eaa.iop.org
DOI: 10.1888/0333750888/2631

Gravitational Lensing

P Saha

From

Encyclopedia of Astronomy & Astrophysics

P. Murdin

ISBN: 0333750888

Institute of Physics Publishing
Bristol and Philadelphia

Gravitational Lensing

Gravity bends light rays in a way analogous to, but quantitatively different from, the way it bends trajectories of passing particles. If light from some bright object passes close enough to some foreground mass, that object's image will be altered. The effect is more like a piece of bathroom glass in the sky than a precision-ground and well-focused lens, but the terms 'gravitational lensing' or simply 'lensing' have stuck. The observable effect varies, depending on the mass of the lens and the relative positions of lens and background source. 'Weak lensing' produces a shape change in the image, typically a stretching along one direction; for very weak lensing such stretching is imperceptible in images of individual sources but statistically measurable from images of many sources close together on the sky. 'Strong lensing' produces multiple images of the same source, one or more being brighter than the unlensed source would have been, and typically all having very distorted shapes.

The bending of light by the Sun can be measured directly, but apart from this all known examples of lensing involve very distant objects. (A peculiar feature of the physics of lensing is that lensing becomes easier to find as one observes more distant objects.) Lensing of stars in the Milky Way and its environs by foreground stellar-sized masses is now routinely observed; this requires very large surveys because of order 1 in 10^6 stars is being strongly lensed. Lensing by foreground galaxies of quasars is proportionately much more common—of order 1 in 10^3 quasars is strongly lensed. GRAVITATIONAL LENSING BY CLUSTERS OF GALAXIES is ubiquitous—a long-exposure image of any rich cluster of galaxies at cosmological distance will show obvious weak lensing and possibly also strong lensing.

What makes lensing particularly attractive is that it depends only on gravity and geometrical things such as source and lens positions; the effects of lensing do not come mixed up with other (possibly poorly understood) physical effects. Thus astrophysical inferences made from lensing will hopefully be particularly robust. Lensing is astrophysically most important as a probe of dark matter (see DARK MATTER: ITS NATURE). However, it has other, sometimes surprising, uses too.

The main part of this article is based on physical arguments. Mathematical material is confined to sections indicated between ◊ and ⧫, and these may be skipped.

Basic physics of lensing

Photons are affected by a gravitational field, but not in the same way as massive particles are. For the details one needs general relativity, and because of this lensing was considered a very difficult subject for a long time (see GENERAL RELATIVITY AND GRAVITATION). However, in astrophysical situations some useful approximations apply, and these make the physics very much simpler than in full general relativity. If anything, lensing is easier than the astrophysical applications of Newtonian dynamics.

There are several ways of expressing the effect of gravity on light, in the context of lensing. Perhaps the most intuitive is through the wavefront. A wavefront is the locus of points with a given light travel time from the source. When a wavefront crosses an observer, they see an image of the source in the direction normal to the wavefront. Normally a wavefront is an expanding sphere, and light travels in straight lines. The effect of a gravitational field is to delay the part of the wavefront that passes through it. This naturally changes the shape of the wavefront and hence the direction normal to it, i.e. the position of the image, which is equivalent to saying that the light ray is bent. Figure 1 illustrates this. Note that if the gravitational delay is large enough, the wavefront can double in on itself, producing multiple images (strong lensing). A good way to gain some intuition about the wavefront is to take a plastic transparency with a blank piece of paper behind it and look at the reflections of a lightbulb. The shape of the plastic produces delays in the lightbulb's wavefront analogous to the gravitational time delays. Notice how changing the shape of the plastic causes images to merge and split, and the appearance of grotesquely stretched images at these transitions. Images of galaxies lensed by clusters show just such effects.

The wavefront formulation is not in practice the most useful because it works in terms of a single source and many observers, whereas in astrophysics one is usually interested in many sources and a single observer. So lensing work generally uses two other formulations and some standard approximations.

The standard approximations are the following. First, the gravitational fields are weak enough that massive particles simply follow Newtonian dynamics. (Weak gravitational fields should not be confused with weak lensing—strong lensing can still happen.) Second, the extent of the lens masses along the line of sight is negligible compared with the distances between source and lens and observer. Then a lens behaves as if its mass had been squashed along the line of sight into a sheet of mass. The mass distribution transverse to the line of sight is still important, however. Third, the lens and source are nearly aligned along the line of sight, in the sense that the angular separation (say θ) on the sky between the lens and source is such that $\sin\theta \simeq \theta$. Fourth, diffraction effects are negligible, because even for radio waves the wavelengths are too small.

With these approximations, the effect of gravity on photons can be described by adopting the following result from general relativity. A point mass M at perpendicular distance R from a light ray bends the light towards it by an angular amount

$$\frac{4GM}{c^2R}$$

where G is the gravitational constant and c the speed of light. For the weak field approximation to apply R must be much larger than $2GM/c^2$ (the Schwarzschild radius), but this condition is easily satisfied because Schwarzschild radii are at most a few km for stars, <1 pc

