Contents

Part III

OPTICS

Optics

Version 1207.1.K.pdf, 28 October 2012

Prior to the twentieth century's quantum mechanics and opening of the electromagnetic spectrum observationally, the study of optics was concerned solely with visible light.

Reflection and refraction of light were first described by the Greeks and further studied by medieval scholastics such as Roger Bacon (thirteenth century), who explained the rainbow and used refraction in the design of crude magnifying lenses and spectacles. However, it was not until the seventeenth century that there arose a strong commercial interest in manipulating light, particularly via the telescope and the compound microscope.

The discovery of Snell's law in 1621 and observations of diffractive phenomena by Grimaldi in 1665 stimulated serious speculation about the physical nature of light. The wave and corpuscular theories were propounded by Huygens in 1678 and Newton in 1704, respectively. The corpuscular theory initially held sway, for 100 years. However, observational studies of interference by Young in 1803 and the derivation of a wave equation for electromagnetic disturbances by Maxwell in 1865 then seemed to settle the matter in favor of the undulatory theory, only for the debate to be resurrected in 1887 with the discovery of the photoelectric effect. After quantum mechanics was developed in the 1920's, the dispute was abandoned, the wave and particle descriptions of light became "complementary", and Hamilton's optics-inspired formulation of classical mechanics was modified to produce the Schrödinger equation.

Many physics students are all too familiar with this potted history and may consequently regard optics as an ancient precursor to modern physics, one that has been completely subsumed by quantum mechanics. Not so! Optics has developed dramatically and independently from quantum mechanics in recent decades, and is now a major branch of classical physics. And it is no longer concerned primarily with light. The principles of optics are routinely applied to all types of wave propagation: from all parts of the electromagnetic spectrum, to quantum mechanical waves, e.g. of electrons and neutrinos, to waves in elastic solids (Part IV of this book), fluids (Part V), plasmas (Part VI) and the geometry of spacetime (Part VII). There is a commonality, for instance, to seismology, oceanography and radio physics that allows ideas to be freely transported between these different disciplines. Even in the study of visible light, there have been major developments: the invention of the laser has led to the modern theory of coherence and has begotten the new field of nonlinear optics.

An even greater revolution has occured in optical technology. From the credit card and white light hologram to the laser scanner at a supermarket checkout, from laser printers to CD's, DVD's and BD's, from radio telescopes capable of nanoradian angular resolution

to Fabry-Perot systems that detect displacements smaller than the size of an elementary particle, we are surrounded by sophisticated optical devices in our everyday and scientific lives. Many of these devices turn out to be clever and direct applications of the fundamental optical principles that we shall discuss in Part III of this book.

Our treatment of optics in this Part III differs from that found in traditional texts, in that we shall assume familiarity with basic classical mechanics and quantum mechanics and, consequently, fluency in the language of Fourier transforms. This inversion of the historical development reflects contemporary priorities and allows us to emphasize those aspects of the subject that involve fresh concepts and modern applications.

In Chap. 7, we shall discuss optical (wave-propagation) phenomena in the *geometric optics* approximation. This approximation is accurate whenever the wavelength and the wave period are short compared with the lengthscales and timescales on which the wave amplitude and the waves' environment vary. We shall show how a wave equation can be solved approximately, with optical rays becoming the classical trajectories of quantum particles (photons, phonons, plasmons, gravitons) and the wave field propagating along these trajectories, and how, in general, these trajectories develop singularities or caustics where the geometric optics approximation breaks down, and we must revert to the wave description.

In Chap. 8 we will develop the theory of *diffraction* that arises when the geometric optics approximation fails and the waves' energy spreads in a non-particle-like way. We shall analyze diffraction in two limiting regimes, called *Fresnel* and *Fraunhofer*, after the physicists who discovered them, in which the wavefronts are approximately planar or spherical, respectively. Insofar as we are working with a linear theory of wave propagation, we shall make heavy use of Fourier methods and shall show how elementary applications of Fourier transforms can be used to design powerful optics instruments.

Most elementary diffractive phenomena involve the superposition of an infinite number of waves. However, in many optical applications, only a small number of waves from a common source are combined. This is known as *interference* and is the subject of Chap. 9. In this chapter we will also introduce the notion of coherence, which is a quantitative measure of the distributions of the combining waves and their capacity to interfere constructively.

The final chapter on optics, Chap. 10, is concerned with *nonlinear phenomena* that arise when waves, propagating through a medium, become sufficiently strong to couple to each other. These nonlinear phenomena can occur for all types of waves (we shall meet them for fluid waves in Sec. 16.3 and plasma waves in Chap. 23). For light (the focus of Chap. 10), nonlinearities have become especially important in recent years; the nonlinear effects that arise when laser light is shone through certain crystals are having a strong impact on technology and on fundamental scientific research. We shall explore several examples.

Chapter 7

Geometric Optics

Version 1207.1.K.pdf, 28 October 2012
Please send comments, suggestions, and errata via email to kip@caltech.edu or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

Box 7.1
Reader's Guide

- This chapter does not depend substantially on any previous chapter.
- Secs. 7.1–7.4 of this chapter are foundations for the remaining Optics chapters: 8, 9, and 10.
- The discussion of caustics in Sec. 7.5 is a foundation for Sec. 8.6 on diffraction at a caustic.
- Secs. 7.2 and 7.3 (plane, monochromatic waves and wavepackets in a homogeneous, time-independent medium, the dispersion relation, and the geometric optics equations) will be used extensively in subsequent Parts of this book, including
  - Chap. 12 for elastodynamic waves
  - Chap. 16 for waves in fluids
  - Sec. 19.7 and Chaps. 21–23, for waves in plasmas
  - Chap. 27 for gravitational waves.

7.1 Overview

Geometric optics, the study of "rays," is the oldest approach to optics. It is an accurate description of wave propagation whenever the wavelengths and periods of the waves are far

smaller than the lengthscales and timescales on which the wave amplitude and the medium supporting the waves vary.

After reviewing wave propagation in a homogeneous medium (Sec. 7.2), we shall begin our study of geometric optics in Sec. 7.3. There we shall derive the geometric-optics propagation equations with the aid of the *eikonal* approximation, and we shall elucidate the connection to Hamilton-Jacobi theory, which we will assume the reader has already encountered. This connection will be made more explicit by demonstrating that a classical, geometric-optics wave can be interpreted as a flux of quanta. In Sec. 7.4, we shall specialize the geometric optics formalism to any situation where a bundle of nearly parallel rays is being guided and manipulated by some sort of apparatus. This is called the *paraxial approximation*, and we shall illustrate it with a magnetically focused beam of charged particles and shall show how matrix methods can be used to describe the particle (i.e. ray) trajectories. In Sec. **??**, we shall explore how imperfect optics can produce multiple images of a distant source, and that as one moves from one location to another, the images appear and disappear in pairs. Locations where this happens are called *caustics*, and are governed by *catastrophe theory*, a topic we shall explore briefly. In Sec. 7.6, we shall describe *gravitational lenses*, remarkable astronomical phenomena that illustrate the formation of multiple images, and caustics. Finally, in Sec. 7.7, we shall turn from scalar waves to the vector waves of electromagnetic radiation. We shall deduce the geometric-optics propagation law for the waves' polarization vector and shall explore the classical version of a phenomenon called the *geometric phase*.

