $\Lambda > 0$ is known as the de Sitter solution. It turns out that the Ricci scalar R is related to the curvature k of an FRW model by

$$R = \frac{6}{c^2a^2}(a\ddot{a} + \dot{a}^2 + c^2k)\,. \tag{A.23}$$

Thus, it is possible to have a flat universe ($k = 0$) with $R \neq 0$.

In the case of nonzero energy density, the Einstein tensor satisfies

$$G_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}\,, \tag{A.24}$$

where $T_{\mu\nu}$ are the components of the **stress-energy tensor**. For an FRW model with energy density ρc^2 and pressure p, the stress-energy tensor is diagonal, with $T_{00} = \rho c^2$ and $T_{ii} = p$ ($i = 1, 2, 3$). For a universe with a nonzero cosmological constant but without matter or radiation,

[1] A factor of $c^4/(16\pi G)$ is often included, but is irrelevant for our purposes. It becomes meaningful when generalizing to a universe with matter or other sources of energy.

$$\rho = \frac{c^2}{8\pi G}\Lambda\,, \quad p = -\rho c^2\,. \tag{A.25}$$

Combining the first of these relations with Eq. (3.128) shows that the scale factor for the de Sitter solution is $a(t) = \exp(c\sqrt{\Lambda/3}\ t)$, with $-\infty < t < \infty$. The Hubble constant for this model is then $H_0 = c\sqrt{\Lambda/3}$. The general form of Einstein's equations can be derived from a variational principle, but the details are beyond the scope of this book.

Appendix B
Functions of a Complex Variable

For problems in two dimensions, it may be convenient to work with complex numbers and complex functions. The results presented here can be found in any textbook on complex analysis. We denote the set of complex numbers by $\mathbb{C}$, and we visualize it as a plane, called the *complex plane.*

B.1 Complex Derivatives and Analytic Functions

Let $\Omega \subseteq \mathbb{C}$, and let f be a function defined on Ω. Given a point $z_0 \in \Omega$, if the limit

$$\lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h} \tag{B.1}$$

exists, we say f is (complex) differentiable at z_0, and we denote by $f'(z_0)$ the value of the limit.

If we write $z = x + iy$, then

$$f(z) = f(x + iy) = u(x, y) + iv(x, y), \tag{B.2}$$

where x, y, u, and v are real. If we restrict h to real values in the limit (B.1), we find that

$$f'(z) = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x}\,. \tag{B.3}$$

On the other hand, if h is purely imaginary, then

$$f'(z) = -i \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}\,. \tag{B.4}$$

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1

Thus, if f is differentiable at z, then u and v must satisfy the *Cauchy-Riemann equations:*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \tag{B.5a}$$

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}\,. \tag{B.5b}$$

Since the value of the limit (B.1) must be the same no matter how h approaches 0, the Cauchy-Riemann equations are a necessary but not sufficient condition for $f'(z)$ to exist.

The *complex conjugate* of z is $z^* = x - iy$. Note that z and z^* are not independent. However, if we write the (purely formal) expressions

$$\frac{\partial}{\partial x} = \frac{\partial z}{\partial x}\frac{\partial}{\partial z} + \frac{\partial z^*}{\partial x}\frac{\partial}{\partial z^*} = \frac{\partial}{\partial z} + \frac{\partial}{\partial z^*} \tag{B.6a}$$

$$\frac{\partial}{\partial y} = \frac{\partial z}{\partial y}\frac{\partial}{\partial z} + \frac{\partial z^*}{\partial y}\frac{\partial}{\partial z^*} = i\left(\frac{\partial}{\partial z} - \frac{\partial}{\partial z^*}\right) \tag{B.6b}$$

and solve, we obtain the so-called *Wirtinger derivatives*:

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right) \tag{B.7a}$$

$$\frac{\partial}{\partial z^*} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)\,. \tag{B.7b}$$

If $f(z, z^*)$ is a rational function in z and z^*, then the Wirtinger derivatives $\partial f/\partial z$ and $\partial f/\partial z^*$ can be calculated in the usual way, treating z and z^* as if they were independent. In particular, $\partial z/\partial z^* = \partial z^*/\partial z = 0$, and $\partial z/\partial z = \partial z^*/\partial z^* = 1$. It is straightforward to verify that u and v satisfy the Cauchy-Riemann equations if and only if $\partial f/\partial z^* = 0$.

We say f is *analytic*[2] at z_0 if it is differentiable at every point in some neighborhood of z_0, and we say f is *holomorphic* in a subset $A \subseteq \Omega$ if it is analytic at each point of A. If f is holomorphic on all of $\mathbb{C}$, we say it is *entire.* We caution the reader that the terms "analytic" and "holomorphic" are used in subtly different ways by different authors.

A standard theorem in complex analysis states that f is holomorphic in A if and only if the partial derivatives of u and v with respect to x and y are all continuous and satisfy the Cauchy-Riemann equations. It is easy to see that these conditions are satisfied if and only if $\partial f/\partial z$ is continuous and $\partial f/\partial z^* = 0$. For

[2]Note that our definition of an analytic function only applies to complex functions. The term is also used in the context of real functions, but with a different definition.

example, any polynomial function of z is analytic in all of $\mathbb{C}$, as are the exponential and trigonometric functions. On the other hand, the functions $p(z) = z^*$ and $q(z) = Re(z) = (z + z^*)/2$ are not analytic anywhere since $\partial p/\partial z^* = 1$ and $\partial q/\partial z^* = 1/2$. The function $r(z) = |z|^2 = zz^*$ satisfies $\partial r/\partial z^* = z$, which vanishes when $z = 0$, but at no other point. Thus, there is no point of $\mathbb{C}$ at which r is analytic. Note that q and r are real valued; the only real-valued analytic functions are constant.

B.2 Contour Integrals and Cauchy's Theorem

A *contour* in $\mathbb{C}$ is a continuous function $P : [a, b] \to \mathbb{C}$, where $a < b$ are real. Given a contour $P = Q + iR$ and a complex function $f = u + iv$, the *contour integral* of f along P is defined as

$$\begin{aligned}\int_P f(z)\mathrm{d}z &= \int_a^b f(P(t))P'(t)\mathrm{d}t \\ &= \int_a^b \left[u(P(t))Q'(t) - v(P(t))R'(t)\right]\mathrm{d}t \\ &\quad + i\int_a^b \left[v(P(t))Q'(t) + u(P(t))R'(t)\right]\mathrm{d}t\,.\end{aligned}$$

Example B.1 Let $f(z) = (z-z_0)^n$ and let $P(t) = z_0 + r(\cos t + i\sin t) = z_0 + re^{it}$ for $0 \le t \le 2\pi$. Geometrically, P is a circle of radius r centered at z_0, and

$$\int_P f(z)\mathrm{d}z = \int_0^{2\pi} r^n e^{int}\cdot ire^{it}\mathrm{d}t = ir^{n+1}\int_0^{2\pi} e^{i(n+1)t}\mathrm{d}t = \begin{cases} 0 & \text{if } n \neq -1 \\ 2\pi i & \text{if } n = -1\,. \end{cases} \tag{B.8}$$