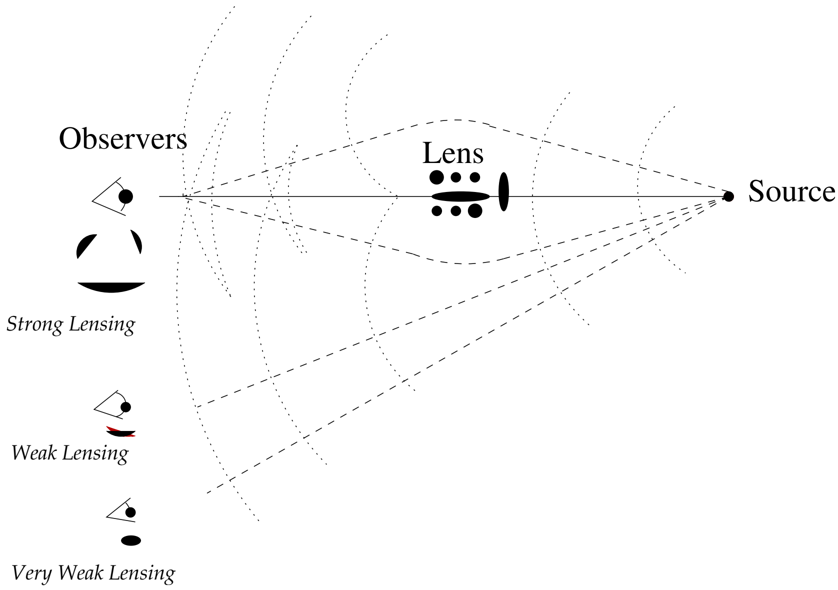

Figure 1. Schematic illustration of the wavefront. Observers see images of the source in the direction normal to the wavefront where the wavefront crosses them. If delayed enough by the lens, part of the wavefront can double in on itself, producing multiple images for some observers (strong lensing); other observers see single images (weak lensing). Images usually experience size and shape changes.

for a galaxy and $\lesssim 10^3$ pc for a cluster of galaxies. (The bending angle of $4GM/c^2R$ was the first prediction of general relativity; Newtonian dynamics would predict half that value. Measurement of the bending angle for the Sun demonstrated that the dynamics of photons is not Newtonian, and was the first test of general relativity.)

Alternatively, one can proceed from another result from general relativity, which is that a photon passing by a point mass M and then reaching an observer from a direction θ relative to the mass will have had its arrival time been delayed by

$$-\frac{4GM}{c^3}\ln\theta$$

by the mass. (Here θ will be very small, so the logarithm will be negative; hence the minus sign.) Now Fermat's principle is that photons take paths which make the arrival time a maximum, minimum or saddle point. In the absence of special delays photons simply minimize the arrival time, i.e. they follow straight lines. However, with a gravitational time delay, photons will follow a different path; this path will be such that the new arrival time (considering both the increased path length and the gravitational delay) is still a maximum, minimum or saddle point. It is often useful to consider the arrival time as a function of the arrival direction of photons from a particular source past some lens: this is known as the 'arrival time surface'. It is related to the wavefront, but is not quite the same thing, for the wavefront is a surface in real space whereas the arrival time surface is a more abstract entity.

The bending angle and arrival time formulations are equivalent. In both cases the effect of a distributed mass (rather than a point mass) is simply the sum of contributions from all points making up the mass distribution.

◊ *Derivation of the arrival time*

Consider a situation as in figure 2, where an observer is viewing a source at distance D_S, with a lens (a mass screen) intervening at distance D_L; D_{LS} is the distance from lens to the source. On galactic scales D_L, D_S, D_{LS} are ordinary distances, but on cosmological scales they must be understood as angular diameter distances, and $D_S \neq D_L + D_{LS}$. (The reason for this complication is that the universe expands significantly over the light travel time.) We use angular coordinates for the transverse position; $\boldsymbol{\theta}_S$ is the position of the source and $\boldsymbol{\theta}$ is its observed position after being lensed. Let $\Sigma(\boldsymbol{\theta})$ be the lens's surface mass density, i.e. mass per unit solid angle. (An important notational point: $\Sigma(\boldsymbol{\theta})$ is often defined as mass per unit physical area, i.e. D_L^{-2} times the convention in this article.) Let $\boldsymbol{\alpha}(\boldsymbol{\theta})$ be the deflection angle. Then, comparing vectors in the source plane, we obtain

$$D_S\boldsymbol{\theta} = D_S\boldsymbol{\theta}_S + D_{LS}\boldsymbol{\alpha}. \qquad (1)$$

(By convention, $\boldsymbol{\alpha}$ is directed outwards from the deflecting mass rather than towards it.) Using $4GM/c^2R$ for the deflection from a point mass, we obtain

$$\boldsymbol{\theta} = \boldsymbol{\theta}_S + \frac{D_{LS}}{D_S}\boldsymbol{\alpha}(\boldsymbol{\theta}) \qquad \boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{4G}{c^2 D_L}\int \frac{\Sigma(\boldsymbol{\theta}')(\boldsymbol{\theta}-\boldsymbol{\theta}')\,\mathrm{d}^2\boldsymbol{\theta}'}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|^2} \qquad (2)$$

This is known as the lens equation. It gives $\boldsymbol{\theta}_S$ as an explicit function of $\boldsymbol{\theta}$, but $\boldsymbol{\theta}$ as an implicit function of $\boldsymbol{\theta}_S$. Moreover,