7.2 Waves in a Homogeneous Medium

7.2.1 Monochromatic, Plane Waves; Dispersion Relation

Consider a monochromatic plane wave propagating through a homogeneous medium. Independently of the physical nature of the wave, it can be described mathematically by

$$\boxed{\psi = Ae^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \equiv Ae^{i\varphi}}\,, \tag{7.1}$$

where ψ is any oscillatory physical quantity associated with the wave, for example, the y-component of the magnetic field associated with an electromagnetic wave. If, as is usually the case, the physical quantity is real (not complex), then we must take the real part of Eq. (7.1). In Eq. (7.1), A is the wave's *complex amplitude*, $\varphi = \mathbf{k}\cdot\mathbf{x} - \omega t$ is the wave's *phase*, t and $\mathbf{x}$ are time and location in space, $\omega = 2\pi f$ is the wave's *angular frequency*, and $\mathbf{k}$ is its *wave vector* (with $k \equiv |\mathbf{k}|$ its *wave number*, $\lambda = 2\pi/k$ its *wavelength*, $\lambda\!\!\!^{-} = \lambda/2\pi$ its reduced wavelength and $\hat{\mathbf{k}} \equiv \mathbf{k}/k$ its *propagation direction*). Surfaces of constant phase φ are orthogonal to the propagation direction $\hat{\mathbf{k}}$ and move in the $\hat{\mathbf{k}}$ direction with the *phase velocity*

$$\boxed{\mathbf{V}_{\rm ph} \equiv \left(\frac{\partial \mathbf{x}}{\partial t}\right)_{\varphi} = \frac{\omega}{k}\hat{\mathbf{k}}}\,, \tag{7.2}$$

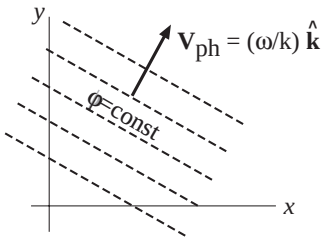


Fig. 7.1: A monochromatic plane wave in a homogeneous medium.

cf. Fig. 7.1. The frequency ω is determined by the wave vector $\mathbf{k}$ in a manner that depends on the wave's physical nature; the functional relationship

$$\boxed{\omega = \Omega(\mathbf{k})} \tag{7.3}$$

is called the wave's *dispersion relation* because (as we shall see in Ex. 7.2) it governs the dispersion (spreading) of a wave packet that is constructed by superposing plane waves.

Some examples of plane waves that we shall study in this book are: ***(i) Electromagnetic waves propagating through an isotropic dielectric medium*** with index of refraction $\mathfrak{n}$ [Eq. 10.20)], for which ψ could be any Cartesian component of the electric or magnetic field or vector potential and the dispersion relation is

$$\omega = \Omega(\mathbf{k}) = Ck \equiv C|\mathbf{k}| \ , \tag{7.4}$$

with $C = c/\mathfrak{n}$ the propagation speed and c the speed of light in vacuum. ***(ii) Sound waves propagating through a solid*** (Sec. 12.2.3) ***or fluid*** (liquid or vapor; Secs. 16.5 and 7.3.1), for which ψ could be the pressure or density perturbation produced by the sound wave (or it could be a potential whose gradient is the velocity perturbation), and the dispersion relation is the same as for electromagnetic waves, Eq. (7.4), but with C now the sound speed. ***(iii) Waves on the surface of a deep body of water*** (depth $\gg \lambda$; Sec. 16.2), for which ψ could be the height of the water above equilibrium, and the dispersion relation is [Eq. (16.9)]:

$$\omega = \Omega(\mathbf{k}) = \sqrt{gk} = \sqrt{g|\mathbf{k}|} \ , \tag{7.5}$$

with g the acceleration of gravity. ***(iv) Flexural waves on a stiff beam or rod*** (Sec. 12.3.4), for which ψ could be the transverse displacement of the beam from equilibrium and the dispersion relation is

$$\omega = \Omega(\mathbf{k}) = \sqrt{\frac{D}{\Lambda}}k^2 = \sqrt{\frac{D}{\Lambda}}\mathbf{k}\cdot\mathbf{k} \ , \tag{7.6}$$

with Λ the rod's mass per unit length and D its "flexural rigidity" [Eq. (12.34)]. ***(v) Alfvén waves in a magnetized, nonrelativistic plasma*** (bending oscillations of the plasma-laden magnetic field lines; Sec. 19.7.2), for which ψ could be the transverse displacement of the field and plasma, and the dispersion relation is [Eq. (19.64)]

$$\omega = \Omega(\mathbf{k}) = \mathbf{a}\cdot\mathbf{k}, \tag{7.7}$$

with $\mathbf{a} = \mathbf{B}/\sqrt{\mu_o \rho}$, $[= \mathbf{B}/\sqrt{4\pi\rho}]$ [1] the *Alfvén speed*, $\mathbf{B}$ the (homogeneous) magnetic field, μ_o the magnetic permitivity of the vacuum, and ρ the plasma mass density.

In general, one can derive the dispersion relation $\omega = \Omega(\mathbf{k})$ by inserting the plane-wave ansatz (7.1) into the dynamical equations that govern one's physical system [e.g. Maxwell's equations, or the equations of elastodynamics (Chap. 12), or the equations for a magnetized plasma (Part VI) or ...]. We shall do so time and again in this book.

