

Foundations of Geometrical Optics: Phenomenology and Principles

By Dwight E. Neuenschwander

Geometrical optics is the study of the trajectory of light modeled as rays. One of the oldest branches of physics, two of its three main laws were articulated before 300 BCE, and the third was stated quantitatively by the year 1000 CE.

Whenever light is modeled as waves, the rays are perpendicular to the wave fronts. But by working only with the rays, geometrical optics does not “know” about diffraction and interference, the hallmarks of wave phenomena. However, wave optics reduces to geometrical optics in the short-wavelength approximation, where the dimensions of apertures and obstacles are vastly greater than the wavelength, and diffraction is negligible.[1]

Throughout this article I will be talking about light, but with the appropriate caveats geometrical optics applies to acoustics as well.

Geometrical optics for light was understood centuries before anyone could answer with data-based confidence the question of what light really is. Geometrical optics therefore does not depend on the underlying physical mechanism. Accordingly, its laws are robust, derivable from at least three approaches: Fermat’s principle, Huygens’ principle, and Maxwell’s equations. First let us look at the three main phenomenological laws of geometrical optics. Then we will discuss the principles that bring coherence to them, and conclude with a productive application.

Phenomenology: Three Laws of Geometrical Optics

Like all sciences, geometrical optics is based on empirical observations, three in this case. They are (1) the law of rectilinear propagation, (2) the law of reflection, and (3) the law of refraction. The first of these describes light propagating freely through a uniform medium. The laws of reflection and refraction describe what happens when the ray interacts with matter, in particular, when a ray encounters the interface between two media.

People have known for millennia that light seems to travel in a straight line.[2] Such a notion was justified by technical and practical matters ranging from hunting and navigation, to

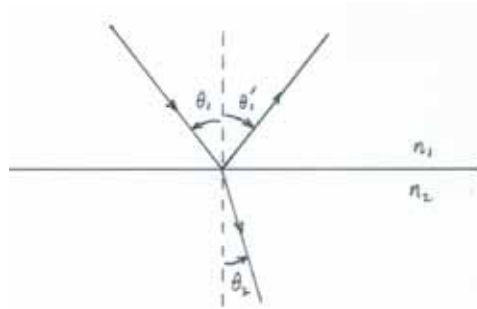


Fig. 1. Illustrating the laws of rectilinear propagation, reflection, and refraction. In this instance, $n_2 > n_1$. The ray can travel either way, bending towards the normal in going from medium 1 to medium 2, and bending away from the normal when traveling in the opposite direction.

surveying and signaling. The law of reflection has been known almost as long. Mirrors were mentioned in the book of *Exodus*, written about 1200 BCE. Mirrors of polished bronze and copper were made in ancient societies; for instance, a mirror in pristine condition, dating from about 1900 BCE, was found in Egyptian excavations of the pyramid of Sesostrius II. The laws of rectilinear propagation and reflection were described by Euclid in his book *Catoptrics*, written about 300 BCE.[3]

The laws of reflection and refraction can be discussed with quantitative precision with the aid of Fig. 1. The horizontal line represents the interface between medium 1 and medium 2. Note that within each medium, the law of rectilinear propagation is respected. We use the convention, introduced by the Arab scholar Alhazen about the year 1000 CE, of measuring the angles of the incident, reflected, and refracted rays from the normal to the interface.

The law of reflection is succinctly stated as the equality of the angle between the normal and incident ray and the normal and reflected rays:

$$\theta_i = \theta_r. \quad (1)$$

Refraction has also been noted with commentary from Antiquity, although it was not described quantitatively until much later. A stick appearing to be bent when partly immersed in

water is an everyday observation which we find mentioned in Plato’s *Republic*, written about 360 BCE. The “burning glass”, a converging lens used to start fires by focusing sunlight, was a well-known feature of ancient technology. For instance, a magnifier was unearthed in the ruins of the palace of Assyrian King Sennacherib (704–681 BCE), and a converging lens was mentioned in a play called *The Clouds*, written by Aristophanes in 424 BCE. Since pre-history all fishermen would have noticed that the apparent depth of a pond is less than its actual depth; the ratio of the depths D_{actual} to D_{apparent} was perhaps the first inkling of what came to be formally called the index of refraction, n :

$$D_{\text{actual}} / D_{\text{apparent}} = n \quad (2)$$

so that $n \geq 1$. For water, $n = 1.33$; for various kinds of glass, n ranges from about 1.45 to a little over 1.95; a value of 1.5 offers a good back-of-the-envelope estimate for common glasses. For air, $n = 1 + \epsilon$, where $\epsilon = 0.00029(\rho / \rho_0)$, ρ denotes the air’s actual density, and ρ_0 its density at standard pressure and temperature.[4]