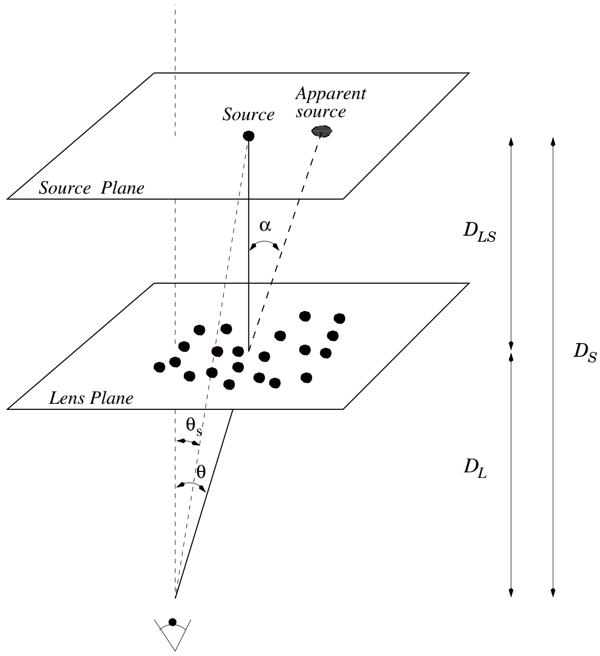

Figure 2. The geometrical meanings of the various distances and angles used in lensing work.

$\boldsymbol{\theta}(\boldsymbol{\theta}_S)$ need not be single valued, so sources can be multiply imaged.

We can change from this bending-angle formalism to the arrival time formalism by noting that the lens equation 2 amounts to equating a gradient to zero:

$$\nabla T(\boldsymbol{\theta}) = 0 \qquad T(\boldsymbol{\theta}) = \tfrac{1}{2}T_0(\boldsymbol{\theta} - \boldsymbol{\theta}_S)^2 - \Psi(\boldsymbol{\theta})$$
$$\Psi(\boldsymbol{\theta}) = (1 + z_L)\frac{4G}{c^3}\int \Sigma(\boldsymbol{\theta}')\ln|\boldsymbol{\theta} - \boldsymbol{\theta}'|\,d^2\boldsymbol{\theta}'$$
$$T_0 = (1 + z_L)\frac{D_L D_S}{c D_{LS}}. \tag{3}$$

Here $T(\boldsymbol{\theta})$ is of course the arrival time, up to an irrelevant constant, for arbitrary deflections: the first term in it is the change in light travel from geometrical considerations alone; the second term is the gravitational time delay (the factor of $1 + z_L$ comes because everything that happens in the lens plane is observed redshifted, and that goes for time delays too). The deflections that actually occur are those that make $T(\boldsymbol{\theta})$ stationary; this is Fermat's principle. Later on, we will rearrange equation (3) to show the dimensions involved a little better. ♦

The Einstein radius

In the absence of lensing the wavefront is spherical. However, in lensing we are always only interested in very small regions of the wavefront, and over a small enough angular region a sphere is indistinguishable from a parabola. Similarly, in the absence of lensing the relevant part of the arrival time surface is also parabolic, with the image of course at the bottom.

Now consider what happens if a point lens is interposed exactly along the line of sight to the source. The arrival time surface will be raised according to the lens's time delay, and if the lens is strong enough the bottom of the parabola will become a dimple. The old minimum will now become a maximum, and the whole edge of the dimple will become a minimum. Observationally, the dimple edge corresponds to a spectacular ring image, called an Einstein ring. Its radius is known as the Einstein radius, say θ_E; if the source is much further away than the lens

$$\theta_E \simeq 0.1 \text{ arcsec} \times \left(\frac{M \text{ in } M_\odot}{\text{lens distance in parsecs}}\right)^{1/2}$$

where $M_\odot$ denotes the solar mass.

Actually the lens need not be a point mass. Just as in Newtonian gravity a sphere attracts externally as if it were a point mass, a circular lens affects light outside the circle as if it were a point mass. So any circular mass can produce an Einstein ring, provided that the Einstein ring is 'outside the mass'. In other words the mass needs to be smaller than a length that projects to θ_E; this length, which we may denote by R_E ($=\theta_E \times$ lens distance) is also called the Einstein radius. (Whether one means θ_E or R_E will always be clear from context.) It turns out that R_E is roughly the geometric mean of the Schwarzschild radius and the lens distance. Thus for the ANDROMEDA GALAXY, R_E is some tens of parsecs—much smaller than the galaxy itself. However, a similar galaxy at cosmological distances will have R_E comparable with or bigger than its own size, and thus can produce strong lensing. Masses that are smaller than their R_E are said to be 'compact' to lensing.

In practice, Einstein rings are rarely observed. The reason is that the above scenario for forming one requires a perfectly circular lens and a perfectly aligned source. These improbable requirements relax a little if the source is very broad, however, and a few examples have been observed. Nevertheless, the *concept* of an Einstein ring is very important, because even if there is no Einstein ring the interesting strong lensing behavior all happens over a scale of θ_E. The larger θ_E, the likelier it is that there are some background sources to strongly lens. Thus from the above expression for θ_E one can easily infer that lensing by massive galaxies ($M \sim 10^{12} M_\odot$ and θ_E perhaps $\sim$1 arcsec) is much likelier than lensing by nearby stars (where $\theta_E \ll 1$ arcsec), while rich clusters of galaxies (with $M \gtrsim 10^{14} M_\odot$ and $\theta_E \sim 10$ arcsec) are even more favorable for lensing.