Example B.2 Let $f(z) = z^2$. For the semicircle $P(t) = e^{it}$ with $0 \le t \le \pi$,

$$\int_P f(z)\mathrm{d}z = \int_0^{\pi} e^{2it}\cdot ie^{it}\mathrm{d}t = i\int_0^{\pi} e^{3it}\mathrm{d}t = -\frac{2}{3}\,. \tag{B.9}$$

On the other hand, if $Q(t) = 1 - t$ for $0 \le t \le 2$, then $Q(0) = P(0)$ and $Q(2) = P(\pi)$. That is, Q and P have the same endpoints. In this case,

$$\int_Q f(z)\mathrm{d}z = \int_0^2 (1-t)^2(-1)\mathrm{d}t = -\frac{2}{3}\,. \tag{B.10}$$

Example B.2 illustrates the general principle that if f is holomorphic in some set $\Omega \subseteq \mathbb{C}$, if P and Q are two contours in Ω with the same endpoints, and if P can be continuously deformed into Q within Ω while keeping the endpoints fixed, then

$\int_P f(z)\mathrm{d}z = \int_Q f(z)\mathrm{d}z$. The contours P and Q are said to be *homologous* in this case. One consequence is that if f is holomorphic in Ω, and if the closed contour P is homologous to the constant contour at $P(a)$, then $\int_P f(z)\mathrm{d}z = 0$. This result is called Cauchy's theorem. Note that Example B.1 does not contradict this statement, since the function $f(z) = 1/(z - z_0)$ is not analytic (nor even defined) at $z = z_0$.

An important property of complex functions is that if the function f is analytic at some point z_0, then it has derivatives of all orders at z_0.[3] This fact follows from Cauchy's integral formula:

$$f(z_0) = \frac{1}{2\pi i}\int_P \frac{f(z)}{z - z_0}\mathrm{d}z\,, \tag{B.11}$$

where P is any positively oriented simple closed curve, z_0 lies inside P, and f is analytic at every point on or inside P. Note that as a function of z_0, $f(z)/(z - z_0)$ is differentiable at each point on P, and so

$$\begin{aligned} f'(z_0) &= \frac{1}{2\pi i}\int_P \frac{\mathrm{d}}{\mathrm{d}z_0}\frac{f(z)}{z - z_0}\mathrm{d}z \\ &= \frac{1}{2\pi i}\int_P \frac{f(z)}{(z - z_0)^2}\mathrm{d}z\,. \end{aligned}$$

Note that $f(z)/(z - z_0)^2$ is also differentiable as a function of z_0. In fact, we can conclude that

$$\frac{\mathrm{d}^n}{\mathrm{d}z_0^n} f(z_0) = \frac{n!}{2\pi i}\int_P \frac{f(z)}{(z - z_0)^{n+1}}\mathrm{d}z\,. \tag{B.12}$$

In particular, f has a *Taylor series* centered at z_0:

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n\,, \tag{B.13}$$

where $a_n = \frac{1}{n!} f^{(n)}(z_0)$. In fact, this Taylor series converges to $f(z)$ at each point z in some neighborhood of z_0.[4]

Note that if $f = u + iv$ is analytic, then $\mathrm{d}^2 f/\mathrm{d}z^2$ exists, and so the second partial derivatives of u and v with respect to x and y all exist. It then follows from the Cauchy-Riemann equations (B.5) that u and v satisfy *Laplace's equation* in two variables. That is,

[3] By contrast, the real function $f(x) = x|x|$ is differentiable everywhere, but $f''(0)$ does not exist.
[4] The analogous statement for real functions is not true. For example, if $F(x) = e^{-1/x^2}$ for $x \neq 0$ and $F(0) = 0$, then $F^{(n)}(0) = 0$ for all n, which means its Taylor series converges to 0 for all x; but $F(x)$ is nonzero for all $x \neq 0$.

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \,. \tag{B.14}$$

A function whose second partial derivatives are continuous and which satisfies Laplace's equation is said to be *harmonic*. It can be shown that a function of x and y is harmonic if and only if it is the real (imaginary) part of an analytic function.

B.3 Meromorphic Functions and the Residue Theorem

We now turn to functions that are analytic at each point of an open set Ω except for one, which we denote by z_0. For example, the function $s(z) = (z - z_0)^{-2}$ is analytic at every point except z_0. Let f be such a function, and suppose that $\lim_{z \to z_0} |f(z)| = \infty$. We say f has a *pole* at z_0 if there exists an integer $n > 0$ such that $\lim_{z \to z_0} (z - z_0)^n f(z)$ exists and is finite. In this case, the function $g(z) = (z - z_0)^n f(z)$ is analytic at z_0. The smallest n for which this limit is finite is called the *order* of the pole. For example, the function $s(z)$ has a pole of order two at z_0.

If f is holomorphic in Ω except for poles, we say f is *meromorphic* in Ω. A meromorphic function can also be expressed by means of a power series, which will include terms of the form $b_n(z - z_0)^{-n}$ for some $n > 0$. Such a series is called a *Laurent series*. Note that the Laurent series for f can be obtained by multiplying the Taylor series for g by $(z - z_0)^{-n}$.

Now let P be a positively oriented simple closed curve, such as the curve $\mathcal{P}_1$ in Fig. B.1. Suppose z_0 lies inside P, and that f is analytic at each point on and inside P except z_0, where it has a pole. Then by integrating the Laurent series for f term-by-term, we find

$$\int_P f(z) \mathrm{d}z = 2\pi i b_1 \,. \tag{B.15}$$

The coefficient b_1 is called the *residue* of f at z_0; we denote it by $\mathrm{Res}(f, z_0)$. This result can be extended to any function that has a finite number of poles inside P: if f has poles $z_1, z_2, \ldots, z_N$ inside P, then

$$\int_P f(z) \mathrm{d}z = 2\pi i \sum_{k=1}^{N} \mathrm{Res}(f, z_k) \,. \tag{B.16}$$

This is called the *residue theorem*. We illustrate it in Fig. B.1.

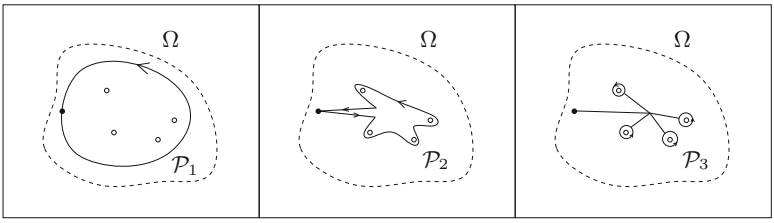


Fig. B.1 The contours $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$ are all homologous. Note that each straight line segment in the third panel is traversed once in each direction, and so their contributions to the value of the integral are all 0. Thus, the integral around $\mathcal{P}_1$ is equal to the sum of the integrals around the circles in the third panel. Each of these integrals is $2\pi i$ multiplied by the corresponding residue

Note the resemblance between the residue theorem and the Poincaré-Hopf index theorem (see Sect. 4.4). Each asserts that the value of a certain integral around a simple closed curve depends only on the behavior of the integrand at certain "exceptional" points: poles of a meromorphic function on the one hand and zeros of a vector field on the other. In particular, the value of such an integral is unchanged if the path of integration is replaced by one that is homologous to it, that is to say, one that can be deformed to it without crossing an exceptional point.

