1

Introduction

> The science of optics, like every other physical science, has two different directions of progress, which have been called the ascending and the descending scale, the inductive and the deductive method, the way of analysis and of synthesis. In every physical science, we must ascend from facts to laws, by the way of induction and analysis; and must descend from laws to consequences, by the deductive and synthetic way. We must gather and groupe appearances, until the scientific imagination discerns their hidden law, and unity arises from variety: and then from unity must re-deduce variety, and force the discovered law to utter its revelations of the future.
>
> *William Rowan Hamilton (1805–1865)*[1]

It is a fact of immediate importance to our everyday experience that light nearly always travels in straight lines from the source to our eyes, perhaps scattering off some object along the way.[2] Without the ability to assume this as a fact about the world around us, our extraordinary talent for instinctively comprehending spatial relationships in everyday life would be severely compromised. Consider how much computer power must be expended to disentangle the multiple images of distant galaxies[3] to map the dark matter distribution in the visible universe [MRE+07]. Imagine what life would be like if we had to do similar mental computations just to navigate around the furniture in our living room.[4]

[1] From "On a general method of expressing the paths of light, and of the planets, by the coefficients of a characteristic function," by WR Hamilton (1833) [Ham33].

[2] That rays travel in straight lines was fully appreciated by the ancient world. See, for example, Euclid's Optics (ca. 300 BCE) [Bur45] which begins with the statement: "Let it be assumed that *lines* drawn directly from the eye pass through a space of great extent." (emphasis added) From this simple insight, Euclid lays the groundwork for *geometrical optics* and *projective geometry*, which form the basis for modern fields like computer animation. See Darrigol [Dar12] for a recent survey of the history of optics from antiquity to the nineteenth century.

[3] Using Einstein's theory of gravitational lensing.

[4] This leads to many other interesting questions, such as: What types of spatial imagery do animals possess that evolved in the dark, or in murky environments such as muddy water? Some, like bats and dolphins, use echolocation, which in many cases can provide a spatial image since ultrasound waves locally travel in straight lines, too. But what type of spatial imagery do animals possess that live largely by sense of smell, such as ants?

How do we build upon this insight that light nearly always travels in straight lines in order to develop a theory with predictive power? More importantly, how can we develop a theory that can encompass those cases where light does *not* travel in straight lines? And, looking further, can we develop a theory that can be extended to other types of waves in other settings? In the following sections, we selectively discuss some foundational concepts that we will find useful for the rest of the book. This is not a historical survey of the development of ray tracing. The history of optics and wave theory is vast, and the story too complex, for us to do more than touch upon the most relevant highlights to begin stocking our toolkit. Some suggestions for further reading are given along the way for readers who want more detail.

1.1 Fermat's principle of stationary time

1.1.1 General comments

A major unifying theme of this book concerns the power of variational principles. These have a venerable history. In the first century CE, the mathematician and inventor Heron of Alexandria posed, and solved, the following problem in planar geometry: Given a line and two points not on the line, what is the shortest path between the two points that *touches* the line? If the points lie on opposite sides of the line, then the path connecting them crosses the line and the shortest path is simply a straight line. However, if they lie on the same side of the line, then Heron proved (without the use of calculus!) that the shortest path obeys the *law of reflection*. This basic principle of optics was therefore known in antiquity, and it was known to satisfy a minimization principle.

Given that light travels in straight lines, and that rays obey the law of reflection, why invoke a *least-time* principle? Because the path of least time and the path of shortest length are only equivalent if the *speed* of light is constant along the path. Feynman famously pointed out that lifeguards must solve for the least-time path every time they rescue a swimmer: they must determine how far to run along the beach before they dive into the water, where their speed of propagation drops dramatically. If they are good at their job, their path obeys Snell's Law, as we'll discuss.[5]

Let's start with the original form of Fermat's principle of least time, and improve it as we pursue the implications. The least-time principle asserts that of all possible paths light might take from the source to the point of observation, it will take the path that requires the least time. We will adopt the convention that the actual paths light follows are called *rays*, to distinguish them from all the possible paths we

[5] More recently, ants have been found to follow the least-time path as well. See Oettler *et al.* [OSZ+13] .