◊ *The Einstein radius and other scales*

Expressions for θ_E and R_E are easily derived by considering the lens equation for a collinear point mass and source, i.e. $\boldsymbol{\theta}_S = 0$ and $\Sigma(\boldsymbol{\theta}) = M\delta(\boldsymbol{\theta})$. The lens equation is trivially solved by $\theta = \theta_E$, with

$$\theta_E^2 = \frac{4GM}{c^2}\frac{D_{LS}}{D_L D_S} \qquad R_E^2 = \frac{4GM}{c^2}\frac{D_L D_{LS}}{D_S}. \tag{4}$$

As mentioned in the previous section

$$R_{\rm E} \sim (\text{Schwarzschild radius} \times D_{\rm L})^{1/2}. \qquad (5)$$

By a Gauss law type argument, for any circular mass distribution $\Sigma(\theta_r)$, $\Psi(\boldsymbol{\theta})$ and $\alpha(\boldsymbol{\theta})$ will be influenced only by interior mass. So any circular distribution of the mass M, *provided that it fits within an Einstein radius*, can produce an Einstein ring.

For given $D_{\rm L}$, $D_{\rm S}$, to obtain a compact object one needs to pack a mass (in projection) into a circle of radius $\theta_{\rm E}$, but the area of the circle is proportional to the mass. It follows that there is a critical density, say $\Sigma_{\rm crit}$, such that if $\Sigma \geq \Sigma_{\rm crit}$ somewhere then there is a compact (sub)object. Working out the algebra we easily find

$$\Sigma_{\rm crit} = \frac{D_{\rm L} D_{\rm S}}{D_{\rm LS}} \frac{c^2}{4\pi G}. \qquad (6)$$

If now we define κ as the projected mass density in units of the critical density, equation (3) can be rewritten as

$$\nabla\tau(\boldsymbol{\theta}) = 0 \qquad \tau(\boldsymbol{\theta}) = \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\rm S})^2 - \psi(\boldsymbol{\theta})$$

$$\psi(\boldsymbol{\theta}) = \frac{1}{\pi}\int \kappa(\boldsymbol{\theta}') \ln|\boldsymbol{\theta} - \boldsymbol{\theta}'|\, {\rm d}^2\boldsymbol{\theta}'$$

$$T(\boldsymbol{\theta}) = (1 + z_{\rm L})\frac{D_{\rm L} D_{\rm S}}{c D_{\rm LS}}\tau(\boldsymbol{\theta}). \qquad (7)$$

From the second line of equation (7) it should be evident that ψ satisfies a two-dimensional Poisson equation

$$\nabla^2\psi = 2\kappa. \qquad (8)$$

Note that the first two lines in equation (7) involve only dimensionless quantities and the scale of $T(\boldsymbol{\theta})$ is set entirely by the distances and lens redshift. ♦

On images and magnification

Let us now see what can happen to the arrival time surface from masses of arbitrary shape. For small masses, the shape changes slightly from being a parabola and the mininum moves a little. However, for large enough mass, a qualitative change occurs, in that a contour of constant arrival time becomes self-crossing. There are two ways in which a self-crossing can develop: as a kink on the outside of a contour line or a kink on the inside. These are illustrated in figure 3. The outer-kink type is topologically a lemniscate and the inner-kink type a limaçon. If the original contour loop enclosed a minimum then a lemniscate produces another minimum, plus a saddle-point at the self-crossing, while a limaçon produces a new maximum plus a saddle point. (Interchange maximum and minimum in the previous sentence if the original loop enclosed a maximum.) The process of contour self-crossing can then repeat around any of the new maxima and minima, producing more and more new images, but always satisfying

$$\text{maxima} + \text{minima} = \text{saddle points} + 1.$$

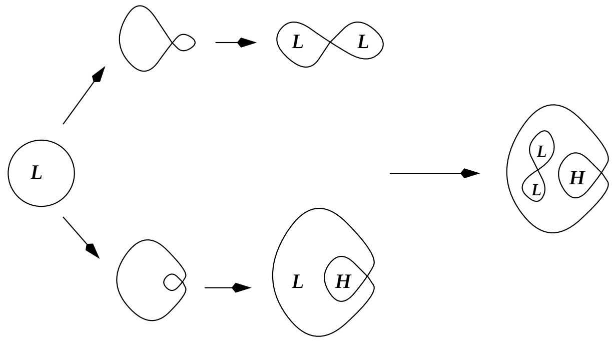

Figure 3. Multiple images via self-crossing contours in the arrival time surface. Here L marks minima and H marks maxima.

The rightmost part of figure 3 shows a typical configuration in observed systems.

Multiple images are not all equally bright. It turns out that each image is magnified by an amount equal to the inverse of the curvature of the arrival time surface. This curvature need not be the same in all directions. Very stretched images (of which an Einstein ring is an extreme example) correspond to very small curvature in one direction with larger curvature in the perpendicular direction. The standard way of quantifying lensing magnification (also called amplification in the literature) is through the 'convergence' which is total change in brightness and the 'shear' which is the amount and direction of stretching.

If the lens's mass density is infinite anywhere (as for any point mass, or say a galaxy with a central BLACK HOLE) the arrival time surface is going to have a needle-sharp but infinitely high maximum at that point, and, while this will formally be an image, because of the needle-sharp curvature the image will be (de)magnified to nothing. For example the five-image configuration in figure 3 is always in practice a four-image system, with the maximum completely demagnified by a very high central density in the lens.