7.2.2 Wave Packets

Waves in the real world are not precisely monochromatic and planar. Instead, they occupy wave packets that are somewhat localized in space and time. Such wave packets can be constructed as superpositions of plane waves:

$$\boxed{\psi(\mathbf{x},t) = \int \mathtt{A}(\mathbf{k}) e^{i\alpha(\mathbf{k})} e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} d^3k \text{ , where } \mathtt{A}(\mathbf{k}) \text{ is concentrated around some } \mathbf{k} = \mathbf{k}_o} \, . \tag{7.8a}$$

Here $\mathtt{A}$ and α (both real) are the modulus and phase of the complex amplitude $\mathtt{A}e^{i\alpha}$, and the integration element is $d^3k \equiv d\mathcal{V}_k \equiv dk_x dk_y dk_z$ in terms of components of $\mathbf{k}$ along Cartesian axes x, y, z. In the integral (7.8a), the contributions from adjacent $\mathbf{k}$'s will tend to cancel each other except in that region of space and time where the oscillatory phase factor changes little with changing $\mathbf{k}$, i.e. when $\mathbf{k}$ is near $\mathbf{k}_o$. This is the spacetime region in which the wave packet is concentrated, and its center is where $\boldsymbol{\nabla}_{\mathbf{k}}(\text{phasefactor}) = 0$:

$$\left(\frac{\partial \alpha}{\partial k_j} + \frac{\partial}{\partial k_j} (\mathbf{k}\cdot\mathbf{x} - \omega t) \right)_{\mathbf{k}=\mathbf{k}_o} = 0 \, . \tag{7.8b}$$

Evaluating the derivative with the aid of the wave's dispersion relation $\omega = \Omega(\mathbf{k})$, we obtain for the location of the wave packet's center

$$x_j - \left(\frac{\partial \Omega}{\partial k_j} \right)_{\mathbf{k}=\mathbf{k}_o} t = - \left(\frac{\partial \alpha}{\partial k_j} \right)_{\mathbf{k}=\mathbf{k}_o} = \text{const} \, . \tag{7.8c}$$

This tells us that the wave packet moves with the *group velocity*

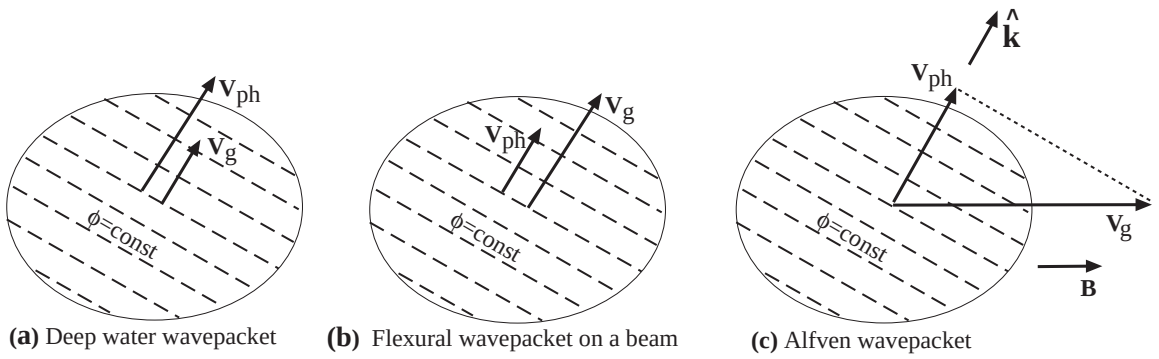
$$\boxed{\mathbf{V}_g = \nabla_{\mathbf{k}} \Omega \, , \quad \text{i.e.} \quad \mathbf{V}_{g\,j} = \left(\frac{\partial \Omega}{\partial k_j} \right)_{\mathbf{k}=\mathbf{k}_o}} \, . \tag{7.9}$$

When, as for electromagnetic waves in a dielectric medium or sound waves in a solid or fluid, the dispersion relation has the simple form (7.4), $\omega = \Omega(\mathbf{k}) = Ck$ with $k \equiv |\mathbf{k}|$, then the group and phase velocities are the same

$$\mathbf{V}_g = \mathbf{V}_{\rm ph} = C\hat{\mathbf{k}} \, , \tag{7.10}$$

and the waves are said to be *dispersionless*. If the dispersion relation has any other form, then the group and phase velocities are different, and the wave is said to exhibit *dispersion*;

[1] Gaussian unit equivalents will be given with square brackets.



(a) Deep water wavepacket **(b)** Flexural wavepacket on a beam **(c)** Alfven wavepacket

Fig. 7.2: (a) A wave packet of waves on a deep body of water. The packet is localized in the spatial region bounded by the thin ellipse. The packet's (ellipse's) center moves with the group velocity $\mathbf{V}_g$. The ellipse expands slowly due to wave-packet dispersion (spreading; Ex. 7.2). The surfaces of constant phase (the wave's oscillations) move twice as fast as the ellipse and in the same direction, $\mathbf{V}_{\rm ph} = 2\mathbf{V}_g$ [Eq. (7.11)]. This means that the wave's oscillations arise at the back of the packet and move forward through the packet, disappearing at the front. The wavelength of these oscillations is $\lambda = 2\pi/k_o$, where $k_o = |\mathbf{k}_o|$ is the wavenumber about which the wave packet is concentrated [Eq. (7.8a) and associated discussion]. (b) A flexural wave packet on a beam, for which $\mathbf{V}_{\rm ph} = \frac{1}{2}\mathbf{V}_g$ [Eq. (7.12)] so the wave's oscillations arise at the packet's front and, traveling more slowly than the packet, disappear at its back. (c) An Alfvén wave packet. Its center moves with a group velocity $\mathbf{V}_g$ that points along the direction of the background magnetic field [Eq. (7.13)], and its surfaces of constant phase (the wave's oscillations) move with a phase velocity $\mathbf{V}_{\rm ph}$ that can be in any direction $\hat{\mathbf{k}}$. The phase speed is the projection of the group velocity onto the phase propagation direction, $|\mathbf{V}_{\rm ph}| = \mathbf{V}_g \cdot \hat{\mathbf{k}}$ [Eq. (7.13)], which implies that the wave's oscillations remain fixed inside the packet as the packet moves; their pattern inside the ellipse does not change.

cf. Ex. 7.2. Examples are (see Fig. 7.2 and the text above, from which our numbering is taken): ***(iii) Waves on a deep body of water*** [dispersion relation (7.5); Fig. 7.2a], for which

$$\mathbf{V}_g = \frac{1}{2}\mathbf{V}_{\rm ph} = \frac{1}{2}\sqrt{\frac{g}{k}}\,\hat{\mathbf{k}}\,. \tag{7.11}$$

(iv) Flexural waves on a beam or rod [dispersion relation (7.6); Fig. 7.2b], for which

$$\mathbf{V}_g = 2\mathbf{V}_{\rm ph} = 2\sqrt{\frac{D}{\Lambda}}\,k\hat{\mathbf{k}}\,. \tag{7.12}$$

(v) Alfvén waves in a magnetized plasma [dispersion relation (7.7); Fig. 7.2c], for which

$$\mathbf{V}_g = \mathbf{a}\,, \quad \mathbf{V}_{\rm ph} = (\mathbf{a}\cdot\hat{\mathbf{k}})\hat{\mathbf{k}}\,. \tag{7.13}$$

Notice that, depending on the dispersion relation, the group speed $|\mathbf{V}_g|$ can be less than or greater than the phase speed, and if the homogeneous medium is anisotropic (e.g., for a magnetized plasma), the group velocity can point in a different direction than the phase velocity.