Appendix C
Orthogonal Functions

We often encounter physical quantities that are most naturally represented as linear combinations of orthogonal basis functions. The choice of a particular basis is determined by the symmetry of a given situation. For example, it is convenient to express temperature fluctuations in the cosmic microwave background in terms of spherical harmonics (Sect. 9.2). In this Appendix, we discuss the orthogonal functions most relevant to astrophysics and cosmology, and we demonstrate how these functions arise as solutions to Laplace's equation (cylindrical Bessel functions and cylindrical and spherical harmonics) and the Helmholtz equation (spherical Bessel functions). More details can be found in any standard text on mathematical physics.

C.1 Completeness

Any function of physical interest is *square integrable*, i.e.,

$$\int_{\mathbb{R}^n} |f(\mathbf{x})|^2 \, \mathrm{d}^n x < \infty \,. \tag{C.1}$$

The set of complex-valued square-integrable functions on $\mathbb{R}^n$ is a *Hilbert space.* One feature common to all the sets of functions presented below is that they are *complete*, which is to say that any square-integrable function can be expressed as a linear combination of a possibly infinite number of basis functions.

The most useful bases are those that are *orthonormal*, i.e., a collection $\{u_1, u_2, \ldots\}$ of complex-valued functions for which

$$\int_{\mathbb{R}^n} u_k(\mathbf{x}) u_{k'}^*(\mathbf{x}) \, \mathrm{d}^n x = \delta_{kk'} \,. \tag{C.2}$$

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1

It follows that the functions u_k are *linearly independent*, meaning that none of them can be expressed as a linear combination of the others. An orthonormal set is complete if and only if any square-integrable function $f(\mathbf{x})$ can be written as

$$f(\mathbf{x}) = \sum_{k=1}^{\infty} a_k u_k(\mathbf{x}) \,, \tag{C.3}$$

where for each k,

$$a_k = \int_{\mathbb{R}^n} f(\mathbf{x}) u_k^*(\mathbf{x}) \mathrm{d}^n x \,. \tag{C.4}$$

Note that the coefficients a_k are finite, since

$$\sum_{k=1}^{\infty} |a_k|^2 = \int_{\mathbb{R}^n} |f(\mathbf{x})|^2 \mathrm{d}^n x < \infty \,. \tag{C.5}$$

In this case, we say $\{u_k\}$ is a *basis* of the Hilbert space of square-integrable functions.

A necessary and sufficient condition for completeness is

$$\sum_{k=1}^{\infty} u_k(\mathbf{x}) u_k^*(\mathbf{x}') = \delta^{(n)}(\mathbf{x} - \mathbf{x}') \,. \tag{C.6}$$

To see this, first suppose that $\{u_n\}$ is complete. Then substituting Eq. (C.4) into Eq. (C.3) gives

$$\int_{\mathbb{R}^n} f(\mathbf{x}') \sum_{k=1}^{\infty} u_k(\mathbf{x}) u_k^*(\mathbf{x}') \mathrm{d}^n x' = f(\mathbf{x}) \equiv \int_{\mathbb{R}^n} f(\mathbf{x}') \delta^{(n)}(\mathbf{x} - \mathbf{x}') \mathrm{d}^n x' \,. \tag{C.7}$$

Since this holds for all (square-integrable) functions $f(\mathbf{x})$, Eq. (C.6) immediately follows.

Now assume that Eq. (C.6) is satisfied. Multiplying this relation by any square-integrable function $f(\mathbf{x}')$ and integrating the result with respect to $\mathbf{x}'$ yields

$$\sum_{k=1}^{\infty} u_k(\mathbf{x}) \int_{\mathbb{R}^n} f(\mathbf{x}') u_k^*(\mathbf{x}') \mathrm{d}^n x' = f(\mathbf{x}) \,. \tag{C.8}$$

Thus, $f(\mathbf{x})$ is of the form (C.3) where the coefficients a_k are given by Eq. (C.4). As before, Eq. (C.5) insures that a_k is finite for all k.

We can generalize the completeness relation (C.6) to the case of an uncountable number of basis functions $u(\mathbf{x}; \mathbf{k})$, parametrized by the continuous index $\mathbf{k} \in \mathbb{R}^n$. Equations (C.3), (C.4), (C.5), and (C.6) then become

$$f(\mathbf{x}) = \int_{\mathbb{R}^n} a(\mathbf{k}) u(\mathbf{x}; \mathbf{k}) \mathrm{d}^n k \tag{C.9a}$$

$$a(\mathbf{k}) = \int_{\mathbb{R}^n} f(\mathbf{x}) u^*_{\mathbf{k}}(\mathbf{x}; \mathbf{k}) \mathrm{d}^n x \tag{C.9b}$$

$$\int_{\mathbb{R}^n} f(\mathbf{x}) \mathrm{d}^n x = \int_{\mathbb{R}^n} |a(\mathbf{k})|^2 \, \mathrm{d}^n k \tag{C.9c}$$

$$\int_{\mathbb{R}^n} u(\mathbf{x}; \mathbf{k}) u^*(\mathbf{x}'; \mathbf{k}) \mathrm{d}^n x = \delta^{(n)}(\mathbf{x} - \mathbf{x}') \,, \tag{C.9d}$$

respectively. As an example, consider the complex exponentials $e^{i\mathbf{k}\cdot\mathbf{x}}$ ($\mathbf{k} \in \mathbb{R}^n$). It is well known, though not trivial to prove, that these functions satisfy the orthogonality relation:

$$\int_{\mathbb{R}^n} \mathrm{d}^n x e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}} = (2\pi)^n \delta^{(n)}(\mathbf{k} - \mathbf{k}') \,. \tag{C.10}$$

By relabeling variables and rearranging the result, we see that

$$\int_{\mathbb{R}^n} \frac{\mathrm{d}^n k}{(2\pi)^n} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} = \delta^{(n)}(\mathbf{x} - \mathbf{x}') \,, \tag{C.11}$$

which is just Eq. (C.9d) for the functions $u(\mathbf{x}; \mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{x}}/(2\pi)^{n/2}$. Thus, the complex exponentials form a complete, orthonormal basis for the square-integrable functions. This result provides the foundation for *Fourier analysis*, as discussed in detail in Appendix D.

We have assumed to this point that the domain of interest is $\mathbb{R}^n$, but this condition can be relaxed. The formalism we have presented can be applied both to Euclidean spaces described by curvilinear coordinates (e.g., cylinders) and to non-Euclidean spaces (e.g., spheres). Many of the functions described in the following sections fall into one of these categories.