When it became possible in the 17th century to measure the speed of light with some precision, the refractive index acquired a dynamical role beyond its geometrical one. If c denotes the speed of light in vacuum (about 3×10^8 m/s),[5] then the speed of light v in a refractive medium of index n is

$$v = c/n. \quad (3)$$

Christian Huygens (1629–1695), in his treatise *Traité de la lumiere* of 1678, argued in favor of light as a wave. He envisioned each point on a wave front as the source of another wave. With this concept, through geometric constructions of great elegance (the “Huygens construction” [6]), he was able to derive the laws of rectilinear propagation, reflection, and refraction. To do so Huygens had to assume Eq. (3).

The law of refraction, which quantifies the change in direction of a ray when going from one medium to another, was accurately described by the 10th-century Arabian mathematician and

optical engineer Abu Sa'd al-Alá ibn Sahl (c. 940–1000), who was retained by the court at Baghdad.[7] In his book *On Burning Mirrors and Lenses*, written about 984, he accurately describes refraction in terms of two right triangles, as shown in Fig. 2 below. There we see a ray undergoing refraction as it goes from medium 1 to medium 2, along with the line drawn by extending the incident ray's direction into the second medium. Let us draw a vertical line that intersects the refracted and extended incident ray to form right triangles OAB and OCD. Their hypotenuses have lengths h_1 and h_2 , respectively. For the refraction of a light ray from air into a crystal of refractive index n , Ibn Sahl stated the law of refraction as

$$h_1/h_2 = 1/n. \quad (4)$$

More generally, at the boundary between any two media, Ibn Sahl's law says

$$h_2/h_1 = n_2/n_1. \quad (5)$$

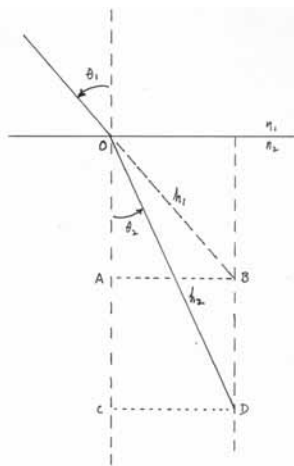


Fig. 2. Ibn Sahl's construction of the law of refraction.

In his treatise Ibn Sahl applied this quantitative description of refraction by lenses, and along with the law of reflection for mirrors, described the design of focusing devices such as converging lenses and parabolic and elliptical mirrors. Ibn Sahl influenced other Arabic scholars of optics, such as Ibn al-Haitham (965–1039), but unfortunately for science in general, his manuscript was dispersed between two libraries until it was assembled and translated by R. Rashed in 1990.[7] Meanwhile, Willebrord Snell (1580–1626), a professor at Leyden University, empirically rediscovered refraction's geometry in 1621, but he did not publish his findings. In *La Dioptrique* of 1637, René Descartes first published Snell's law (but neglecting to give Snell credit) in its now-familiar form,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \quad (6)$$

Although he wrote the law of refraction without trig functions, we can see that our Arabian colleague Ibn Sahl had it right—starting with Snell's law, Eq. (6), and using Ibn Sahl's triangles, we derive Eq. (5). The argument can also be reversed. Therefore Ibn Sahl's law and Snell's law are equivalent.

Theory: Principles and Inferences Derived from Them

A collection of phenomenological laws, arrived at inductively from repeated experience, give us some practical knowledge of *what* happens. Such statements are called “laws of nature”, but perhaps they would be better acknowledged as “correlations”. By themselves such statements do not give us understanding of *why* a system behaves as it does, nor do they say *how* phenomenological laws are related to one another. For that we need to go deeper, from “laws” to “principles”. We need to postulate principles from which the laws can be derived, which pull the laws together into a connected, coherent system and lay out their limitations. In other words, we need a theory.

One of the first attempts at theoretical natural philosophy was taken by Hero of Alexandria (c. 10–70 CE), who suggested that the optical laws of rectilinear propagation and reflection could be understood from a single unifying principle: that rays travel from one point to another along the path of minimum distance. In the case of reflection, the light must go from the source to the observer's eye while also touching the mirror. To give the shortest distance with this constraint, the angle of incidence and reflection from the mirror must be equal. Because the shortest distance between two points is a straight line,[8] Hero's principle explains the laws of rectilinear propagation and reflection as consequences of the “least distance” axiom.

Hero's principle cannot account for refraction, but the principle could be effectively generalized, as done in 1657 when Pierre de Fermat (1601–1665) articulated a “least time principle” that we now call Fermat's principle. It asserts that of all conceivable paths between two fixed points, the path actually followed by a light ray is the one for which the elapsed time is minimized.