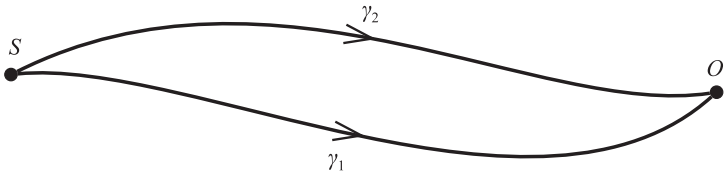

Figure 1.1 The point source S and the point of observation O can be connected by an infinite number of paths, γ. Fermat's principle of least time states that the light *ray* follows that particular path between S and O for which the time taken is minimal.

might imagine. We should emphasize, of course, that the source will in general emit light in many different directions. The principle relates to that particular ray which travels from the source point to the observation point and does not concern itself with the rest, although in principle the observation point is arbitrary. We will return to this important issue in a moment.

Fermat used the least-time principle around 1657 to derive what we now call Snell's Law, which we will discuss in a moment. Here we note that, although Galileo famously described an attempt to measure the speed of light by flashing lanterns about a mile apart in the 1630s, the finiteness of the speed of light was not firmly established until 1676. The astronomer Roemer had detected a regular variation over some years in the timing of the observed eclipses of Io, a moon of Jupiter, relative to the predicted times. The slight advance or retardation of the times for Io to disappear and reappear behind Jupiter depended upon whether Earth and Jupiter were on the same or opposite sides of the Sun. Roemer correctly attributed this apparent change to the finite speed of light, which led to the estimate $c \approx$ 140,000 miles per second (see Hockey *et al.* [Sch07]).

1.1.2 Uniform media

Start with the simplest case, where light has the constant speed c and no obstacles are present. We assume the light is emitted by a point source at S, and the observer is at the point O (see Figure 1.1). The travel time from S to O is

$$T[\gamma] = \frac{L[\gamma]}{c}, \tag{1.1}$$

where $L[\gamma]$ is the length of the path γ from source to observer. It is blindingly obvious[6] to everyone but theoretical physicists and mathematicians that the shortest path between two points is a straight line. But, of course, it is a worthwhile exercise in variational principles to prove it (see Problem 1.1).

[6] Pun intended.

1.1.3 Snell's Law

Thus, the least-time path and the shortest-length path are the same in the simplest case where the light speed is constant. But the least-time principle leads to something new: Suppose S and O lie in two different regions. The source lies in region 1, where the speed of light is c/n_1. The observation point is in region 2, where the speed of light is c/n_2.[7] We leave it as an exercise for the reader to prove that the least-time principle in this case leads to what we now call Snell's Law[8] for the bending of rays at the interface

$$\boxed{n_1 \sin\theta_1 = n_2 \sin\theta_2.} \tag{1.2}$$

An approximate form of Snell's Law had been established by Ptolemy (ca. CE 90–168), in terms of the ratios of the angles. This is only correct for rays that are nearly perpendicular to the interface. It appears that the correct form involving the sines of the angles was discovered by the Persian astronomer-mathematician Ibn Sahl around the year 984 CE (see Hockey *et al.* [Ber07]), though this form was not known in the West until the early seventeenth century, when the invention of the telescope (1608) would have motivated the development of improved theories and measurements of light refraction for lens design.[9] Snell's Law was established empirically, of course, and there are multiple claims to primacy in the literature of the early mid 1600s. What is certain is that Fermat was aware of the law of refraction (1.2), and showed that his principle of least time could be used to derive it.

The combination of straight-line rays within uniform regions, and Snell's Law at the interface between regions, forms the basis for ray optics and lens theory. Kepler was the first to provide a theoretical explanation for the telescope, using ray theory for compound lens systems. This theory first appeared in Kepler's *Dioptrice* (1611).[10]

Observation shows that the refractive index depends upon the color of light. When this effect is included, the theory of prisms emerges and – through the possibility of a double internal reflection within raindrops – the theory of the rainbow.[11] Thus, many of the design principles for microscopes, telescopes,

[7] The constants n_1 and n_2 are the *refractive indices*; θ_1 and θ_2 are the angles formed by rays in the two regions and the local normal at the interface. See Figure 3.4, Section 3.5.

[8] See Problem 1.2.