C.2 Laplace's Equation in Cylindrical Coordinates

We begin by writing Laplace's equation in cylindrical coordinates (R, ϕ, z):

$$\nabla^2 u = \frac{1}{R} \frac{\partial}{\partial R} \left(R \frac{\partial u}{\partial R} \right) + \frac{1}{R^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} = 0 \,. \tag{C.12}$$

C.2.1 Cylindrical Harmonics

We first consider solutions which are independent of z. Laplace's equation then reduces to

$$\nabla^2 u = \frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial u}{\partial R}\right) + \frac{1}{R^2}\frac{\partial^2 u}{\partial \phi^2} = 0\,. \tag{C.13}$$

Substituting a separable solution of the form $u(R,\phi) = f(R)g(\phi)$ and then dividing by fg yield

$$\frac{1}{f}\left(R\frac{\mathrm{d}f}{\mathrm{d}R} + R^2\frac{\mathrm{d}^2 f}{\mathrm{d}R^2}\right) + \frac{1}{g}\frac{\mathrm{d}^2 g}{\mathrm{d}\phi^2} = 0\,. \tag{C.14}$$

Since the first term depends only on R and the second only on ϕ, each must be constant, say, $\pm\mu^2$, respectively:

$$\frac{1}{f}\left(R\frac{\mathrm{d}f}{\mathrm{d}R} + R^2\frac{\mathrm{d}^2 f}{\mathrm{d}R^2}\right) = \mu^2 \tag{C.15a}$$

$$\frac{1}{g}\frac{\mathrm{d}^2 g}{\mathrm{d}\phi^2} = -\mu^2. \tag{C.15b}$$

We first solve the azimuthal equation, obtaining

$$g(\phi) = A\cos(\mu\phi) + B\sin(\mu\phi) \equiv C\cos[\mu(\phi - \phi_*)]\,, \tag{C.16}$$

where A, B, C, and ϕ_* are constants. In order that $u(R,\phi)$ be single valued, we require $g(\phi + 2\pi) = g(\phi)$. This means that μ must be an integer m. The radial equation can then be written as

$$Rf'(R) + R^2 f''(R) = m^2 f(R). \tag{C.17}$$

Substituting a solution of the form $f(R) = DR^\delta$, we find $\delta = \pm m$. For $m = 0$, the two linearly independent solutions are $f(R) = C_0$ (a constant) and $f(R) = B_0 \ln R$. Thus, the general solution of Laplace's equation in polar coordinates is

$$u(R,\phi) = B_0 \ln R + \sum_{m=-\infty}^{\infty} C_m R^m \cos[m(\phi - \phi_m)]\,. \tag{C.18}$$

C.2.2 Cylindrical Bessel Functions

Returning to three dimensions, we again use the technique of separation of variables and assume a solution to Eq. (C.12) of the form $u(R,\phi,z) \equiv f(R)g(\phi)h(z)$. Making this substitution and dividing by fgh yield

$$\frac{1}{fR}\frac{\mathrm{d}}{\mathrm{d}R}\left(R\frac{\mathrm{d}f}{\mathrm{d}R}\right)+\frac{1}{gR^2}\frac{\mathrm{d}^2g}{\mathrm{d}\phi^2}+\frac{1}{h}\frac{\mathrm{d}^2h}{\mathrm{d}z^2}=0\,. \tag{C.19}$$

Since the third term is the only term which depends on z and is independent of R and ϕ, it must be constant. Thus, $h''=k^2h$, and

$$\frac{1}{fR}\frac{\mathrm{d}}{\mathrm{d}R}\left(R\frac{\mathrm{d}f}{\mathrm{d}R}\right)+\frac{1}{gR^2}\frac{\mathrm{d}^2g}{\mathrm{d}\phi^2}+k^2=0\,. \tag{C.20}$$

Multiplying by R^2 gives

$$\frac{R}{f}\frac{\mathrm{d}}{\mathrm{d}R}\left(R\frac{\mathrm{d}f}{\mathrm{d}R}\right)+k^2R^2+\frac{1}{g}\frac{\mathrm{d}^2g}{\mathrm{d}\phi^2}=0. \tag{C.21}$$

Since the third term depends only on ϕ, $g''=\mu^2g$. This implies

$$\boxed{R\frac{\mathrm{d}}{\mathrm{d}R}\left(R\frac{\mathrm{d}f}{\mathrm{d}R}\right)+(k^2R^2-m^2)f=0}\,, \tag{C.22}$$

which is known as Bessel's equation. In the solution to Laplace's equation, μ is restricted to integer values, as in the preceding subsection. However, in other applications this is not the case, and we consider the solution for general complex μ.

In terms of $x=kR$ and $F(x)=f(R)$, Bessel's equation takes the form

$$x\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}F}{\mathrm{d}x}\right)+(x^2-\mu^2)F=0 \tag{C.23}$$

and has the power series solution

$$J_\mu(x)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\Gamma(k+\mu+1)}\left(\frac{x}{2}\right)^{2k+\mu}\,. \tag{C.24}$$

The functions $J_\mu(x)$ and $J_{-\mu}(x)$ are linearly independent, except in the limit $\mu\to m$, where m is an integer. This follows from the identity $J_m(x)=(-1)^mJ_{-m}(x)$,

which can be proved from Eq. (C.24), keeping in mind that the gamma function has simple poles at each of the nonpositive integers. It is customary to work with the two solutions $J_\mu(x)$ and

$$Y_\mu(x) = \frac{J_\mu(x)\cos(\mu\pi) - J_{-\mu}(x)}{\sin(\mu\pi)}\,, \tag{C.25}$$

known as Bessel functions of the first and second kind (or as Bessel and Neumann functions, respectively), which remain linearly independent, even in the limit of integer μ.

The Bessel functions have the useful integral representation

$$J_\mu(z) = \frac{(\frac{z}{2})^\mu}{\sqrt{\pi}\,\Gamma(\mu+\frac{1}{2})}\int_0^\pi e^{iz\cos\phi}\sin^{2\mu}\phi\,\mathrm{d}\phi \tag{C.26}$$

for complex z. Expanding the exponential in the integrand in powers of $iz\cos\phi$ and then integrating over ϕ lead to the series representation (C.24).

We also mention two important orthogonality properties of the Bessel functions. It can be shown that

$$\int_0^a J_\mu(x_{\mu n}R/a)J_\mu(x_{\mu n'}R/a)R\,\mathrm{d}R = \frac{1}{2}a^2 J_{\mu+1}^2(x_{\mu n})\delta_{n,n'}\,, \tag{C.27}$$

where $x_{\mu n}$ with $n = 1, 2, 3, \ldots$ is the nth zero of $J_\mu(x)$. In the limit $a \to \infty$, $n \to \infty$, $n' \to \infty$ with fixed $k \equiv x_{\mu n}/a$ and $k' \equiv x_{\mu n'}/a$, the orthogonality property (C.27) takes the form

$$\int_0^\infty J_\mu(kR)J_\mu(k'R)R\,\mathrm{d}R = k^{-1}\delta(k-k')\,. \tag{C.28}$$

C.3 Helmholtz's Equation in Spherical Coordinates

In addition to the cylindrical harmonics and Bessel functions, we will need their spherical counterparts. We introduce them in the context of Helmholtz's equation for the propagation of waves. The pressure amplitude $u(x,t)$ of a sound wave propagating along the x-axis satisfies the one-dimensional wave equation,

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}\,, \tag{C.29}$$

where c is the speed of sound. In three dimensions, the wave equation is

$$\nabla^2 u = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}\,, \tag{C.30}$$

where $u \equiv u(\mathbf{x}, t)$. For a wave that is periodic in time, we can write $u(\mathbf{x}, t) = e^{i\omega t} u(\mathbf{x})$. On assuming this time dependence and substituting $\omega \equiv ck$, the wave equation reduces to Helmholtz's equation:

$$\nabla^2 u = -k^2 u \,. \tag{C.31}$$

In terms of spherical coordinates (r, θ, ϕ), Helmholtz's equation takes the form

$$\begin{aligned} \nabla^2 u &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 u}{\partial \phi^2} \\ &= -k^2 u \,. \end{aligned} \tag{C.32}$$