That the least-time principle *must* give a refracted line as in Fig. 1 can be beautifully illustrated with an analogy (which depends on Eq. 3) that I borrow from Richard Feynman. [9] It goes like this: A lifeguard on the beach sees a swimmer in trouble offshore, some distance down the beach from the lifeguard's position. Because the lifeguard can run faster than she can swim, to reach the swimmer in the

minimum *time*, the lifeguard will not take the shortest *distance* between her initial location and the flailing swimmer; rather, she will follow a path similar to the refracted ray of Fig. 1. To minimize the time when the speed varies with medium, the path *must* exhibit refraction.

To express Fermat's principle quantitatively, while allowing for the general possibility of a refractive index n that varies with position, suppose the light goes from a fixed initial point a to another fixed final point b . By Eq. (3) a time increment may be written $dt = ds/v = n ds/c$, where ds denotes an increment of length. The elapsed time T for the entire trip for a to b will be the integral

$$T = \int_a^b n ds/c. \quad (7)$$

The distance $n ds$ is called the optical length. Fermat's principle postulates that the actual trajectory of the ray between fixed points will be the one for which cT , or equivalently, the integrated optical length $\int n ds$, is a minimum. Fermat's principle subsumes Hero's principle as the special case of uniform n , but Fermat also goes beyond Hero to include nonuniform media. For geometrical optics, the shortest *time*, in other words the shortest *optical length*, is more fundamental than the shortest *distance*.

Fermat's principle already contains the law of rectilinear propagation in a uniform media, and the laws of refraction and reflection also follow at once from it. Consider a light ray that travels between fixed endpoints in media throughout which the index of refraction n is a piecewise constant.[10] Let the plane made by the incident and refracted ray define an xy coordinate plane with the x -axis along the boundary between two media (see Fig. 3). Also, let the index of refraction be n_1 above the x -axis and n_2 below it. By the law of rectilinear propagation, the rays above and below the x -axis are straight lines. Fermat's principle says we must minimize the optical length cT , where

$$cT = n_1 s_1 + n_2 s_2 \quad (8)$$

and s denotes the length of a rectilinear line segment. Let a light ray move from (a,b) to (A,B) by way of some point $(\epsilon,0)$. The time T can be written as a function of this variable ϵ :

$$cT(\epsilon) = n_1 [(\epsilon - a)^2 + b^2]^{1/2} + n_2 [(A - \epsilon)^2 + B^2]^{1/2}. \quad (9)$$

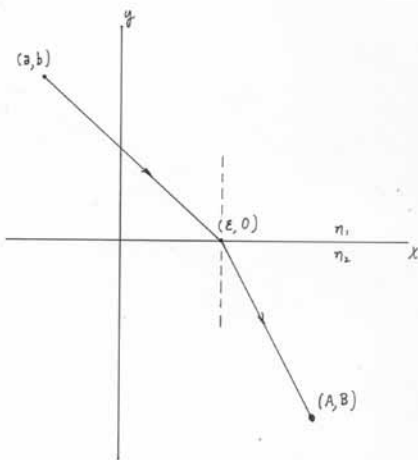


Fig. 3. The geometry of the derivation of Snell's law and the law of reflection using Fermat's principle.

According to Fermat's principle, ϵ will be such that $T(\epsilon)$ is a minimum. In calculus we learned this means that the first derivative must vanish. Setting $dT/d\epsilon = 0$ yields

$$0 = n_1(\epsilon - a)/s_1 - n_2(A - \epsilon)/s_2 .$$

From the geometry of Fig. 3 this can be written in terms of sines of the angles, to give Snell's law, Eq. (6).

If the wave is reflected, so that we replace B with $-B$ and set $n_1 = n_2$, then our derivation gives the law of reflection.

In general, some of the energy that arrives at the interface will be reflected and some of it refracted. But to derive those results from Maxwell's equations (the so-called Fresnel equations for light[11]) requires an excursion into wave optics, in particular, an appeal to the boundary conditions on Maxwell's equations, which lies outside the scope of this discussion.