[9] See Willach [Wil08] for an interesting account of the evolving crafts of glass and lens manufacture in the late Middle Ages and how these contributed to the invention of the telescope. Going even further back, for those interested in the history of lenses in the classical world, a survey of what is known about the use of lenses in the Graeco-Roman world can be found in Plantzos [Pla97]. We thank our colleague Professor Lily Panoussi for bringing this reference to our attention.

[10] This book, written in Latin, has not yet been translated into English, although high-quality scans of the original are available online. It is interesting to note that the first figure to appear in the text concerns a means to measure the refractive properties of materials.

[11] Descartes discussed what is now considered the correct geometrical explanation in his 1637 *Discourse on Method*. Earlier scholars in China and the Middle East also realized the importance of internal reflection for explaining the rainbow by using glass spheres as laboratory models of raindrops.

cameras, eyeglasses, and the explanation of some of the most beautiful of atmospheric phenomena, follow from the principle of least time.

Isaac Newton (1643–1727) made important contributions to the theory of optics and revolutionized the design of telescopes using methods based upon the assumption that light was composed of particles, supplemented by the notion that the speed depended upon their color [New10]. The ray theory of light fits nicely with this hypothesis, and the success of the predictions using a ray theory seemed to confirm the particle hypothesis. (However, Fermat and the least-time principle are not even mentioned in Newton's *Optics*.)

It was not until the work of Thomas Young (1773–1829) and Fresnel (1788–1827) that light was shown convincingly to be a wave phenomenon, capable of diffraction and interference like water waves.[12] This built upon the much earlier work of Huygens (1629–1695).[13] This then leads to a puzzle: if light is a wave, *why does ray theory work so well*? We will see in later sections that Hamilton provided an answer to this question by showing how to construct wave fields, including interference patterns, using an entirely new type of ray theory.

1.1.4 Distributed sources

We should point out that nonpoint sources are dealt with at this stage by simple superposition. Each point on the extended source S' is treated as a point source independent of the others. This is easy to understand if light is composed of particles, but it is a more subtle issue if light is a wave. Use of the superposition assumption leads to the theory of imaging optics.

This simple approach to the analysis of distributed sources is valid only if the light emission is *incoherent* from one point on the source to the next. Speaking imprecisely for the moment, by a *coherent* source we mean one whose rays have a well-defined phase θ at almost all points along each ray, and that this phase function is *smoothly varying* along the ray and between neighboring rays. Coherent wave fields are the primary topic of this book, though we will return to incoherent fields in Section 3.5.5, where we summarize a ray phase space theory for them. We note here that the *incoherence* of visible light in everyday life is almost always a good assumption.[14]

What is lacking in the theory so far, of course, is that we have not discussed how to compute the light intensity. A ray could arrive at the observation point with zero intensity, in which case the existence of the ray itself is largely meaningless.

[12] Young was able to produce coherent light by using a pinhole smaller than the transverse coherence length of sunlight. The coherence length of an extended incoherent source is the wavelength divided by the solid angle of the source [Wol07].

[13] See, for example, Fresnel's essay in [Fre00].

[14] The reader should verify that the previous statement is correct. When, outside of a physics laboratory, do you encounter coherent visible light? What physical conditions are required to produce it?

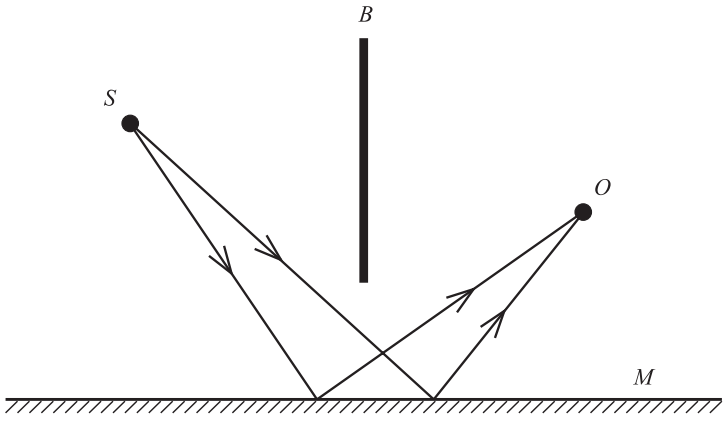

Figure 1.2 The point source S and the point of observation O can no longer be connected by a straight-line path, but the paths must instead go around the obstacle B. A mirror M is present and we consider only paths that reflect from the mirror. The least-time path is the light ray that obeys the law of reflection.