Surface brightness (photon flux per unit sky area per unit telescope area) is conserved by lensing. Magnification changes only angular sizes on the sky. Thus a constant surface brightness sheet stays a constant brightness sheet when lensed. If this were not the case, the microwave background would be wildly lensed by large-scale structure. Polarization of light is also unaffected by lensing.

For unresolved sources, magnification is not observed directly; only their total luminosity changes under lensing. This leads to an interesting effect known as magnification (or amplification) bias. Imagine an unlensed patch of sky that has been surveyed down to some brightness limit for some type of unresolved objects. Now imagine the same patch lensed with some magnification factor A, for definiteness say $A > 1$. This magnification will increase the area of the patch by A, and if the survey area stays the same the source area will be reduced by A. Thus the number of objects available will also be reduced by A. However, at the same time all the objects will be brightened by A, so some objects that were originally below the survey brightness limit will now climb above it. The net effect depends on the luminosity function: if going fainter leads to enough new objects (a steep luminosity function) then the net effect will be an increase in the number of sources detected, otherwise there will be a net decrease.

Magnification

Magnification is formalized as the derivative of the image position with respect to the source position. It is a tensor, which we denote by $\mathbf{M}$. We have

$$\mathbf{M}^{-1} = \frac{\partial \boldsymbol{\theta}_S}{\partial \boldsymbol{\theta}} = \frac{\partial^2}{\partial \boldsymbol{\theta}^2} T(\boldsymbol{\theta}). \tag{9}$$

In Cartesian coordinates

$$\mathbf{M}^{-1} = \begin{bmatrix} 1 - \dfrac{\partial^2 \psi}{\partial \theta_x^2} & \dfrac{\partial^2 \psi}{\partial \theta_x \partial \theta_y} \\ \dfrac{\partial^2 \psi}{\partial \theta_y \partial \theta_x} & 1 - \dfrac{\partial^2 \psi}{\partial \theta_y^2} \end{bmatrix} \tag{10}$$

It is helpful to write $\mathbf{M}^{-1}$ in terms of its eigenvalues, and the usual form is like

$$\mathbf{M}^{-1} = (1-\kappa)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \gamma \begin{pmatrix} \cos 2\phi & \sin 2\phi \\ \sin 2\phi & -\cos 2\phi \end{pmatrix}. \tag{11}$$

The first term in equation (1) is the trace part—and comparing equations (10) and (8) shows that it must be κ—while the second term is traceless. The κ part produces an isotropic expansion or contraction, while the γ part produces a stretching in the ϕ direction and a shrinking in the perpendicular direction; κ is known as 'convergence' and γ as 'shear'.

The determinant of $\mathbf{M}$ can be thought of as a scalar magnification:

$$|\mathbf{M}| = [(1-\kappa)^2 - \gamma^2]^{-1}. \tag{12}$$

The places where one of the eigenvalues of $\mathbf{M}^{-1}$ becomes zero (and in consequence $|\mathbf{M}|$ is infinite) are in general curves and are known as critical curves. When critical curves are mapped onto the source plane through the lens equation, they give caustics; a point source lying on a caustic is infinitely magnified. Some straightforward algebra shows that for a point mass

$$|\mathbf{M}|^{-1} = 1 - \frac{\theta_E^4}{\theta_r^4} \tag{13}$$

and that for a circular mass distribution $\Sigma \propto \theta_r^{-1}$

$$|\mathbf{M}|^{-1} = 1 - \frac{\theta_E}{\theta_r}. \tag{14}$$

The latter case is known as the 'isothermal lens' after the isothermal sphere in stellar dynamics.

One way of proving that surface brightness is conserved by lensing is by setting up a correspondence with Hamiltonian dynamics. Let us write z for the axial coordinate in figure 2, and $\boldsymbol{w}$ for the transverse coordinates, and allow for observers at arbitrary $\boldsymbol{w}$, z. A photon trajectory can be expressed as $\boldsymbol{w}(z)$, and an observer at $\boldsymbol{w}, z$ sees the photon coming from direction $\boldsymbol{\theta} = \mathrm{d}\boldsymbol{w}/\mathrm{d}z$ (in the small-angle regime). In these coordinates the lens equation $\boldsymbol{\theta}_S = \boldsymbol{\theta} - \boldsymbol{\nabla}\psi(\boldsymbol{\theta})$ becomes

$$\begin{aligned} &\boldsymbol{\theta}(\boldsymbol{w}_{\rm obs}, z_{\rm obs}) - \boldsymbol{\theta}(\boldsymbol{w}_{\rm source}, z_{\rm source}) \\ &\qquad = (z_{\rm obs} - z_{\rm lens}) \frac{\partial}{\partial \boldsymbol{w}} \psi \left(\frac{\boldsymbol{w}_{\rm lens}}{z_{\rm obs} - z_{\rm lens}} \right). \end{aligned} \tag{15}$$

However, this is a solution of Hamilton's equations for the Hamiltonian

$$H = \tfrac{1}{2}\boldsymbol{\theta}\cdot\boldsymbol{\theta} - (z_{\rm obs} - z_{\rm lens})\psi\left(\frac{\boldsymbol{w}_{\rm lens}}{z_{\rm obs} - z_{\rm lens}}\right)\delta(z - z_{\rm lens}) \tag{16}$$

with z as a formal time variable and $\boldsymbol{\theta} = \mathrm{d}\boldsymbol{w}/\mathrm{d}z$ identified as the momentum. By Liouville's theorem (or equivalently the collisionless Boltzmann equation), the photon density (say f) in the phase space of $(\boldsymbol{w}, \boldsymbol{\theta})$ space is conserved along photon trajectories. This f must be conserved by the act of placing the lens there too—think of f before and after going through the lens. In fact f is nothing but the surface brightness, because θ corresponds to sky area and $\boldsymbol{w}$ to telescope area. This proves surface brightness conservation, although we must be careful with two things about interpretation. First, because the result is 'along photon trajectories' we must always be looking at photons from the same source; so if the image is moved in the sky by lensing we must follow it when we measure surface brightness. Second, z here is just a formal variable and is neither time nor REDSHIFT, so the above does not imply that surface brightness is independent of redshift.