As mentioned above, geometrical optics is a robust topic, in that its empirical laws can be derived from at least three distinct starting points, each taken as fundamental principles: Fermat's principle, Huygens' principle, and Maxwell's equations. The principles of Fermat and Huygens apply to rays in general and can be used for acoustics as well as the optics of light. But as already mentioned, light as waves is a crucial *assumption* of Huygens's principle. The laws of rectilinear propagation, reflection, and (with Eq. 3) refraction follow from its geometric construction whereby each point on a wave front is a point source for another wave.[6]

In contrast, the existence and properties of electromagnetic waves are *deductive consequences* of Maxwell's equations, which relate electric and magnetic fields to their charge and current sources and relate these fields to each

other. When Maxwell's first-order differential equations are combined into second-order ones, we obtain wave equations in which the wave's speed equals the speed of light in the medium. Within this paradigm light is *shown* to be a wave in the electromagnetic field. Furthermore, the boundary conditions implied by Maxwell's equations also lead deductively to the laws of reflection and refraction, and to the Fresnel equations as well, mentioned above, for the relative amplitudes of the reflected and refracted waves.[11] Since we are concerned in this article with ray optics only, we leave these wave principles for another day, but we note here how they demonstrate the robustness of the laws of rectilinear propagation, reflection, and refraction of rays.

Of these three sets of principles—Fermat, Huygens, and Maxwell—only Fermat's principle can be said to be “pure ray optics”, because unlike the other two, it makes no assumptions or predictions about the underlying mechanism represented by the rays.

Ubiquitous applications of our three laws of geometric optics are found in mirrors and lenses, to which we next turn our attention.

Thin Lenses and Spherical Mirrors

Because a ray's entrance into and exit from a lens involves two refractions, and because the interfaces must be curved to focus a bundle of rays, we must set up a program to handle a succession of refractions through curved surfaces. Toward that end we derive, via Fermat's principle, the so-called “lens maker's equation”.

Let two media of uniform refractive indices n_1 and n_2 lie respectively to the left and right of a curved boundary between them (see Fig. 4). Let this curved boundary have radius of curvature R , with the center of curvature located at point C . Let the object, the source of a ray, be located at point A , and let o denote the distance AB (the object distance). Let the ray travel from A and encounter the boundary at point P , where the ray gets refracted back toward the AA' axis, which it crosses at point A' . The ray APA' and another ray ABA' intersect at point A' and thus form the image. Let i denote the distance BA' (the image distance). In going from A to A' via P the ray travels a path of length $s_1 + s_2$. Fermat's principle says the path actually followed will minimize the elapsed time T , where

$$cT = n_1 s_1 + n_2 s_2 . \quad (10)$$

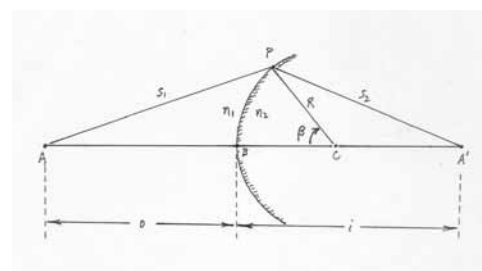


Fig. 4. The geometry used in the derivation of Eq. (11).

As the location of P changes, so does the angle β (angle BCP). The law of cosines for triangles $AA'P$ and $CA'P$ allows us to write s_1 and s_2 in terms of $\cos \beta$.

Fermat's principle instructs us to set equal to zero the derivative of T with respect to β . For P off the AA' axis, but for $\beta \ll 1$, we obtain

$$n_1/o + n_2/i = (n_2 - n_1)/R. \quad (11)$$

(Incidentally, one can also obtain this result from Snell's law directly by applying it to the refraction at point P , then using geometry on the triangles mentioned above and making small-angle approximations.)

We build a lens by considering *two* curved surfaces with which our light ray will refract, and apply Eq. (11) to the two consecutive refractions. Let the first image be located at distance i' to the right of the first surface. The refraction at the first surface is as shown in Fig. 4, where its radius of curvature is now denoted R_1 . In this refraction the ray moves from air of essentially unit refractive index, into glass of index $n > 1$. Equation (11) becomes

$$1/o + n/i' = (n - 1)/R_1 . \quad (12)$$

The image of the first refraction becomes the object for the second refraction. Let the on-axis thickness of the lens be τ , in which case the object distance for the second refraction is approximately $\tau - i'$. The second refraction has the ray coming from a medium of index n and going back into air, where it forms the final image at the distance i from the second interface, whose radius of curvature is R_2 . Eq. (11), when applied to this second refraction, gives

$$n/(\tau - i') + 1/i = (1 - n)/R_2. \quad (13)$$

In the thin lens approximation, where τ is negligible compared to all other distances, Eqs. (12) and (13) together give the lens maker's equation,

$$1/o + 1/i = 1/f_{\text{lens}} \quad (14)$$

where

$$1/f_{\text{lens}} = (n - 1)(1/R_1 - 1/R_2) \quad (15)$$

with f_{lens} the so-called focal length of the lens. This name comes from the observation that an object at infinity, whose rays are all parallel as they approach the lens, has $1/o = 0$, so that $i = f$; all such rays converge, or focus, to the same place. By symmetry, each lens has two focal points, located to either side, each at the distance $|f|$ from the (“thin”) lens. That two symmetrically placed focal points exist will become evident (I hope) from the sign convention discussions which follow below.