It is an observational fact that the convergence of rays increases the intensity of light. Think of the focusing of sunlight using a lens.[15] Likewise, as rays diverge, the intensity decreases. It is physically reasonable to conjecture that a conservation law applies, for example one modeled on treating the energy density in the light field as a conserved fluid. If light were composed of discrete and long-lived particles, the conservation law would follow directly from particle number conservation. But if light is a wave, the derivation is less obvious. This will be discussed at length later in the book, where we discuss *wave energy* and *wave momentum* (Section 3.3.2), and we derive *action conservation laws* for wave fields, using both coherent (Section 3.1) and incoherent (Section 3.5.5) formulations.

1.1.5 Stationarity vs. minimization: the law of reflection

Return to a point source S and a fixed point of observation O, but now block the direct straight-line ray path and add a plane mirror (see Figure 1.2). This will turn out to be a situation where the naive formulation of the principle of least time will fail us, but it will help guide us to a better formulation. If we consider only paths that pass from S to O *after reflection from* M, the path of least time reflects from M in such a way that the angle of incidence equals the angle of reflection (as measured

[15] The first mention of the use of lenses for burning and cauterizing for medicinal purposes is believed to be Aristophanes, *The Clouds*, first performed ca. 423 BCE. The reference occurs in lines 767–769, see p. 64 of [Ari12]. The use of mirrors to focus light for the same purpose was also understood, as evidenced by the famous legend of Archimedes setting fire to the Roman fleet at Syracuse, ca. 214–212 BCE.

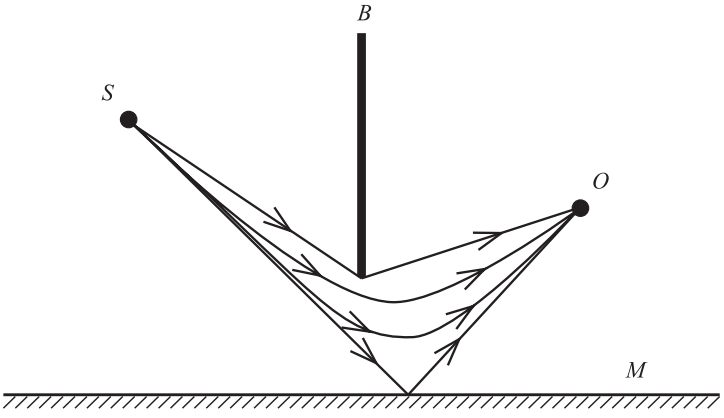

Figure 1.3 The point source S and the point of observation O separated by the obstacle B, but now we consider all paths. The sequence of paths shown that do *not* reflect from the mirror all have travel times less than the ray.

with respect to the local normal at the point of reflection).[16] This is simply Heron's problem, mentioned earlier. With the law of reflection, a host of new applications emerge: optical instruments such as reflecting telescopes (building upon Newton's original breakthrough design) can be analyzed.

But why should we restrict ourselves to paths that reflect from M? Why can't we proceed as before and consider *all* paths that go from S to O? In that case, it should be clear that some of the paths shown in Figure 1.3 have travel times less than the ray that reflects from M. In fact, of the sequence of paths shown, the one with the shortest travel time is the one that travels in a straight line from S to the edge of B, and then on to O. Why don't we use *that* path as our least-time ray and ignore the path that reflects from M?[17]

Suppose we now remove the obstacle. The straight-line path is once again the least-time path, but we also know from experience that the ray that reflects off the mirror will also reach the point O. A helpful way to view what is going on – and one that is quite physical – is to think of the source S as emitting rays that travel in straight lines in all possible directions. If they encounter the mirror M, they bounce and satisfy the law of reflection. Almost all of the infinitude of rays emitted from S will miss the point O. There are *two* that make it to O: the straight-line path and the one that bounces off the mirror and satisfies the law of reflection along the way.

[16] See Problem 1.3.

[17] There is good reason to consider the path that bends around the edge of B as a ray, but it requires careful treatment at the edge where it encounters the obstacle. In fact, the light can *diffract* around the edge if it is sharp enough, so some light could reach O from S by this route. But this takes us beyond a simple ray picture. However, we note that the Huygens–Fresnel theory of wave propagation and diffraction starts with such considerations. See Fresnel [Fre00].