Seeking a separable solution, we substitute $u(r, \theta, \phi) \equiv f(r)g(\theta)h(\phi)$ and multiply by $r^2 \sin^2\theta/(fgh)$, which leads to

$$\frac{1}{u}\nabla^2 u = \frac{\sin^2\theta}{f}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2 \frac{\mathrm{d}f}{\mathrm{d}r}\right) + \frac{\sin\theta}{g}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta \frac{\mathrm{d}g}{\mathrm{d}\theta}\right) + \frac{1}{h}\frac{\mathrm{d}^2 h}{\mathrm{d}\phi^2} = -k^2 r^2 \sin^2\theta \,. \tag{C.33}$$

The term involving h is the only one that depends on ϕ, so it must be a constant. Thus,

$$h(\phi) \propto e^{\pm i m \phi} \,, \tag{C.34}$$

where the requirement of periodicity restricts m to integer values. Substituting this into Helmholtz's equation and dividing by $\sin^2\theta$ give

$$\frac{1}{f}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2 \frac{\mathrm{d}f}{\mathrm{d}r}\right) + k^2 r^2 + \frac{1}{g \sin\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta \frac{\mathrm{d}g}{\mathrm{d}\theta}\right) - \frac{m^2}{\sin^2\theta} = 0 \,. \tag{C.35}$$

The first two terms of Eq. (C.35) depend only on r, while the final two depend only on θ. Consequently,

$$\frac{1}{f}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2 \frac{\mathrm{d}f}{\mathrm{d}r}\right) + k^2 r^2 = \ell(\ell+1) \tag{C.36}$$

$$\frac{1}{g \sin\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta \frac{\mathrm{d}g}{\mathrm{d}\theta}\right) - \frac{m^2}{\sin^2\theta} = -\ell(\ell+1), \tag{C.37}$$

for some real number ℓ. (Why we call this constant $\ell(\ell+1)$ will become apparent in the next subsection.)

C.3.1 Spherical Harmonics

In terms of $x \equiv \cos\theta$ and $G(x) \equiv g(\theta)$, Eq. (C.37) takes the form

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\frac{\mathrm{d}G}{\mathrm{d}x}\right] - \left(\ell(\ell+1) - \frac{m^2}{1-x^2}\right) G = 0\,. \tag{C.38}$$

We first solve this differential equation in the special case $m = 0$, known as Legendre's equation. Substituting a power series solution $G(x) = \sum_{k=0}^{\infty} a_k x^{\alpha+k}$, one can show that α must be 0 or 1 and that the series diverges unless it terminates after a finite number of terms, which restricts ℓ to 0, 1, 2, 3,... The corresponding polynomial solutions $P_\ell(x)$ are normalized so that $P_\ell(1) = 1$, known as Legendre polynomials, are given by Rodrigues' formula[5]:

$$P_\ell(x) = \frac{1}{2^\ell \ell!}\frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell}[(x^2-1)^\ell]\,. \tag{C.39}$$

It follows that $P_\ell(x)$ is an even (odd) function of x for ℓ even (odd). The first four Legendre polynomials are

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2-1) \\ P_3(x) &= \frac{1}{2}(5x^3-3x)\,. \end{aligned} \tag{C.40}$$

For $m \geq 0$, the solution to Eq. (C.38) is the **associated Legendre polynomial**

$$P_{\ell m}(x) = (-1)^m (1-x^2)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}x^m}P_\ell(x)\,, \tag{C.41}$$

which reduces to $P_\ell(x)$ for $m = 0$ and vanishes for all integers $m > \ell$. The associated Legendre polynomials are extended to negative m by the relation

$$P_{\ell,-m} = (-1)^m \frac{(\ell-m)!}{(\ell+m)!}\,P_{\ell m}\,. \tag{C.42}$$

[5]For a given ℓ, Legendre's equation has two linearly independent solutions $P_\ell(x)$ and $Q_\ell(x)$. The $Q_\ell(x)$ diverge logarithmically at $x = \pm 1$ and are not needed in many physical applications.

Thus, the allowed range of m is $-\ell, -\ell+1, \ldots, \ell$. Finally we note the orthogonality relation

$$\int_{-1}^{1} P_{\ell m}(x) P_{\ell' m}(x) \mathrm{d}x = \frac{2(\ell+m)!}{(2\ell+1)(\ell-m)!} \delta_{\ell\ell'} \,. \tag{C.43}$$

It is convenient to combine the angular functions $g(\theta) \propto P_{\ell m}(\cos\theta)$ and $h(\phi) \propto e^{im\phi}$ in the spherical harmonic $Y_{\ell m}(\theta, \phi)$ defined by

$$\boxed{Y_{\ell m}(\theta, \phi) = \sqrt{\frac{(2\ell+1)}{4\pi} \frac{(\ell-m)!}{(\ell+m)!}} P_{\ell m}(\cos\theta) e^{im\phi}} \tag{C.44}$$

for all integers $\ell \geq 0$ and $-\ell \leq m \leq \ell$. The spherical harmonics have the property

$$Y_{\ell,-m}(\theta, \phi) = (-1)^m Y^*{}_{\ell m}(\theta, \phi) \,. \tag{C.45}$$

Note that $Y_{\ell m}(\theta, \phi)$ and $Y_{\ell -m}(\theta, \phi)$ are linearly independent, whereas $P_{\ell m}(\cos\theta)$ and $P_{\ell -m}(\cos\theta)$ are not, as is clear from Eq. (C.42). The spherical harmonics form a complete set of angular functions and have the orthonormality property

$$\int_0^{\pi} \mathrm{d}\theta \sin\theta \int_0^{2\pi} \mathrm{d}\phi \, Y^*{}_{\ell m}(\theta, \phi) Y_{\ell' m'}(\theta, \phi) = \delta_{\ell\ell'} \delta_{mm'} \,, \tag{C.46}$$

which follows from Eq. (C.43) and the identity $\int_0^{2\pi} e^{i(m-m')\phi} \mathrm{d}\phi = 2\pi \delta_{mm'}$. They also satisfy the useful addition theorem

$$P_\ell(\cos\vartheta) = \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y^*{}_{\ell m}(\theta, \phi) Y_{\ell m}(\theta', \phi') \,, \tag{C.47}$$

where ϑ is the angle between (θ, ϕ) and (θ', ϕ'), which satisfies $\cos\vartheta = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi')$.

The first few spherical harmonics are given by

$$\begin{aligned}
Y_{00} &= \frac{1}{\sqrt{4\pi}} \\
Y_{1,\pm1} &= \mp\sqrt{\frac{3}{8\pi}} \sin\theta \, e^{\pm i\phi} \\
Y_{10} &= \sqrt{\frac{3}{4\pi}} \cos\theta \\
Y_{2,\pm2} &= \sqrt{\frac{15}{32\pi}} \sin^2\theta \, e^{\pm 2i\phi}
\end{aligned}$$

$$Y_{2,\pm1} = \mp\sqrt{\frac{15}{8\pi}}\,\sin\theta\,\cos\theta\,e^{\pm i\phi}$$

$$Y_{20} = \sqrt{\frac{5}{4\pi}}\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right). \tag{C.48}$$

C.3.2 Spherical Bessel Functions

On substituting $f(r) \equiv F(r)/\sqrt{kr}$, Eq. (C.36) becomes

$$r\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}F}{\mathrm{d}r}\right) + \left[k^2r^2 - \left(\ell + \frac{1}{2}\right)^2\right]F = 0\,, \tag{C.49}$$

which is Bessel's Equation (C.22) with $R = r$ and the integer m replaced by the half-odd-integer $\ell + \frac{1}{2}$. According to Eqs. (C.22), (C.24), and (C.25), the general solution of Eq. (C.36) is given by

$$f(r) = a\,j_\ell(kr) + b\,y_\ell(kr)\,, \tag{C.50}$$

where a and b are arbitrary constants and the functions

$$j_\ell(x) \equiv \sqrt{\frac{\pi}{2x}}\,J_{\ell+1/2}(x)\,, \quad y_\ell(x) \equiv \sqrt{\frac{\pi}{2x}}\,Y_{\ell+1/2}(x) \tag{C.51}$$

are called spherical Bessel functions of order ℓ.