Similarly, for a spherical mirror having radius of curvature R , one may show that the relation between the object and image distances is again given by Eq. (14), but for mirrors the focal length is

$$f_{\text{mirror}} = \frac{1}{2} R. \quad (16)$$

A mirror has only one focal point, either in front of a concave mirror or behind a convex mirror.

Unlike the usual meaning of “distance”, the object distance o , image distance i , and focal length f are algebraic quantities that can be positive or negative, as discussed below.

Thin or not, converging lenses are thickest in the middle, and diverging lenses are thinnest in the middle. Converging lenses have a positive focal lengths, and diverging lenses have negative focal lengths.[12] Concave mirrors (looking at the front of a shiny spoon) have positive focal length, while convex mirrors (back of the spoon) have negative focal length.

Important: Do not memorize any of these sign conventions. Although the adjectives “real” and “virtual” and the sign conventions for o , i , and f can cause plenty of confusion, I assure you that there is a system to it, one version of which I will share with you below, with a couple of illustrative examples. But before doing any calculations it is always useful to represent the system visually as much as possible. In the case of thin lenses or spherical mirrors, this means drawing ray diagrams.

Note carefully that I use the convention in all my ray diagrams that the original ray comes in from the left. This is an arbitrary choice, like the choice of a coordinate system: the important thing is to be consistent. Let us start our ray tracing with a converging lens (see Figs. 5a and 5b). An incoming ray parallel to the lens axis passes through the focal point on the far side of the lens. Why does it do this? As noted above, if all the rays coming from an object were parallel to the lens axis, then the object would be located at infinity, and consequently, $i = f$ by Eq. (14). If the object is not at infinity, then not all the rays emanating from it will approach the lens parallel

to its axis, but one of the rays will be parallel, and will pass through the far-side focal point after interacting with the lens. By symmetry, an incoming ray that passes through the near-side focal point will emerge from the lens moving parallel to its axis. In contrast, a ray that passes through the lens center will emerge essentially undeflected (to this ray the center of the lens is essentially a flat window pane). The image is formed where these rays (and others not drawn) intersect. Notice what happens if the object lies between a focal point and the lens (Fig. 5b).

For a diverging lens, an incoming ray parallel to the lens axis will be deflected away from the axis, such that the extension of it backwards passes through the first focal point (see Fig. 5c). Similar considerations hold for spherical mirrors (see Figs. 6a and 6b).

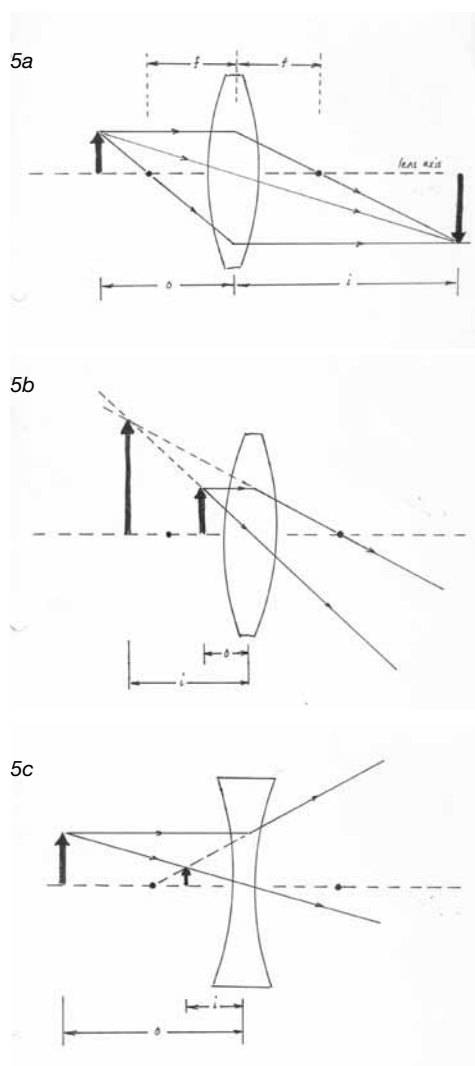


Fig. 5. Ray tracing for a converging lens (a, b) and a diverging lens (c)