It should be obvious, physically, that the energy contained in a wave packet must remain always with the packet and cannot move into the region outside the packet where the wave

amplitude vanishes. Correspondingly, *the wave packet's energy must propagate with the group velocity* $\mathbf{V}_g$ *and not with the phase velocity* $\mathbf{V}_{\rm ph}$. Correspondingly, when one examines the wave packet from a quantum mechanical viewpoint, *its quanta must move with the group velocity* $\mathbf{V}_g$. Since we have required that the wave packet have its wave vectors concentrated around $\mathbf{k}_o$, the energy and momentum of each of the packet's quanta are given by the standard quantum mechanical relations

$$\boxed{\mathcal{E} = \hbar\Omega(\mathbf{k}_o) \text{ and } \mathbf{p} = \hbar\mathbf{k}_o} \,. \tag{7.14}$$

****

EXERCISES

Exercise 7.1 *Practice: Group and Phase Velocities*

Derive the group and phase velocities (7.10)–(7.13) from the dispersion relations (7.4)–(7.7).

Exercise 7.2 ***Example: Gaussian Wave Packet and its Dispersion*

Consider a one-dimensional wave packet, $\psi(x,t) = \int \mathtt{A}(k) e^{i\alpha(k)} e^{i(kx-\omega t)} dk$ with dispersion relation $\omega = \Omega(k)$. For concreteness, let $\mathtt{A}(k)$ be a narrow Gaussian peaked around k_o: $\mathtt{A} \propto \exp[-\kappa^2/2(\Delta k)^2]$, where $\kappa = k - k_o$.

(a) Expand α as $\alpha(k) = \alpha_o - x_o\kappa$ with x_o a constant, and assume for simplicity that higher order terms are negligible. Similarly expand $\omega \equiv \Omega(k)$ to quadratic order, and explain why the coefficients are related to the group velocity V_g at $k = k_o$ by $\Omega = \omega_o + V_g\kappa + (dV_g/dk)\kappa^2/2$.

(b) Show that the wave packet is given by

$$\psi \propto \exp[i(\alpha_o + k_o x - \omega_o t)] \int_{-\infty}^{+\infty} \exp[i\kappa(x - x_o - V_g t)] \exp\left[-\frac{\kappa^2}{2}\left(\frac{1}{(\Delta k)^2} + i\frac{dV_g}{dk}t\right)\right] d\kappa \,. \tag{7.15a}$$

The term in front of the integral describes the phase evolution of the waves inside the packet; cf. Fig. 7.2.

(c) Evaluate the integral analytically (with the help of a computer, if you wish). Show, from your answer, that the modulus of ψ is given by

$$\boxed{|\psi| \propto \exp\left[-\frac{(x - x_o - V_g t)^2}{2L^2}\right] \,, \quad \text{where } L = \frac{1}{2\Delta k}\sqrt{1 + \left(\frac{dV_g}{dk}(\Delta k)^2\, t\right)^2}} \tag{7.15b}$$

is the packet's half width.

(d) Discuss the relationship of this result, at time $t = 0$, to the uncertainty principle for the localization of the packet's quanta.

(e) Equation (7.15b) shows that the wave packet spreads (i.e. disperses) due to its containing a range of group velocities. How long does it take for the packet to enlarge by a factor 2? For what range of initial half widths can a water wave on the ocean spread by less than a factor 2 while traveling from Hawaii to California?

****

7.3 Waves in an Inhomogeneous, Time-Varying Medium: The Eikonal Approximation and Geometric Optics

Suppose that the medium in which the waves propagate is spatially inhomogeneous and varies with time. If the lengthscale $\mathcal{L}$ and timescale $\mathcal{T}$ for substantial variations are long compared to the waves' reduced wavelength and period,

$$\mathcal{L} \gg \lambda\!\!\!^{-} = 1/k \ , \quad \mathcal{T} \gg 1/\omega \ , \tag{7.16}$$

then the waves can be regarded locally as planar and monochromatic. The medium's inhomogeneities and time variations may produce variations in the wave vector $\mathbf{k}$ and frequency ω, but those variations should be substantial only on scales $\gtrsim \mathcal{L} \gg 1/k$ and $\gtrsim \mathcal{T} \gg 1/\omega$. This intuitively obvious fact can be proved rigorously using a two-lengthscale expansion, i.e. an expansion of the wave equation in powers of $\lambda\!\!\!^{-}/\mathcal{L} = 1/k\mathcal{L}$ and $1/\omega\mathcal{T}$. Such an expansion, in this context of wave propagation, is called the *geometric optics approximation* or the *eikonal approximation* (after the Greek word $\epsilon\iota\kappa\omega\nu$ meaning image). When the waves are those of elementary quantum mechanics, it is called the *WKB approximation*. The eikonal approximation converts the laws of wave propagation into a remarkably simple form in which the waves' amplitude is transported along trajectories in spacetime called *rays*. In the language of quantum mechanics, these rays are the world lines of the wave's quanta (photons for light, phonons for sound, plasmons for Alfvén waves, gravitons for gravitational waves), and the law by which the wave amplitude is transported along the rays is one which conserves quanta. These ray-based propagation laws are called the laws of *geometric optics*.

In this section we shall develop and study the eikonal approximation and its resulting laws of geometric optics. We shall begin in Sec. 7.3.1 with a full development of the eikonal approximation and its geometric-optics consequences for a prototypical dispersion-free wave equation that represents, for example, sound waves in a weakly inhomogeneous fluid. In Sec. 7.3.3 we shall extend our analysis to cover all other types of waves. In Sec. 7.3.4 and a number of exercises we shall explore examples of geometric-optics waves, and in Sec. 7.3.5 we shall discuss conditions under which the eikonal approximation breaks down, and some non-geometric-optics phenomena that result from the breakdown. Finally, in Sec. 7.3.6 we shall return to nondispersive light and sound waves, and deduce Fermat's principle and explore some of its consequences.

7.3.1 Geometric Optics for a Prototypical Wave Equation

Our prototypical wave equation is

$$\frac{\partial}{\partial t}\left(W\frac{\partial\psi}{\partial t}\right) - \boldsymbol{\nabla}\cdot(WC^2\boldsymbol{\nabla}\psi) = 0\ . \tag{7.17}$$

Here $\psi(\mathbf{x},t)$ is the quantity that oscillates (the *wave field*), $C(\mathbf{x},t)$ will turn out to be the wave's slowly varying *propagation speed*, and $W(\mathbf{x},t)$ is a slowly varying *weighting function* that depends on the properties of the medium through which the wave propagates. As we shall see, W has no influence on the wave's dispersion relation or on its geometric-optics rays, but does influence the law of transport for the waves' amplitude.