Lensing in the Milky Way

The Einstein radius for lensing of a Milky Way star by another star or a brown dwarf is, for typical distances, only ~1 AU for R_E and <1 marcsec for θ_E. Nevertheless, because of stellar motions, stars do occasionally manage to come within an Einstein radius (in projection of course) of lenses in the Milky Way. The fraction of source stars in a field that are doing this at a given time is known as the microlensing optical depth, and it is of order 10^{-6} in the Milky Way. ('Microlensing' is a term used when stellar-mass lenses are important.) As the source enters and then exits the Einstein radius of the lens, the image brightens and then dims again in a characteristic way, over timescales of weeks to months. Such lensing is now being detected by large surveys that monitor $\sim 10^7$ stars, searching for the characteristic behavior of the brightness. The aim of these surveys is to measure the optical depth in different directions and thence infer something about the mass density in compact objects in the Milky Way.

Microlensing optical depth

Approximating a star or a brown dwarf as a point lens, there will be two images, at

$$|\boldsymbol{\theta}| = \tfrac{1}{2}[|\boldsymbol{\theta}_S| \pm (|\boldsymbol{\theta}_S|^2 + 4\theta_E^2)^{1/2}]. \quad (17)$$

The image separation is too small to resolve. What will be observed is a brightening equal to the combined magnification of both images. Using equation (13) for the magnification, and adding the absolute values of $|\mathbf{M}|$ at the two image positions, we obtain the total brightness amplification

$$A_{\rm tot} = \frac{u^2+2}{u(u^2+4)^{1/2}} \qquad u = \frac{|\boldsymbol{\theta}_S|}{\theta_E}. \quad (18)$$

If the background source star has impact parameter b and velocity v (projected onto the lens plane) with respect to the lens, then

$$u = \frac{(b^2+v^2t^2)^{1/2}}{R_E}. \quad (19)$$

Inserting equation (19) into (18) gives us $A_{\rm tot}(u)$, i.e. the light curve, plotted for three different b in figure 4. The height of a measured light curve immediately gives R_E/b, and the width gives R_E/v.

We can express the microlensing optical depth (usually denoted by τ) as follows. Using equation (4) for R_E, the total area covered by the Einstein rings of lenses at distances between D_L and $D_L + dD_L$ in a patch of sky is

$$\frac{4\pi GM}{c^2 D_S} D_L D_{LS} \rho(D_L)\, dD_L \times \langle\text{area of patch}\rangle. \quad (20)$$

To obtain the fraction of sky covered by Einstein rings, we divide equation (20) by the area and integrate over D_L, giving

$$\tau = \frac{4\pi G}{c^2 D_S} \int_0^{D_S} D_L D_{LS} \rho(D_L)\, dD_L. \quad (21)$$

The really nice thing about this formula is that it does not depend on the mass distribution of the lenses, as long as each mass fits within its own Einstein radius (diffuse gas clouds do not count, nor does any kind of diffuse dark matter). So τ estimated from light curve monitoring could be used to make inferences about ρ.

Although trying to resolve the images in microlensing seems hopeless with foreseeable technology, it may be possible to track the moving double image indirectly. By combining the positions and magnifications of the two images, we have for the centroid

$$\theta_{\rm cen} = \frac{u(3+u^2)}{2+u^2}\theta_E. \quad (22)$$

Since images can be centroided much better (10^2–10^3 times more precisely) than they can be resolved, measuring the proper motion of the centroid seems a realistic prospect.

Lensing by galaxy clusters

At the other end of the lensing scale, we have clusters of galaxies, which are the most effective lensers of all. A rich cluster at cosmological distance will be effective for lensing over a large enough patch of sky to include plenty of background objects (faint blue galaxies) to lens. A number of clusters show all the regimes of lensing, with multiple-image systems in the inner region, arcs and arclets (highly stretched single images) outside of that and statistical stretching in the outer regions. Each of these regimes contributes different types of information about the mass distribution. Now galaxy clusters are generally not in dynamical equilibrium (there have not been enough crossing times since they formed), so their mass distributions tend to be more complicated than those of single galaxies. This makes it all the more important to have good ways of mapping cluster masses, and makes clusters particularly interesting to researchers in lensing.

Lensing constraints on clusters

There are several approaches in the literature to mass-mapping clusters from lensing data. The following sketches one approach.

Cluster lensing data provide three types of information:

(i) multiple image positions $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and so on;
(ii) relative brightnesses at $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and so on;
(iii) elongated images of sources may be obvious (arcs and arclets) or statistical (requiring a large number of background galaxies to measure).

Going back to equation (7), these can be translated into three types of constraints on $\kappa(\boldsymbol{\theta})$.