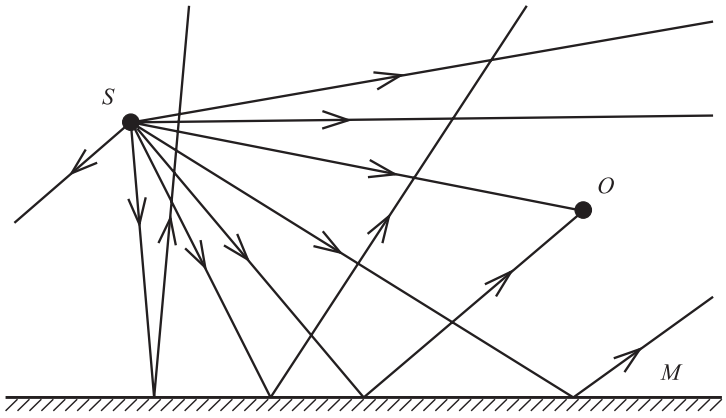

Figure 1.4 The point source S is now shown emitting rays in all directions, and no obstacle lies between S and O. In this case, it is clear that *two* ray paths make it from S to O, and that we should understand Fermat's principle not as a global minimization principle, but as a local *stationarity principle* with respect to neighboring paths.

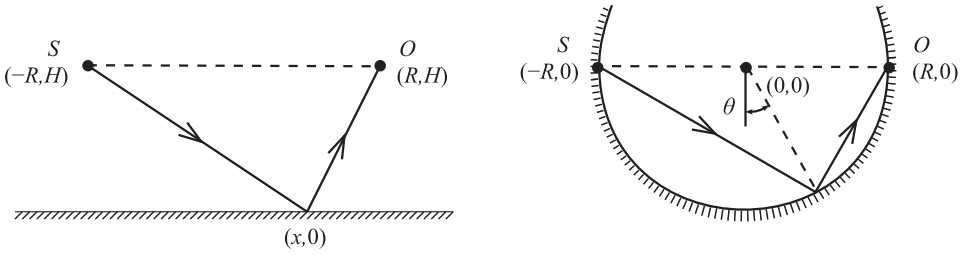

Figure 1.5 The law of reflection from a mirror surface arises by selecting the *shortest* path for the flat mirror (left), but the *longest* for the curved one (right).

The straight-line path is the *global* winner in the race, but the second path is *locally minimal* when compared to neighboring paths *that also reflect off the mirror*.

With these results in mind, a better formulation of Fermat's principle is to define rays as *any path for which the travel time is stationary with respect to small variations*, and accept that there will sometimes be more than one ray that travels from S to O, rather than insisting upon a unique global minimum. For rays encountering mirrors, the variation is carried out only among neighboring paths that also reflect; hence it is an example of a *constrained variation* in the family of paths, not a general variation. Only by using a constrained variation at a mirror do we recover the expected physical result.

As another example which demonstrates that the least-time principle is inadequate, consider the flat and curved mirrors in Figure 1.5. In Problem 1.4 it is shown

that the ray is the shortest path (among those that reflect from the mirror) for the flat mirror, but the *longest* straight-line path that reflects once off the curved mirror shown to the right in the figure. This can be seen by starting with the straight-line path (the ray) from S to O, and then considering a nearby path that reflects once just before it reaches O. This path must have longer flight time than the straight-line path. Now move the reflection point away from O toward S. When the reflection point is in the neighborhood of S, the straight-line path becomes minimal once again. Therefore, the flight time must have reached a maximum in between. These two examples once again emphasize that we should formulate Fermat's principle as a principle of stationary time.

Before leaving the topic of reflection, we should mention that the *specular* reflection we have described here is typical of highly polished surfaces. Rough surfaces (rough on the length scale of a wavelength of light) can lead to more diffuse types of reflection, first characterized by Lambert (1728–1777). This type of reflection is, in fact, more common for everyday surfaces, and diffusion models for reflection by textured surfaces are commonly used in computer graphics.[18]

1.1.6 Smoothly varying media

Thus, Fermat's principle, now properly understood as a *stationarity principle* rather than a minimization principle, unifies many important results that had appeared to be distinct and brings them under one theory. It is natural, then, to extend Fermat's theory to situations where the refractive index varies continuously. Let's consider an important special case of a two-dimensional *layered* medium.