The wave equation (7.17) describes sound waves propagating through a static, inhomogeneous fluid (Ex. 16.12), in which case ψ is the wave's pressure perturbation δP, $C(\mathbf{x}) = \sqrt{(\partial P/\partial\rho)_s}$ is the adabiatic sound speed, and the weighting function is $W(\mathbf{x}) = \rho/C^2$, with ρ the fluid's unperturbed density. This wave equation also describes waves on the surface of a lake or pond or the ocean, in the limit that the slowly varying depth of the undisturbed water $h_o(\mathbf{x})$ is small compared to the wavelength (shallow-water waves; e.g. tsunamis); see Ex. 16.2. In this case $W = 1$ and $C = \sqrt{gh_o}$ with g the acceleration of gravity. In both cases, sound-waves in a fluid and shallow-water waves, if we turn on a slow time dependence in C and W, then additional terms enter the wave equation (7.17). For pedagogical simplicity we leave those terms out, but in the analysis below we do allow W and C to be slowly varying in time, as well as in space: $W = W(\mathbf{x},t)$ and $C = C(\mathbf{x},t)$.

Associated with the wave equation (7.17) are an energy density $U(\mathbf{x},t)$ and energy flux $\mathbf{F}(\mathbf{x},t)$ given by

$$U = W\left[\frac{1}{2}\dot{\psi}^2 + \frac{1}{2}C^2(\boldsymbol{\nabla}\psi)^2\right]\ ,\quad \mathbf{F} = -WC^2\dot{\psi}\boldsymbol{\nabla}\psi\ ; \tag{7.18}$$

see Ex. 7.4. Here and below the dot denotes a time derivative, $\dot{\psi} \equiv \psi_{,t} \equiv \partial\psi/\partial t$. It is straightforward to verify that, *if C and W are independent of time t*, then the scalar wave equation (7.17) guarantees that the U and $\mathbf{F}$ of Eq. (7.18) satisfy the law of energy conservation

$$\frac{\partial U}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{F} = 0\ ; \tag{7.19}$$

cf. Ex. 7.4.

We now specialize to a weakly inhomogeneous and slowly time-varying fluid and to *nearly* plane waves, and we seek a solution of the wave equation (7.17) that *locally* has approximately the plane-wave form $\psi \simeq Ae^{i\mathbf{k}\cdot\mathbf{x}-\omega t}$. Motivated by this plane-wave form, (i) we express the waves as the product of a real amplitude $A(\mathbf{x},t)$ that varies slowly on the length and time scales $\mathcal{L}$ and $\mathcal{T}$, and the exponential of a complex phase $\varphi(\mathbf{x},t)$ that varies rapidly on the timescale $1/\omega$ and lengthscale λ:

$$\psi(\mathbf{x},t) = A(\mathbf{x},t)e^{i\varphi(\mathbf{x},t)}\ ; \tag{7.20}$$

and (ii) we *define* the wave vector (field) and angular frequency (field) by

$$\mathbf{k}(\mathbf{x},t) \equiv \boldsymbol{\nabla}\varphi\ ,\quad \omega(\mathbf{x},t) \equiv -\partial\varphi/\partial t\ . \tag{7.21}$$

Box 7.2
Bookkeeping Parameter in Two-Lengthscale Expansions

When developing a two-lengthscale expansion, it is sometimes helpful to introduce a "bookkeeping" parameter σ and rewrite the anszatz (7.20) in a fleshed-out form

$$\boxed{\psi = (A + \sigma B + \ldots)e^{i\varphi/\sigma}} \,. \tag{1}$$

The numerical value of σ is unity so it can be dropped when the analysis is finished. We use σ to tell us how various terms scale when $\lambda\!\!\!^{-}$ is reduced at fixed $\mathcal{L}$ and $\mathcal{R}$: A has no attached σ and so scales as $\lambda\!\!\!^{-0}$, B is multiplied by σ and so scales proportional to $\lambda\!\!\!^{-}$, and φ is multiplied by σ^{-1} and so scales as $\lambda\!\!\!^{-1}$. When one uses these σ's in the evaluation of the wave equation, the first term on the second line of Eq. (7.22) gets multipled by σ^{-2}, the second term by σ^{-1}, and the omitted terms by σ^0. These factors of σ help us in quickly grouping together all terms that scale in a similar manner, and identifying which of the groupings is leading order, and which subleading, in the two-lengthscale expansion. In Eq. (7.22) the omitted σ^0 terms are the first ones in which B appears; they produce a propagation law for B, which can be regarded as a *post-geometric-optics correction*.

Occasionally (e.g. Ex. 7.9) the wave equation itself will contain terms that scale with $\lambda\!\!\!^{-}$ differently from each other, and one should always look out for this possibility

In addition to our two-lengthscale requirement $\mathcal{L} \gg 1/k$ and $\mathcal{T} \gg 1/\omega$, we also require that A, $\mathbf{k}$ and ω vary slowly, i.e., vary on lengthscales $\mathcal{R}$ and timescales $\mathcal{T}'$ long compared to $\lambda\!\!\!^{-} = 1/k$ and $1/\omega$.[2] This requirement guarantees that the waves are locally planar, $\varphi \simeq \mathbf{k}\cdot x - \omega t + \text{constant}$.

We now insert the Eikonal-approximated wave field (7.20) into the wave equation (7.17), perform the differentiations with the aid of Eqs. (7.21), and collect terms in a manner dictated by a two-lengthscale expansion (see Box 7.2):

$$\begin{aligned} 0 &= \frac{\partial}{\partial t}\left(W\frac{\partial\psi}{\partial t}\right) - \boldsymbol{\nabla}\cdot(WC^2\boldsymbol{\nabla}\psi) \qquad (7.22)\\ &= \left(-\omega^2 + C^2k^2\right)W\psi + \left[-2(\omega\dot{A} + C^2k_jA_{,j})W - (W\omega)_{,t}A - (WC^2k_j)_{,j}A\right]e^{i\varphi} + \ldots \,. \end{aligned}$$

The first term on the second line, $(-\omega^2 + C^2k^2)W\psi$ scales as $\lambda\!\!\!^{-2}$ when we make the reduced wavelength $\lambda\!\!\!^{-}$ shorter and shorter while holding the macroscopic lengthscales $\mathcal{L}$ and $\mathcal{R}$ fixed; the second term (in square brackets) scales as $\lambda\!\!\!^{-1}$; and the omitted terms scale as $\lambda\!\!\!^{-0}$. This is what we mean by "collecting terms in a manner dictated by a two-lengthscale expansion". Because of their different scaling, the first, second, and omitted terms must vanish separately; they cannot possibly cancel each other.