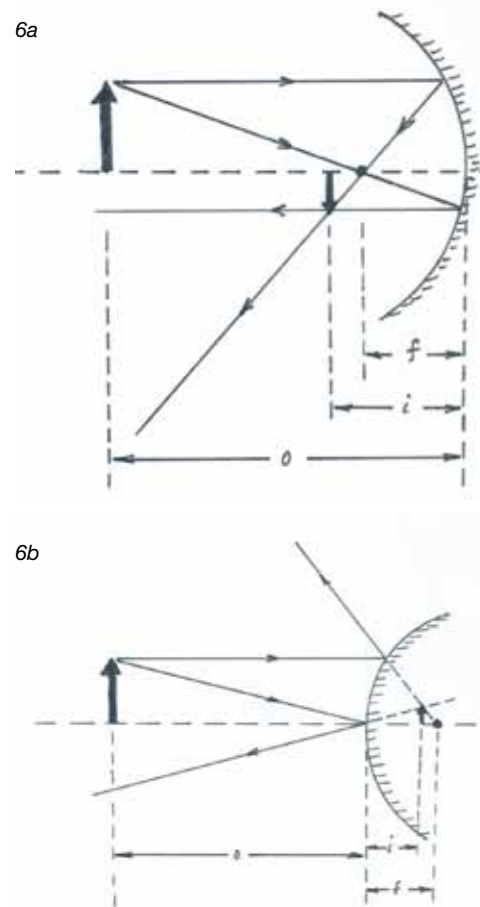


Fig. 6. Ray tracing for (a) a concave and (b) a convex mirror.

Let us consider the appearance of the image relative to the object, in particular, its lateral magnification. If the axial symmetry axis of the lens defines an x -axis, then a y -axis would be perpendicular to it, and my arrows that represent the object and image lie parallel or anti-parallel to this y -axis. Suppose (see Fig. 7) that the tip of the object arrow lies at y -coordinate h_o , and the tip of the image arrow lies at y -coordinate h_i .

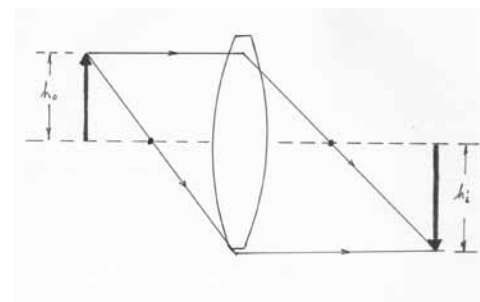


Fig. 7. The lateral coordinates used in discussing magnification.

In terms of the object and image distances, the ratio of object and image heights can be written at $h_i = mh_o$, which defines the lateral magnification. By trigonometry, the lateral magnification m is also equal to

$$m = -i/o \quad (17)$$

which is an algebraic quantity because o and i (like h_i and h_o) can have either sign. A positive magnification means that the image is upright relative to the object, but $m < 0$ means the image is inverted. The absolute value of m gives the factor by which the image size is laterally rescaled relative to the object.

Now let us consider a couple of numerical examples. I will use the sign conventions that are explained in the next section.

Consider two converging lenses that are placed 50 cm apart. Suppose both lenses have the same focal length, say 10 cm. Place an object 15 cm to the left of Lens 1. This lens, according to Eq. (14), produces an image at $i_1 = 30$ cm. Because $i_1 > 0$, this means the first image is produced 30 cm to the right of Lens 1 (the reader should draw a ray diagram).

Now this intermediate image becomes the object for Lens 2. Since this second object is to the left of Lens 2, its object distance is $(50 - 30)$ cm = +20 cm (it is a conventionally placed object). Running Eq. (14) for Lens 2, we find $i_2 = 20$ cm.

What is the magnification of the two-lens system? The first lens gives a magnification $m_1 = -i_1/o_1 = -(+30 \text{ cm})/(+15 \text{ cm}) = -2$, the image inverted and rescaled by a factor of 2 compared to the object. Lens 2 sees this intermediate image as its object and gives it a magnification $m_2 = -i_2/o_2 = -(+20 \text{ cm})/(+20 \text{ cm}) = -1$, inverted and the same size as its object. Together, the two lenses give for the final image the magnification $m = m_1 m_2 = (-2)(-1) = +2$, upright and doubled in size compared to the original object. Notice that as a succession of rescalings, magnifications combine multiplicatively.

Mundane but Important: Real Side, Virtual Side, and Sign Conventions

I want to share with you here the simplest system I can think of for keeping straight the sign conventions for image distances, focal lengths, and object distances. Note carefully that I am about to show you a *convention*; like the choice of which direction—up or down—will be positive in free-fall problems, there is more than one way to do it, but they are all OK if each is internally consistent. I am sharing with you a personal choice for lens and mirror problems, a choice that makes sense to me and which seems

to have worked well with my students.