According to Eqs. (C.24), (C.25), and (C.51), $y_\ell(x)$ diverges as $x^{-(\ell+1)}$ for $x \to 0$, whereas $j_\ell(x) \sim x^\ell$ remains finite. Both types of spherical Bessel functions are expressible in terms of $\sin x$ and $\cos x$ multiplied by polynomials in x^{-1}. The spherical Bessel functions for l =1 and 2 have the explicit form

$$j_0(x) = \frac{\sin x}{x} \tag{C.52}$$

$$j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x} \tag{C.53}$$

$$y_0(x) = -\frac{\cos x}{x} \tag{C.54}$$

$$y_1(x) = -\frac{\cos x}{x^2} - \frac{\sin x}{x}\,. \tag{C.55}$$

In terms of spherical Bessel functions, the orthogonality properties (C.27) and (C.28) take the form

$$\int_0^a j_\ell(\xi_{\ell n} r/a)\, j_\ell(\xi_{\ell n'} r/a) r^2 \, \mathrm{d}r = \frac{1}{2} a^3 j_{\ell+1}^2(\xi_{\ell n}) \delta_{n,n'} \,, \tag{C.56}$$

where $\xi_{\ell n}$ with $n = 1, 2, 3, \ldots$ is the nth zero of $j_\ell(x)$ and

$$\int_0^\infty j_\ell(kr)\, j_\ell(k' r) r^2 \, \mathrm{d}r = \frac{\pi}{2k^2} \delta(k - k') \,. \tag{C.57}$$

Appendix D
Fourier Analysis

In this Appendix, we use the complex exponentials e^{ikx} as a basis for representing periodic functions as Fourier series and aperiodic functions as Fourier integrals.

D.1 Fourier Series

The exponential function e^{ikx} is periodic in x with period L if $e^{ikx} = e^{ik(x+L)}$, which implies $e^{ikL} = 1$, or $k = 2\pi n/L$, where n is an integer. The functions $e^{2\pi inx/L}$ form a very useful set of basis functions with the orthogonality property

$$\int_{-L/2}^{L/2} \mathrm{d}x\, e^{2\pi i(m-n)x/L} = L\delta_{mn}\,. \tag{D.1}$$

An arbitrary function $f(x)$ can be expanded in terms of the basis functions in a *Fourier series*

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi inx/L}\,. \tag{D.2}$$

Multiplying both sides of this equation by $e^{-2\pi imx/L}$ and integrating over x with the help of Eq. (D.1), we find that the coefficients c_m are determined by

$$c_n = \frac{1}{L}\int_{-L/2}^{L/2} \mathrm{d}x\, f(x)e^{-2\pi inx/L}\,. \tag{D.3}$$

The coefficients c_n are, in general, complex and, if $f(x)$ is real, satisfy $c_{-n} = c_n^*$.

Since $\exp(2\pi inx/L) = \cos(2\pi nx/L) + i\sin(2\pi nx/L)$, the Fourier series (D.2) can be rewritten as

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{2\pi n x}{L}\right) + b_n \sin\left(\frac{2\pi n x}{L}\right) \right], \tag{D.4}$$

where $a_n = c_n + c_{-n}$ and $b_n = i\,(c_n - c_{-n})$ for $n \neq 0$ and $a_0 = c_0$. Clearly, if $f(x)$ is an even (odd) function of x, the b_n (the a_n) all vanish.

D.2 Fourier Integrals

Any function $f(x)$ for which $\int_{-\infty}^{\infty} |f(x)|^2\,\mathrm{d}x$ converges can be expanded on the interval $-\infty < x < \infty$ as a *Fourier integral*. This representation follows from Eqs. (D.2) and (D.3) in the limit $L \to \infty$. Defining $k = 2\pi n/L$, we note that the spacing $\Delta k = 2\pi/L$ between two adjacent k values goes to zero for $L \to \infty$, and we use this to convert the sum over n to an integral over k. Using the definition of k, defining $\tilde{f}(k) = L\,c_n$, and inserting the factor $L\Delta k/(2\pi) = 1$, we rewrite Eqs. (D.2) and (D.3) as

$$f(x) = \lim_{L\to\infty} \sum_{\{k\}} \frac{L\Delta k}{2\pi}\, \frac{\tilde{f}(k)}{L}\, e^{ikx} \tag{D.5}$$

$$\tilde{f}(k) = \lim_{L\to\infty} \int_{-L/2}^{L/2} \mathrm{d}x\; f(x) e^{-ikx}\,. \tag{D.6}$$

This yields the desired Fourier integral representation

$$f(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi}\, \tilde{f}(k) e^{ikx} \tag{D.7}$$

$$\tilde{f}(k) = \int_{-\infty}^{\infty} \mathrm{d}x\; f(x) e^{-ikx}\,, \tag{D.8}$$

where $\tilde{f}(k)$ is called the *Fourier transform* of $f(x)$. In N dimensions,

$$f(\mathbf{x}) = \int_{\mathbb{R}^N} \frac{\mathrm{d}^N k}{(2\pi)^N}\, \tilde{f}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} \tag{D.9}$$

$$\tilde{f}(\mathbf{k}) = \int_{\mathbb{R}^N} \mathrm{d}^N x\; f(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}\,. \tag{D.10}$$

Some differential equations are easily solved by the use of Fourier transforms, and for this the Fourier transform of derivatives of a function is needed. Differentiating Eq. (D.7) n times, we see that the Fourier transform of $\mathrm{d}^n f(x)/\mathrm{d}x^n$ is $(ik)^n \tilde{f}(k)$. Similarly, the Fourier transforms of $\boldsymbol{\nabla} f(\mathbf{x})$ and of $\boldsymbol{\nabla}^2 f(\mathbf{x})$ are $i\mathbf{k}\tilde{f}(\mathbf{k})$ and $-\mathbf{k}\cdot\mathbf{k}\,\tilde{f}(\mathbf{k})$, respectively.