[2]Note: these variations can arise both (i) from the influence of the medium's inhomogeneity (which puts limits $\mathcal{R} \lesssim \mathcal{L}$ and $\mathcal{T}' \lesssim \mathcal{T}$ on the wave's variations, and also (ii) from the chosen form of the wave. For example, the wave might be traveling outward from a source and so have nearly spherical phase fronts with radii of curvature $r \simeq$ (distance from source); then $\mathcal{R} = \min(r, \mathcal{L})$.

The vanishing of the first term in the eikonal-approximated wave equation (7.22) says that the waves' frequency field $\omega(\mathbf{x},t) \equiv -\partial\varphi/\partial t$ and wave-vector field $\mathbf{k} \equiv \boldsymbol{\nabla}\varphi$ satisfy the dispersionless dispersion relation

$$\omega = \Omega(\mathbf{k},\mathbf{x},t) \equiv C(\mathbf{x},t)k \ , \tag{7.23}$$

where (as throughout this chapter) $k \equiv |\mathbf{k}|$. Notice that, as promised, this dispersion relation is independent of the weighting function W in the wave equation. Notice further that *this dispersion relation is identical to that for a precisely plane wave in a homogeneous medium, Eq. (7.4), except that the propagation speed C is now a slowly varying function of space and time.* This will always be so:

One can always deduce the geometric-optics dispersion relation by (i) considering a precisely plane, monochromatic wave in a precisely homogeneous, time-independent medium and deducing $\omega = \Omega(\mathbf{k})$ in a functional form that involves the medium's properties (e.g. density); and then (ii) allowing the properties to be slowly varying functions of $\mathbf{x}$ and t. The resulting dispersion relation, e.g. Eq. (7.23), then acquires its $\mathbf{x}$ and t dependence from the properties of the medium.

The vanishing of the second term in the eikonal-approximated wave equation (7.22) says that the waves' real amplitude A is transported with the group velocity $\mathbf{V}_g = C\hat{\mathbf{k}}$ in the following manner:

$$\frac{dA}{dt} \equiv \left(\frac{\partial}{\partial t} + \mathbf{V}_g \cdot \boldsymbol{\nabla}\right) A = -\frac{1}{2W\omega}\left[\frac{\partial(W\omega)}{\partial t} + \boldsymbol{\nabla}\cdot(WC^2\mathbf{k})\right] A \ . \tag{7.24}$$

This propagation law, by contrast with the dispersion relation, does depend on the weighting function W. We shall return to this propagation law shortly and shall understand more deeply its dependence on W, but first we must investigate in detail the directions along which A is transported.

The time derivative $d/dt = \partial/\partial t + \mathbf{V}_g \cdot \boldsymbol{\nabla}$ appearing in the propagation law (7.24) is similar to the derivative with respect to proper time along a world line in special relativity, $d/d\tau = u^0\partial/\partial t + \mathbf{u}\cdot\boldsymbol{\nabla}$ (with u^α the world line's 4-velocity). This analogy tells us that the waves' amplitude A is being propagated along some sort of world lines (trajectories). Those world lines (called the waves' *rays*), in fact, are governed by Hamilton's equations of particle mechanics with the dispersion relation $\Omega(\mathbf{x},t,\mathbf{k})$ playing the role of the Hamiltonian and $\mathbf{k}$ playing the role of momentum:

$$\boxed{\frac{dx_j}{dt} = \left(\frac{\partial\Omega}{\partial k_j}\right)_{\mathbf{x},t} \equiv V_{g\,j}} \ , \quad \boxed{\frac{dk_j}{dt} = -\left(\frac{\partial\Omega}{\partial x_j}\right)_{\mathbf{k},t}} \ , \quad \boxed{\frac{d\omega}{dt} = \left(\frac{\partial\Omega}{\partial t}\right)_{\mathbf{x},\mathbf{k}}} \ . \tag{7.25}$$

The first of these Hamilton equations is just our definition of the group velocity, with which [according to Eq. (7.24)] the amplitude is transported. The second tells us how the wave vector $\mathbf{k}$ changes along a ray, and together with our knowledge of $C(\mathbf{x},t)$, it tells us how the group velocity $\mathbf{V}_g = C\hat{\mathbf{k}}$ for our dispersionless waves changes along a ray, and thence tells us the ray itself. The third tells us how the waves' frequency changes along a ray.

To deduce the second and third of these Hamilton equations, we begin by inserting the definitions $\omega = -\partial\varphi/\partial t$ and $\mathbf{k} = \boldsymbol{\nabla}\varphi$ [Eqs. (7.21)] into the dispersion relation $\omega = \Omega(\mathbf{x}, t; \mathbf{k})$ for an arbitrary wave, thereby obtaining

$$\boxed{\frac{\partial\varphi}{\partial t} + \Omega(\mathbf{x}, t; \boldsymbol{\nabla}\varphi) = 0} \,. \tag{7.26a}$$

This equation is known in optics as the *eikonal equation*. It is formally the same as the Hamilton-Jacobi equation of classical mechanics[3] if we identify Ω with the Hamiltonian and φ with Hamilton's principal function; cf. Ex. 7.9. This suggests that, to derive the second and third of Eqs. (7.25), we can follow the same procedure as is used to derive Hamilton's equations of motion: We take the gradient of Eq. (7.26a) to obtain

$$\frac{\partial^2\varphi}{\partial t \partial x_j} + \frac{\partial\Omega}{\partial k_l}\frac{\partial^2\varphi}{\partial x_l \partial x_j} + \frac{\partial\Omega}{\partial x_j} = 0 \,, \tag{7.26b}$$

where the partial derivatives of Ω are with respect to its arguments $(\mathbf{x}, t; \mathbf{k})$; we then use $\partial\varphi/\partial x_j = k_j$ and $\partial\Omega/\partial k_l = V_{gl}$ to write this as $dk_j/dt = -\partial\Omega/\partial x_j$. This is the second of Hamilton's equations (7.25), and it tells us how the wave vector changes along a ray. The third Hamilton equation, $d\omega/dt = \partial\Omega/\partial t$ [Eq. (7.25)] is obtained by taking the time derivative of the eikonal equation (7.26a).