The crucial concept in the convention I'll describe here is the distinction between the “real side” and the “virtual side” of the lens or mirror. Recall that I am locating the original object to the left of the lens or mirror (or to the left of the first lens or mirror in a multicomponent system). Therefore I am assuming the light to be incident from the left *before* it encounters the lens or mirror. After it interacts with a lens, the light goes *through* it to the other side (viz., the light is on the right *after* encountering the lens, Fig. 8a). In contrast, after it reflects from a mirror, the light is still on the *same* side (the left in my convention) after the interaction (Fig. 8b).

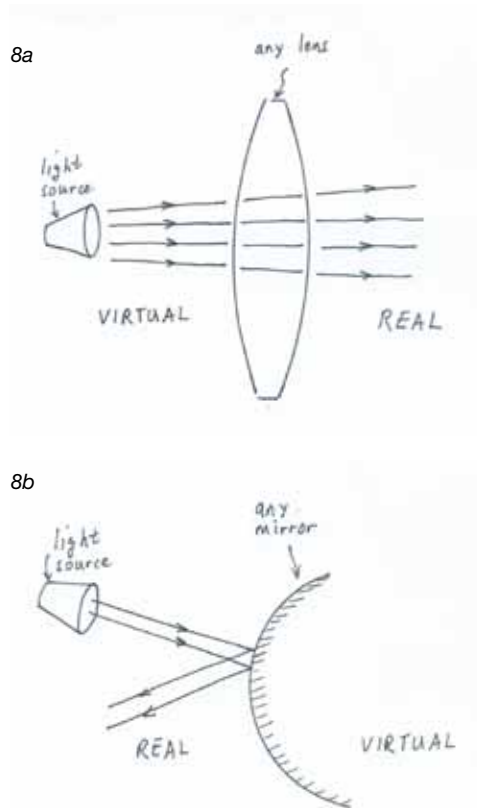


Fig. 8. The definitions of “real” and “virtual” sides of (a) a lens and (b) a mirror.

Definition: The “real side” of the lens or mirror is the side where the light energy “really is” after it encounters the lens or mirror. The “virtual side” is merely the side opposite the real side. Now we can make sign conventions for i and f . For these, quantities “on the real side” will be positive, and quantities “on the virtual side” will be negative. I will deal with object distances separately.

Whether lens or mirror, an image on the real side means $i > 0$, and $i < 0$ means the image forms on the virtual side.

A concave mirror has $f > 0$ because its *center*

of curvature lies on the real side. A convex mirror has $f < 0$ because its center of curvature lies on the virtual side.

The sign conventions for the focal lengths of lenses use the same system, where signs of radii of curvature are determined by the location of their centers of curvature. But unlike mirrors, two centers of curvature have to be considered for lenses. Thereby the sign of f can be understood from Eq. (15), which I rearrange as

$$1/f_{\text{lens}} = (n - 1)(R_2 - R_1)/(R_2 R_1) \quad (18)$$

A converging lens (thickest in the middle) has $f > 0$ for the following reason: The left lens surface (the first surface the light encounters) has its center of curvature on the real side, which makes its radius of curvature positive, $R_1 > 0$. The second surface has its center of curvature on the virtual side, so that $R_2 < 0$. With these signs for the radii of curvature, Eq. (18) gives a positive focal length.

A diverging lens (thinnest in the middle) has $f < 0$. Why? Because the left lens surface has its center of curvature on the virtual side, which makes $R_1 < 0$. The second surface has its center of curvature on the real side, so that $R_2 > 0$. In Eq. (18) these give a negative focal length.

In the case of the object distance o , I prefer *not* to use the language of “real” and “virtual”. Recall that I am using the convention of the original rays coming in from the left. In my diagrams and calculations, that is where the object is originally located, relative to the (first) lens or mirror. Let this arrangement be called the “conventional object placement”. When the object is conventionally placed, then $o > 0$, by definition. Thus, for the first refraction or reflection, o is always positive. Non-conventionally placed objects, and thus the possibility of $o < 0$, occur in my convention only with systems that have more than one lens and/or mirror. Such multiple-component setups are essential to science; we call them “telescopes” and “microscopes”.

We saw that in the interactions of a ray with an optical system having multiple lenses or mirrors, the image produced by the first lens or mirror becomes the object for the next lens or mirror. If that intermediate image, i.e., the object for the second lens, is conventionally placed relative to the second lens or mirror, then the second object distance is positive. But if that intermediate image (second object) lies to the right of the second lens or mirror, in other words, if it is “anti-conventionally placed”, then for it $o < 0$. Let us revisit the two-lens example above and modify it to show this situation.