D.3 Convolution Theorem

Convolution integrals are frequently encountered in applications. For example, they can be used to solve linear differential equations. Let L_x be a linear differential operator, and consider the equation

$$L_x[y(x)] = f(x)\,. \tag{D.11}$$

To solve this equation, we first look for a so-called *Green's function* g such that $L_x[g(x-x')] = \delta(x-x')$. Then

$$L_x[f(x')g(x-x')] = f(x')L_x[g(x-x')] = f(x')\delta(x-x')\,. \tag{D.12}$$

Integrating both sides, we have

$$\begin{aligned}\int L_x[f(x')g(x-x')]\mathrm{d}x' &= L_x\left[\int f(x')g(x-x')\mathrm{d}x'\right] \\ &= \int f(x')\delta(x-x')\mathrm{d}x' = f(x). \end{aligned}\tag{D.13}$$

Hence,

$$y_p(x) = \int f(x')g(x-x')\mathrm{d}x' \tag{D.14}$$

is a *particular solution* to Eq. (D.11). It is a *superposition* of Green's functions. Recall that any solution of Eq. (D.11) has the form $y(x) = y_p(x) + h(x)$, where h satisfies the homogeneous equation $L_x[h(x)] = 0$. The Green's function method can also be used to solve higher-dimensional ordinary and partial differential equations.

An integral of the form (D.14) is said to be a *convolution*. For functions $f, g : \mathbb{R}^N \to \mathbb{R}$, their convolution is given by[6]

$$\mathscr{C}_{pq}(\mathbf{x}) \equiv \int_{\mathbb{R}^N} \mathrm{d}^N x'\; p(\mathbf{x}')q(\mathbf{x}-\mathbf{x}') = \int_{\mathbb{R}^N} \mathrm{d}^N x'\; p(\mathbf{x}-\mathbf{x}')q(\mathbf{x}') \equiv \mathscr{C}_{qp}(\mathbf{x})\,. \tag{D.15}$$

According to Eq. (D.10), the Fourier transform of $\mathscr{C}_{pq}(\mathbf{x})$ is

$$\tilde{\mathscr{C}}_{pq}(\mathbf{k}) = \int_{\mathbb{R}^N} \mathrm{d}^N x\, e^{-i\mathbf{k}\cdot\mathbf{x}} \int_{\mathbb{R}^N} \mathrm{d}^N x'\; p(\mathbf{x}')q(\mathbf{x}-\mathbf{x}')$$

[6] The convolution of f and g is often written as $f * g$, but this notation can be confusing if an argument of the resulting function is explicitly included.

$$= \int_{\mathbb{R}^N} \mathrm{d}^N x' e^{-i\mathbf{k}\cdot\mathbf{x}'} p(\mathbf{x}') \int_{\mathbb{R}^N} \mathrm{d}^N x \, e^{-i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} q(\mathbf{x}-\mathbf{x}')$$

$$= \int_{\mathbb{R}^N} \mathrm{d}^N x' \, e^{-i\mathbf{k}\cdot\mathbf{x}'} p(\mathbf{x}') \int_{\mathbb{R}^N} \mathrm{d}^N y \, e^{-i\mathbf{k}\cdot\mathbf{y}} q(\mathbf{y}) \,, \tag{D.16}$$

where in the second line, the order of the x and x' integrations has been reversed and in the third line, the integration variable $\mathbf{x}$ has been replaced by $\mathbf{y} = \mathbf{x} - \mathbf{x}'$. Making use of Eq. (D.10), we express the third line in terms of Fourier transforms in the simple form

$$\tilde{\mathscr{C}}_{pq}(\mathbf{k}) = \tilde{p}(\mathbf{k})\tilde{q}(\mathbf{k}) \,, \tag{D.17}$$

where $\tilde{p}$ and $\tilde{q}$ are the Fourier transforms of p and q. This is the *convolution theorem.*

Appendix E
Computational Techniques

While the principles of gravitational lensing can be gleaned from analytic examples, many applications require computational approaches. In this Appendix we describe computational methods that can be used to solve the lens equation and find critical curves and caustics. The key concepts have been discussed previously in connection with the original version of Keeton's `gravlens` software (see Keeton 2001a,b, 2010, 2016), but they have been updated for a new open-source, Python-based version of the software.

E.1 Tiling the Image and Source Planes

Many analyses involve searching the image plane (e.g., for images and/or critical curves), so a first step is to make a grid and cover the plane with tiles. It would be ideal to have an infinite number of infinitesimal tiles, but the more tiles we use, the more computer time we need. To balance computational effort and numerical precision, we can use an adaptive algorithm that makes tiles small only where it is important to have good resolution—which, for lensing purposes, is near critical curves. Figure E.1 shows an example that starts with a coarse regular grid (depicted by the red points) and uses adaptive subgridding to improve the resolution near the critical curve. Specifically, if the magnification changes sign across a grid cell, then the critical curve must cross the cell, so that cell should be subdivided.[7] The subdivision can be arbitrary, but the example in the figure uses 2×2 subgrids and goes down three layers beyond the initial grid.

The points on the grid can be connected into triangular tiles using a computational geometry technique known as Delaunay triangulation (Delaunay 1934). The

[7] If we only check the corners of a cell, we may miss cells in which the critical curve crosses a small portion of one edge, but such situations should be relatively rare.

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1

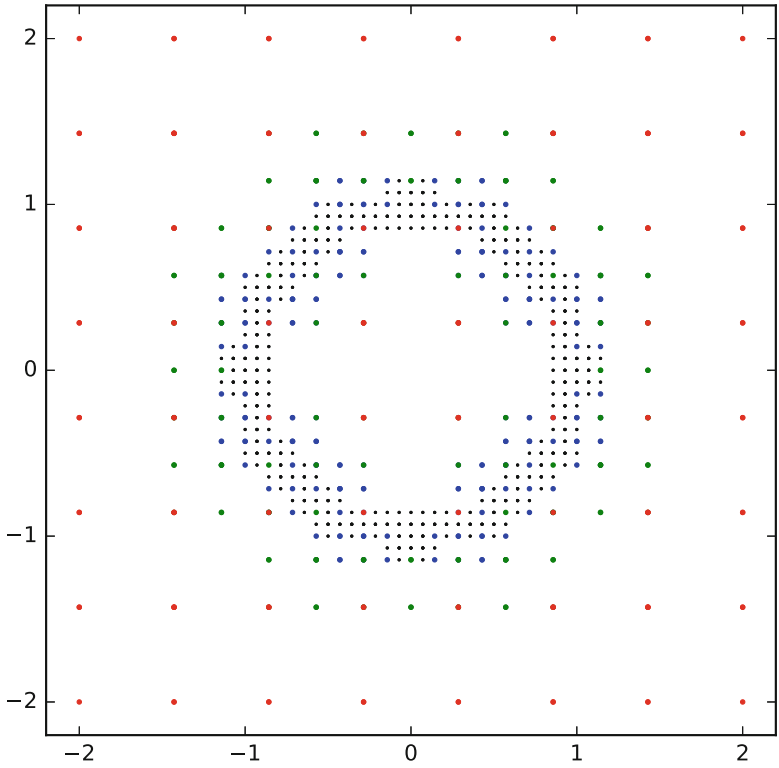


Fig. E.1 Illustration of adaptive gridding for a point mass lens with an Einstein radius of unity. The starting grid is shown by the red points. If the magnification changes sign across a grid cell, then the critical curve must cross the cell, so that cell is subdivided to obtain the green points. Further subdivisions yield the blue and black points

left panel of Fig. E.2 shows a triangulation of the example grid. To obtain good resolution near the centers of lenses, it may be valuable to use a polar grid as shown in the middle panel of Fig. E.2. Delaunay triangulation can easily handle a combination of Cartesian and polar grids, as shown in the right panel of Fig. E.2.