Not only is the waves' amplitude A propagated along the rays, so is their phase:

$$\frac{d\varphi}{dt} = \frac{\partial\varphi}{\partial t} + \mathbf{V}_g \cdot \boldsymbol{\nabla}\varphi = -\omega + \mathbf{V}_g \cdot \mathbf{k} \,. \tag{7.27}$$

Since our dispersionless waves have $\omega = Ck$ and $\mathbf{V}_g = C\hat{\mathbf{k}}$, this vanishes. Therefore, *for the special case of dispersionless waves (e.g., sound waves in a fluid and electromagnetic waves in an isotropic dielectric medium), the phase is constant along each ray*

$$\boxed{d\varphi/dt = 0} \,. \tag{7.28}$$

7.3.2 Connection of Geometric Optics to Quantum Theory

Although the waves $\psi = Ae^{i\varphi}$ are classical and our analysis is classical, their propagation laws in the eikonal approximation can be described most nicely in quantum mechanical language.[4] Quantum mechanics insists that, associated with any wave in the geometric optics regime, there are real quanta: the wave's quantum mechanical particles. If the wave is electromagnetic, the quanta are photons; if it is gravitational, they are gravitons; if it is sound, they are phonons; if it is a plasma wave (e.g. Alfvén), they are plasmons. When we multiply the wave's $\mathbf{k}$ and ω by Planck's constant, we obtain the particles' momentum and energy,

$$\boxed{\mathbf{p} = \hbar\mathbf{k} \,, \quad \mathcal{E} = \hbar\omega} \,. \tag{7.29}$$

[3] See, for example, Goldstein, Safko and Poole (2002).

[4] This is intimately related to the fact that quantum mechanics underlies classical mechanics; the classical world is an approximation to the quantum world.

Although the originators of the 19th century theory of classical waves were unaware of these quanta, once quantum mechanics had been formulated, the quanta became a powerful conceptual tool for thinking about classical waves.

In particular, we can regard the rays as the world lines of the quanta, and by multiplying the dispersion relation by $\hbar$ we can obtain the Hamiltonian for the quanta's world lines

$$\boxed{H(\mathbf{x},t;\mathbf{p}) = \hbar\Omega(\mathbf{x},t;\mathbf{k}=\mathbf{p}/\hbar)} \,. \tag{7.30}$$

Hamilton's equations (7.25) for the rays then become, immediately, Hamilton's equations for the quanta, $dx_j/dt = \partial H/\partial p_j$, $dp_j/dt = -\partial H/\partial x_j$, $d\mathcal{E}/dt = \partial H/\partial t$.

Return, now, to the propagation law (7.24) for the waves' amplitude, and examine its consequences for the waves' energy. By inserting the ansatz $\psi = \Re(Ae^{i\varphi}) = A\cos(\varphi)$ into Eqs. (7.18) for the energy density U and energy flux $\mathbf{F}$, and averaging over a wavelength and wave period so $\overline{\cos^2\varphi} = \overline{\sin^2\varphi} = 1/2$, we find that

$$U = \frac{1}{2}WC^2k^2A^2 = \frac{1}{2}W\omega^2A^2 \;, \quad \mathbf{F} = U(C\hat{\mathbf{k}}) = U\mathbf{V}_g \;. \tag{7.31}$$

Inserting these into the expression $\partial U/\partial t + \boldsymbol{\nabla}\cdot\mathbf{F}$ for the rate at which energy (per unit volume) fails to be conserved, and using the propagation law (7.24) for A, we obtain

$$\frac{\partial U}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{F} = U\frac{\partial \ln C}{\partial t} \;. \tag{7.32}$$

Thus, as the propagation speed C slowly changes at fixed location in space due to a slow change in the medium's properties, the medium slowly pumps energy into the waves or removes it from them at a rate per unit volume $U\partial\ln C/\partial t$.

This slow energy change can be understood more deeply using quantum concepts. The number density and number flux of quanta are

$$\boxed{n = \frac{U}{\hbar\omega} \;, \quad \mathbf{S} = \frac{\mathbf{F}}{\hbar\omega} = n\mathbf{V}_g} \,. \tag{7.33}$$

By combining these with the the energy (non)conservation equation (7.32), we obtain

$$\frac{\partial n}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{S} = n\left[\frac{\partial \ln C}{\partial t} - \frac{d\ln\omega}{dt}\right] \;.$$

The third Hamilton equation tells us that $d\omega/dt = \partial\Omega/\partial t = \partial(Ck)/\partial t = k\partial C/\partial t$, whence $d\ln\omega/dt = \partial\ln C/\partial t$, which, when inserted into the above equation, implies that *the quanta are conserved:*

$$\boxed{\frac{\partial n}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{S} = 0} \,. \tag{7.34a}$$

Since $\mathbf{S} = n\mathbf{V}_g$ and $d/dt = \partial/\partial t + \mathbf{V}_g\cdot\boldsymbol{\nabla}$, we can rewrite this conservation law as a propagation law for the number density of quanta:

$$\boxed{\frac{dn}{dt} + n\boldsymbol{\nabla}\cdot\mathbf{V}_g = 0} \,. \tag{7.34b}$$

The propagation law for the waves' amplitude, Eq. (7.24), can now be understood much more deeply: *The amplitude propagation law is nothing but the law of conservation of quanta in a slowly varying medium, rewritten in terms of the amplitude. This is true quite generally, for any kind of wave (Sec. 7.3.3); and the quickest route to the amplitude propagation law is often to express the wave's energy density U in terms of the amplitude and then invoke conservation of quanta*, Eq. (7.34b).

In Ex. 7.3 we shall show that the conservation law (7.34b) is equivalent to

$$\boxed{\frac{d(nC\mathcal{A})}{dt} = 0\ , \quad \text{i.e.,} \quad nC\mathcal{A} \text{ is a constant along each ray}}\ . \tag{7.34c}$$

Here $\mathcal{A}$ is the cross sectional area of a bundle of rays surrounding the ray along which the wave is propagating. Equivalently, by virtue of Eqs. (7.33) and (7.31) for the number density of quanta in terms of the wave amplitude A,

$$\frac{d}{dt} A\sqrt{CW\omega\mathcal{A}} = 0 \quad \text{i.e.,} \quad A\sqrt{CW\omega\mathcal{A}} \text{ is a constant along each ray}\ . \tag{7.34d}$$

In the above Eqs. (7.33)–(7.34), we have boxed those equations that are completely general (because they embody conservation of quanta) and have not boxed those that are specialized to our prototypical wave equation.

****

EXERCISES

Exercise 7.3 *** Derivation and Example: Amplitude Propagation for Dispersionless Waves Expressed as Constancy of Something Along a Ray*

(a) In connection with Eq. (7.34b), explain why $\boldsymbol{\nabla} \cdot \mathbf{V}_g = d\ln\mathcal{V}/dt$, where $\mathcal{V}$ is the tiny volume occupied by a collection of the wave's quanta.