Suppose we have a two-dimensional system that is uniform in the x-direction, but the refractive index varies in y: $n(y)$. Our source lies at $x = 0$, and could be extended in y (for example, it could be a building or a mountain). The point of observation is at x_1 and $y = y_1$. Draw a path, $\gamma = y : x \mapsto y(x)$ from a point on the source to the point of observation

$$\mathbf{r}(x) = [x, y(x)], \qquad \mathbf{r}(0) = (0, y_0), \quad \mathbf{r}(x_1) = (x_1, y_1). \tag{1.3}$$

The time required for light to travel along the path is given by the integral

$$T[y] = \frac{1}{c}\int_0^{x_1} n[y(x)]\sqrt{1 + [y'(x)]^2}\, dx, \qquad y'(x) \equiv \frac{dy}{dx}. \tag{1.4}$$

[18] Rayleigh scattering should also be mentioned because it also involves a breakdown in the simple law of reflection. Rayleigh scattering involves the scattering of light by particles that are smaller than a wavelength. This type of scattering shows a strong frequency dependence and explains why the sky is blue; see for example Jackson [Jac98].

Here, we use the notation $T[y]$ to denote the fact that the elapsed time is a *functional* of the path $y(x)$.[19] Requiring $T[y]$ to be stationary with respect to small variations in the path leads to the Euler–Fermat equation

$$\boxed{\sqrt{1+[y'(x)]^2}\frac{dn}{dy} - \frac{d}{dx}\left(\frac{n\,y'}{\sqrt{1+[y'(x)]^2}}\right) = 0,} \tag{1.5}$$

which can be reorganized as

$$\boxed{y''(x) = \left\{1+[y'(x)]^2\right\}\frac{d\ln n(y)}{dy}.} \tag{1.6}$$

Hence, in a uniform medium ($n' = 0$), we recover the previous result that the light paths are straight lines: $y''(x) = 0$. In a nonuniform medium ($n' \neq 0$), however, the light paths curve toward the region of higher refractive index: concave upward $[y''(x) > 0]$ for $n'(y) > 0$ and concave downward $[y''(x) < 0]$ for $n'(y) < 0$. Further aspects of the continuous case, including a derivation of Snell's Law for a continuous layered medium, the trapping of waves in channels, and mirages are examined in Problems 1.5 through 1.8. In Problem 1.9 the general three-dimensional case is examined.

It is important to emphasize once more that this type of ray theory depends only upon the refractive index $n(\mathbf{x})$, and implicitly assumes the waves are incoherent (due to the lack of any reference to a phase function). There is no *dispersion relation* between the wave frequency and wavevector, because there is no wavevector in Fermat's theory.[20] The "wave equation" was unknown to Fermat and contemporaries. All we need to know to apply Fermat's theory is the wave speed, $v(\mathbf{x}) = c/n(\mathbf{x})$, so we can compute the travel time along any path.

Ray tracing of this sort can also be applied to other types of waves. For example, computer aided tomography (CAT) reconstructs "images" by measuring the attenuation of X-rays along ray paths, while positron emission tomography (PET) scans map the spatial distribution of the intensity of gamma ray emission. In acoustics, ray theory is important for the theory of reverberation. It is used in the design of concert halls and recording studios, and it forms the basis for ultrasound imaging. A striking example of these ideas is the design of *whispering galleries*, where sound rays skim along the gently curved wall of a room.[21]

[19] See Appendix B for a discussion of functionals.
[20] The wavevector is the gradient of the phase $\theta(\mathbf{x})$, as we will discuss in the next section.
[21] In the audible range of frequencies (approximately 20 Hz to 20 kHz), and in typical rooms (a few, to a few tens of meters across) the lower frequencies are not well modeled using rays, but the higher frequencies often can be treated as traveling in straight lines, satisfying the law of reflection. This is because the wavelength of sound waves is $\lambda = c_s/f$ and, with a sound speed of ~ 300 m/s, we have λ ranging from 15 m for the lowest audible frequencies down to 1.5 cm for the highest audible frequencies. The information needed for human speech recognition lies in the mid-range of frequencies.