(i) Image positions imply $\boldsymbol{\nabla}\tau(\boldsymbol{\theta}_1) = \mathbf{0}$, $\boldsymbol{\nabla}\tau(\boldsymbol{\theta}_2) = 0$ and so on. From equation (7) these are clearly dimensionless linear constraints on $\boldsymbol{\theta}_S$ and $\kappa(\boldsymbol{\theta})$. (If there is more than one multiple-image system, there will be a separate $\boldsymbol{\theta}_S$ for each of these.)
(ii) Relative brightnesses are constraints of the type

$$\boldsymbol{\nabla\nabla}\tau(\boldsymbol{\theta}_1) = \langle\text{measured number}\rangle \times \boldsymbol{\nabla\nabla}\tau(\boldsymbol{\theta}_2).$$

(iii) Elongated images (whether arclets or statistical elongations) provide constraints of the type

$$k\frac{\partial^2}{\partial\theta_{x'}^2}\tau(\boldsymbol{\theta}_1) \le \frac{\partial^2}{\partial\theta_{y'}^2}\tau(\boldsymbol{\theta}_1)$$

where $\theta_{x'}$ and $\theta_{y'}$ are a coordinate system rotated to align with the observed direction of the elongation. This equation expresses the statement that the magnification along $\theta_{x'}$ is at least k times the magnification along the $\theta_{y'}$ direction. Provided that the elongation direction is well-measured, the constraint is linear in $\kappa(\boldsymbol{\theta})$.

Types (i) and (ii) are strong lensing constraints and available only in the inner regions of rich clusters. Type (iii)

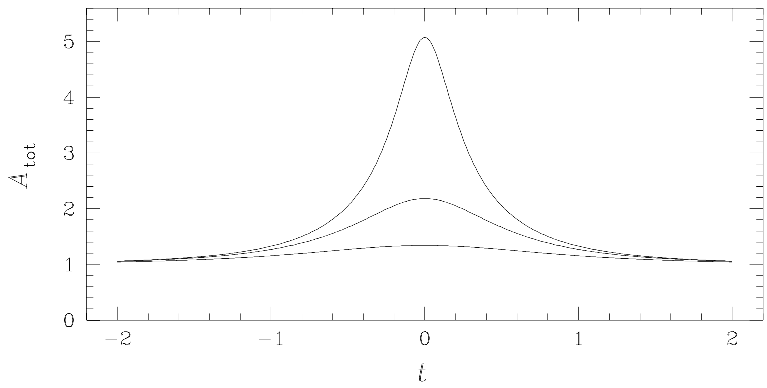

Figure 4. Light curves for impact parameters of R_E (lowest), $0.5R_E$ and $0.2R_E$. The unit of time is how long it takes the source to move a distance R_E.

constraints apply also in the the weak lensing regime and can be extracted over a much larger area (∼5 times further out).

Expressed as above, cluster mass reconstruction becomes an underdetermined problem with linear, plus possibly some quadratic, constraints in both strong and weak lensing regimes. It is thus a relative of image reconstruction problems. The standard way of proceeding to derive an image (or a mass map in this case) would be to pixellate $\kappa(\boldsymbol{\theta})$, select some figure of merit on κ such as smoothness or entropy and then optimize it subject to the data constraints.

Multiple-image quasars

These occur when a QUASAR (or other small but very bright object) is strongly lensed by a foreground galaxy and produces two or four images with arcsec order separations. (A third or fifth image is invariably demagnified to undetectability by the lens's high-density core.) These are of some interest for constraining the masses of the lensing galaxies, but they are perhaps much more interesting to astrophysicists for two other reasons.

The first is that since quasars are often very time variable and the different images have different arrival times, the images will show the same time variability, but with offsets. These offsets, or differences in arrival times, are unique among lensing observables, because they are dimensional. All other lensing observables, i.e. image positions and magnifications, are dimensionless. Measuring a dimensional quantity immediately sets a scale for the system. That scale is the distance scale, which in a cosmological context amounts to HUBBLE'S CONSTANT. Unfortunately, an accurate measurement of variability offsets does not immediately give an accurate determination of Hubble's constant, because the offsets also depend on the distribution of mass in the lens plane; the latter still needs to be determined or modelled somehow. Still, the prospect of measuring Hubble's constant from lensing is attractive because it involves no extra physics apart from gravitation, thus hopefully leaving less room for the sort of systematic errors that plague Hubble constant estimates.

The second has to do with the extremely small size of quasars in optical continuum. Now the mass distribution of a galaxy is not perfectly smooth; it becomes granular on the scale of individual stars. This produces a very complicated network of critical lines (in the lens plane), and a corresponding complicated network of caustics in the source plane (like the pattern at the bottom of a swimming pool). The optical continuum emitting regions of quasars are small enough to fit between the caustics, but the line emitting regions straddle several caustics. As proper motions move the caustic network, the continuum region will sometimes cross a caustic, and show a sudden change in brightness; the time taken for the brightness to change is the time it take to cross the caustic. This is the phenomenon of quasar microlensing: continuum shows it but lines do not. (It is just the gravitational version of stars twinkling and planets not twinkling.) This has been observed, and modelling the caustic network and putting in plausible values for the proper motion leads to an estimate of the intrinsic size of the continuum source. The conclusion is that the optical continuum regions of a quasar—the heart that emits much more light than the entire host galaxy—cannot be much larger than the solar system.