The idea can be generalized to composite models that have multiple mass components. Figure E.3 shows an example with two point masses, using polar grids centered on the two masses along with a Cartesian grid to achieve good coverage of the full image plane.

The tiling of the image plane leads naturally to a tiling of the source plane. Each grid point in the image plane is mapped to the source plane, and the connections are mapped as well: if $\boldsymbol{\theta}_i$ connects to $\boldsymbol{\theta}_j$, then $\boldsymbol{\beta}_i$ connects to $\boldsymbol{\beta}_j$.[8] In a strong lensing

[8]The Delaunay triangulation should *not* be computed directly in the source plane, because that would lose information about the lens mapping.

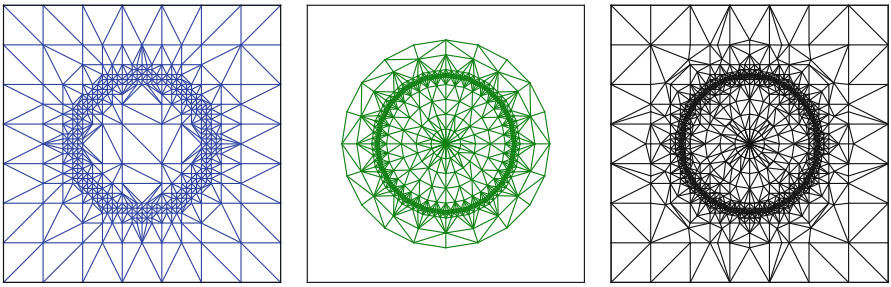


Fig. E.2 The left panel shows a Delaunay triangulation of the adaptive Cartesian grid from Fig. E.1 The middle panel shows an adaptive polar grid. The right panel shows the result when the Cartesian and polar grids are combined before performing the Delaunay triangulation

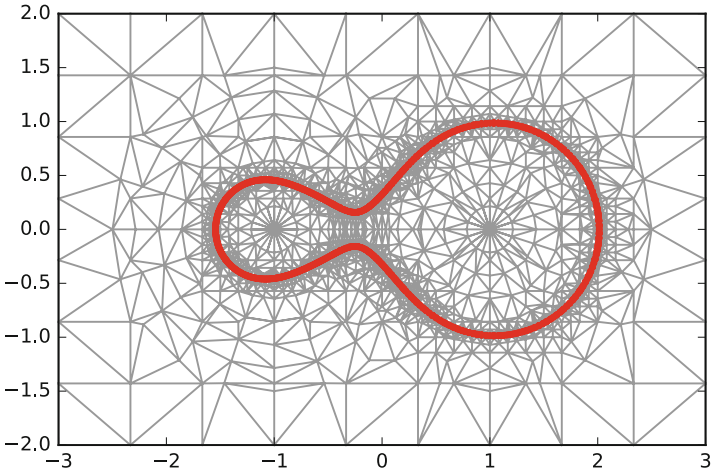


Fig. E.3 An example with two point masses. A Cartesian grid is combined with polar grids centered on each point mass. The lensing critical curves are shown in red (see Sect. E.2)

scenario, the grid folds over into itself as depicted in Figs. E.4 and E.5, such that any multiply-imaged region of the source plane is covered by multiple tiles.

E.2 Finding Critical Curves and Caustics

The tiling of the image plane can serve as a starting point to find the lensing critical curves. For each tile edge, we can check the magnification at the two endpoints. If they have different signs, then the critical curve must cross that segment, and we can use a numerical root finder to find the crossing point. By checking all tile edges, we can find the collection of critical points and then map them to the source plane to obtain points on the caustic(s). Critical curves and caustics found in this way are shown starting in Fig. E.3.

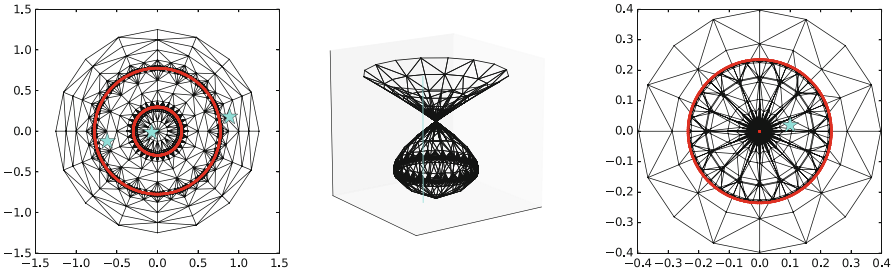


Fig. E.4 The left panel shows the grid (black) and critical curves (red) in the image plane for a nonsingular isothermal sphere lens. The middle panel depicts the lens mapping. Each part of the image plane is effectively pulled toward the origin (recall $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})$), such that the central region folds over on itself. The folded grid gets projected down onto the source plane, yielding the overlapping grid shown in the right panel (with the caustics in red). The cyan points and line illustrate a solution of the lens equation (see Sect. E.3)

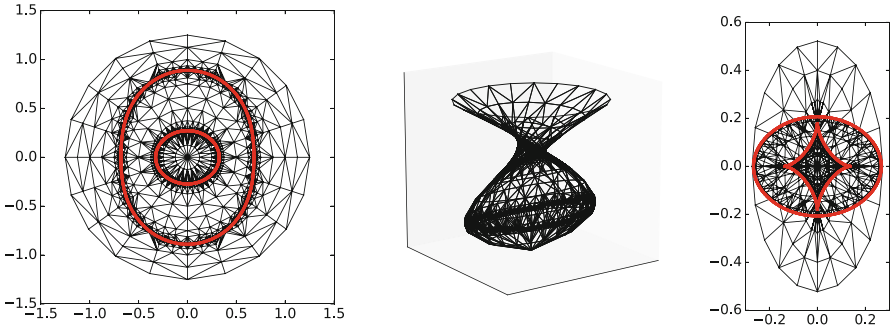


Fig. E.5 Similar to Fig. E.4 but the lens model includes shear so the folding is more complicated

E.3 Solving the Lens Equation

Last but not least, the tiling provides a way to solve the lens equation, as depicted in Fig. E.4. We pick a source position, indicated by the cyan star in the right panel. We find the tiles that overlap that source position, as shown by the cyan line in the middle panel. The positions of those tiles in the image plane provide estimates of the positions of the images, which can be refined with a numerical root finder. The resulting image positions are shown by the cyan stars in the left panel of the figure.

References

Delaunay, B. (1934). Sur la sphère vide. A la mémoire de Georges Voronoï. *Bulletin de l'Académie des Sciences de l'URSS, VII*(6), 793.

Keeton, C. R. (2001a). *A Catalog of Mass Models for Gravitational Lensing*, astro-ph/0102341.

Keeton, C. R. (2001b). *Computational Methods for Gravitational Lensing*, astro-ph/0102340.
Keeton, C. R. (2010). On modeling galaxy-scale strong lens systems. *General Relativity and Gravitation, 42*, 2151.
Keeton, C. R. (2016). Methods for strong lens modelling. In E. Mediavilla, J. A. Muñoz, F. Garzón, & T. J. Mahoney (Eds.), *Astrophysical applications of gravitational lensing* (p. 213). Cambridge: Cambridge University Press.

Index

A. B. Congdon, C. R. Keeton, *Principles of Gravitational Lensing*,
Springer Praxis Books, https://doi.org/10.1007/978-3-030-02122-1