Return to our two converging lenses, both with $f = 10$ cm, and the same original object

conventionally located 15 cm before the first lens. We saw that its image occurs at $i_1 = 30$ cm and that $m_1 = -2$. But now, instead of the second lens being located 50 cm to the right of the first lens, suppose it is located 20 cm from it. Thus, $o_2 = -10$ cm. The thin lens equation, Eq. (14), gives $i_2 = +5$ cm, and thus $m_2 = +\frac{1}{2}$. Both lenses together give a magnification $m = -1$. The final image is real, inverted, and the same size as the original object.

With these principles and sign conventions in mind, you, dear reader, are now invited to consider all manner of applications, such as telescope and microscope designs. In the case of telescopes, some are refractors and some are reflectors. Some of the refractors are designed for astronomical work, and some are for terrestrial observations. In the former it's all right if the image is inverted, but not for the latter. There are at least two designs for making a refracting telescope that gives an upright final image (e.g., the Galilean telescope). Among the reflecting telescopes, which use a primary mirror (and sometimes a secondary one) followed by a converging lens eyepiece (or a camera), we find the Newtonian, Cassegrain, and Gregorian designs.

The field of geometrical optics is, of course, far too huge to cover here. Its applications range from simple devices such as the *camera obscura* (so called by Renaissance artists such as Johann Vermeer who used them, but better known to us as the pinhole camera), to the large-scale design features of the Hubble space telescope. But departures from geometrical optics, as revealed by phenomena such as chromatic aberrations and diffraction, where the wave nature of light reveals itself—not to mention quantum optics and photonics—makes geometrical optics an excellent place to begin the study of physics itself. From the wave/particle dualism that has formed recurring paradigms across the history of physics; from seeing Newtonian mechanics as the short-wavelength limit of quantum mechanics; and with Fermat's principle serving as a model for other fundamental variational principles such as Hamilton's principle for mechanics and its extensions to relativity and field theory—geometrical optics strikes a cord that resonates across all the elegant connections in physics.

Acknowledgment

I am grateful to Robert Hilborn for bringing the work of Ibn Sahl to my attention, and to Thomas Olsen for his thoughtful suggestions regarding a draft of this manuscript.

[1] The relation between wave optics and geometrical optics is echoed in the relation between quantum mechanics and Newtonian mechanics. In both cases the latter is the short-wavelength limit of the former. In the quantum mechanics case, the wavelength of which we speak is the de Broglie wavelength of the harmonic wave that corresponds to the motion of a free particle.

[2] Although “everyone knows” that light travels in a straight line in a uniform medium, does it *really*? For example, Einstein's theory of general relativity predicts that a massive body will deflect a ray, even in vacuum. However, the effect is so weak that we can ignore it in everyday life.

[3] While there are historical sources galore, succinct historical notes as a secondary source are found in the various editions of *Optics* by Eugene Hecht (Addison-Wesley, e.g., 2002 for the 4th ed.).

[4] Bruno Rossi, *Optics* (Addison-Wesley, 1965), p. 98, cites as the standard conditions for ρ_0 a pressure of 76 cm Hg and a temperature of 288 K.

[5] Equation (3) can also be used for acoustics, where c denotes the speed of sound under standardized conditions.

[6] See any general physics textbook or optics text for discussions of the Huygens construction.

[7] E.g., see R. Rashed, “A pioneer in anaclastics: Ibn Sahl on burning mirrors and lenses,” *Isis* **81**, 464–491 (1990).

[8] With the advent of non-Euclidean geometries, we realize that the shortest distance between two points, called the geodesic, may or may not be “straight” as we understand “straightness” in Euclidean space. Of course, Hero had in mind Euclidean geometry, and so do we in everyday life.

[9] Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics* (Addison-Wesley, 1963), Vol. I, Ch. 26, p. 4.

[10] Since a medium with a continuously varying index of refraction can be conceptualized as infinitesimal layers, each of which has a uniform n , no loss of generality ensues from considering only two media.

[11] See any intermediate or advanced textbook on classical electrodynamics, such as David Griffith's, *Introduction to Electrodynamics*, 3rd ed. (Prentice-Hall, 1999).

[12] These correlations between the shapes of

lenses and whether the rays converge or diverge can be understood only if we step outside of geometrical optics for a moment and visit the wave model of light. The speed of light outside the lens is essentially c , but inside the lens the speed drops to c/n . Consider plane waves striking the lens. If the lens is thickest in the middle, then the part of the wave front passing through the lens center emerges *after* the parts of the same wave front that entered near the edge. The emerging wave front is thereby reshaped into a converging configuration. A lens thinnest at the center will make the wave fronts passing near the center emerge *before* the parts that pass through the edges, resulting in a diverging wave.