Arrival times and Hubble's constant

In equation (7) $\tau(\boldsymbol{\theta})$ is dimensionless while $T(\boldsymbol{\theta})$ has dimensions of time, and the factor relating them is $\propto H_0^{-1}$. (This factor actually depends on Ω and Λ as well, but a case such as $z_L < 0.5$, $z_S > 2$, the Ω and Λ dependences are very weak.) So if we can measure the arrival time difference $T(\boldsymbol{\theta}_1) - T(\boldsymbol{\theta}_2)$ between two images (and accurate measurements are now starting to appear) and we know $\kappa(\boldsymbol{\theta})$, we can determine H_0.

The difficulty here is that we need to know $\kappa(\boldsymbol{\theta})$, because for a given $T(\boldsymbol{\theta}_1) - T(\boldsymbol{\theta}_2)$ the inferred H_0 can vary by a factor of several depending on the $\kappa(\boldsymbol{\theta})$ assumed. However, naturally, we can use the lensing data themselves to help constrain $\kappa(\boldsymbol{\theta})$. The constraints (i) and (ii) from the cluster lensing case are available here also, since we have a multiple-image system. Constraints of the type (iii) may possibly be available. If there are more than two images then a fourth type of constraint exists, which comes from the ratio of arrival time differences between different pairs of images, i.e.

$$T(\boldsymbol{\theta}_1) - T(\boldsymbol{\theta}_2) = \langle\text{measured number}\rangle \times [T(\boldsymbol{\theta}_2) - T(\boldsymbol{\theta}_3)].$$

Some algebra from equation (7) shows that this is a dimensionless linear constraint on $\boldsymbol{\theta}_S$ and $\kappa(\boldsymbol{\theta})$.

That the lensing constraints are mostly linear and at worst quadratic is a useful simplification, but still the number of constraints possible is small and leaves $\kappa(\boldsymbol{\theta})$ underdetermined, implying some residual uncertainty in H_0 even for perfect lensing data. Hopefully this uncertainty can be reduced by well-motivated modelling or other observations constraining $\kappa(\boldsymbol{\theta})$. However, whether the residual uncertainty in H_0 for any system can be made small enough to be interesting is at present an unanswered question.

Bibliography

The lensing literature being as large as it now is, the following cannot hope to be comprehensive. However, here are pointers to a number of key papers in the development of the field, and to some recent surveys of research in the field.

The first person to take gravitational lensing seriously as an observational prospect appears to have been Zwicky, in

Zwicky F 1937 *Phys. Rev.* **51** 290, 679

It is interesting to compare these papers with

Einstein A 1936 *Science* **84** 506

Einstein sounds almost apologetic for publishing the results for a point lens, because stellar lenses seemed improbable to observe. Zwicky, writing the following year, points out that for galaxies and clusters of galaxies the situation is very different, and lensing becomes 'virtually a certainty'.

Discoveries of the first two lensed systems (both multiply imaged quasars) appeared in

Walsh D, Carswell R F and Weymann R J 1979 *Nature* **279** 381

Weymann R J, Latham D, Angel J R P, Green R F, Liebert J W, Turnshek D A, Turnshek D E and Tyson J A 1980 *Nature* **285** 641

The first arcs in clusters were reported in

Soucail G, Fort B, Mellier Y and Picat J P 1987 *Astron. Astrophys.* **172** L14

Lynds R and Petrosian V 1987 *Bull. Am. Astron. Soc.* **18** 1014

The study of statistical weak lensing by clusters was initiated by

Tyson J A, Wenk R A and Valdes F 1990 *Astrophys. J.* **349** L1

The first quasar microlensing event is in

Irwin M J, Webster R L, Hewett P C, Corrigan R T and Jedrzejewski R I 1989 *Astron. J.* **98** 1989

The first galactic microlenses were published in

Alcock C *et al* 1993 *Nature* **365** 621

Aubourg E *et al* 1993 *Nature* **365** 623

Udalski A *et al* 1993 *Acta Astron.* **43** 289

Our present understanding of lensing theory is due very largely to Sjur Refsdal. The now-standard approximations and basic methods of calculation, and the relation between time delays and Hubble's constant, are set out in

Refsdal S 1964 *Mon. Not. R. Astron. Soc.* **128** 295

Refsdal S 1964 *Mon. Not. R. Astron. Soc.* **128** 307

These long preceded the discovery of the first lens. Quickly following the discovery of the first lensed system

Chang K and Refsdal S 1979 *Nature* **282** 561

predicted microlensing in lensed quasars. Further developments of the formalism appeared in

Kayser R and Refsdal S 1983 *Astron. Astrophys.* **128** 156

introducing the wavefront picture, and

Blandford R D and Narayan R 1986 *Astrophys. J.* **310** 568

Schneider P 1985 *Astron. Astrophys.* **143** 413

which introduce Fermat's principle by very different routes.

The most comprehensive book on lensing is

Schneider P, Ehlers J A and Falco E E 1992 *Gravitational Lenses* (Berlin: Springer)

The most recent conference proceedings covering all aspects of lensing is

Brainerd T G and Kochanek C S (ed) 2000 Gravitational lensing: recent progress and future goals *ASP Conference Series*

For eight unsolved problems in gravitational lensing, see

Blandford R D 1997 *Unsolved Problems in Astrophysics* ed J N Bahcall and J P Ostriker (Princeton, NJ: Princeton University Press)

P Saha