(b) Choose for the collection of quanta those that occupy a cross sectional area $\mathcal{A}$ orthogonal to a chosen ray, and a longitudinal length Δs along the ray, so $\mathcal{V} = \mathcal{A}\Delta s$. Show that $d\ln\Delta s/dt = d\ln C/dt$ and correspondingly, $d\ln\mathcal{V}/dt = d\ln(C\mathcal{A})/dt$.

(c) Thence, show that the conservation law (7.34b) is equivalent to the constancy of $nC\mathcal{A}$ along a ray, Eq. (7.34c).

(d) From this, derive the constancy of $A\sqrt{CW\omega\mathcal{A}}$ along a ray (where A is the wave's amplitude), Eq. (7.34d).

Exercise 7.4 ***** ***T2*** *Example: Energy Density and Flux, and Adiabatic Invariant, for a Dispersionless Wave*

(a) Show that the prototypical scalar wave equation (7.17) follows from the variational principle

$$\delta \int \mathcal{L} dt d^3x = 0 \ , \tag{7.35a}$$

where $\mathcal{L}$ is the Lagrangian density

$$\mathcal{L} = W \left[\frac{1}{2} \left(\frac{\partial \psi}{\partial t} \right)^2 - \frac{1}{2} C^2 \left(\boldsymbol{\nabla} \psi \right)^2 \right] \ . \tag{7.35b}$$

(not to be confused with the lengthscale $\mathcal{L}$ of inhomogeneities in the medium).

(b) For any scalar-field Lagrangian $\mathcal{L}(\dot{\psi}, \boldsymbol{\nabla}\psi, \mathbf{x}, t)$, there is a *canonical*, relativistic procedure for constructing a stress-energy tensor:

$$T_\mu{}^\nu = -\frac{\partial \mathcal{L}}{\partial \psi_{,\nu}} \psi_{,\mu} + \delta_\mu{}^\nu \mathcal{L} \ . \tag{7.35c}$$

Show that, if $\mathcal{L}$ has no explicit time dependence (e.g., for the Lagrangian (7.35b) if $C = C(\mathbf{x})$ and $W = W(\mathbf{x})$ do not depend on time t), then the field's energy is conserved, $T^{0\nu}{}_{,\nu} = 0$. A similar calculation shows that if the Lagrangian has no explicit space dependence (e.g., if C and W are independent of x), then the field's momentum is conserved, $T^{j\nu}{}_{,\nu} = 0$. Here and throughout this chapter we use Cartesian spatial coordinates, so spatial partial derivatives (denoted by commas) are the same as covariant derivatives.

(c) Show that expression (7.35c) for the field's energy density $U = T^{00} = -T_0{}^0$ and its energy flux $F_i = T^{0i} = -T_0{}^i$ agree with Eqs. (7.18).

(d) Now, regard the wave amplitude ψ as a generalized (field) coordinate. Use the Lagrangian $L = \int \mathcal{L} d^3x$ to define a field momentum Π conjugate to this ψ, and then compute a *wave action*

$$J \equiv \int_0^{2\pi/\omega} \int \Pi (\partial\psi/\partial t) d^3x \, dt \ , \tag{7.35d}$$

which is the continuum analog of Eq. (7.42) below. The temporal integral is over one wave period. Show that this J is proportional to the wave energy divided by the frequency and thence to the number of quanta in the wave. [*Comment:* It is shown in standard texts on classical mechanics that, for approximately periodic oscillations, the particle action (7.42), with the integral limited to one period of oscillation of q, is an *adiabatic invariant*. By the extension of that proof to continuum physics, the wave action (7.35d) is also an adiabatic invariant. This means that the wave action, and thence also the number of quanta in the waves, are conserved when the medium [in our case the index of refraction $\mathfrak{n}(\mathbf{x})$] changes very slowly in time—a result asserted in the text, and a result that also follows from quantum mechanics. We shall study the particle version (7.42) of this adiabatic invariant in detail when we analyze charged particle motion in a slowly varying magnetic field in Sec. 20.7.4.]

Exercise 7.5 *Problem: Propagation of Sound Waves in a Wind*

Consider sound waves propagating in an atmosphere with a horizontal wind. Assume that the sound speed C, as measured in the air's local rest frame, is constant. Let the wind velocity $\mathbf{u} = u_x \mathbf{e}_x$ increase linearly with height z above the ground: $u_x = Sz$, where S is the constant shearing rate. Just consider rays in the $x - z$ plane.

(a) Give an expression for the dispersion relation $\omega = \Omega(\mathbf{x}, t; \mathbf{k})$. [Hint: in the local rest frame of the air, Ω should have its standard sound-wave form.]

(b) Show that k_x is constant along a ray path and then demonstrate that sound waves will not propagate when

$$\left| \frac{\omega}{k_x} - u_x(z) \right| < c \ . \tag{7.36}$$

(c) Consider sound rays generated on the ground which make an angle θ to the horizontal initially. Derive the equations describing the rays and use them to sketch the rays distinguishing values of θ both less than and greater than $\pi/2$. (You might like to perform this exercise numerically.)

****

7.3.3 Geometric Optics for a General Wave

With the simple case of non-dispersive sound waves (previous two subsections) as our model, we now study an arbitrary kind of wave in a weakly inhomogeneous and slowly time varying medium — e.g. any of the examples in Sec. 7.2.1: light waves in a dielectric medium, deep water waves, flexural waves on a stiff beam, or Alfvén waves. Whatever may be the wave, we seek a solution to its wave equation using the eikonal approximation $\psi = Ae^{i\varphi}$ with slowly varying amplitude A and rapidly varying phase φ. Depending on the nature of the wave, ψ and A might be a scalar (e.g. sound waves), a vector (e.g. light waves), or a tensor (e.g. gravitational waves).

When we insert the ansatz $\psi = Ae^{i\varphi}$ into the wave equation and collect terms in the manner dictated by our two lengthscale expansion [as in Eq. (7.22) and Box 7.2], the leading order term will arise from letting every temporal or spatial derivative act on the $e^{i\varphi}$. This is precisely where the derivatives would operate in the case of a plane wave in a homogeneous medium, and here as there the result of each differentiation is $\partial e^{i\varphi}/\partial t = -i\omega e^{i\varphi}$ or $\partial e^{i\varphi}/\partial x_j = ik_j e^{i\varphi}$. Correspondingly, *the leading order terms in the wave equation here will be identical to those in the homogeneous plane wave case: they will be the dispersion relation* multiplied by something times the wave,

$$[-\omega^2 + \Omega^2(\mathbf{x}, t; \mathbf{k})] \times (\text{something}) A e^{i\varphi} = 0 \ , \tag{7.37a}$$

with the spatial and temporal dependence in Ω^2 entering through the medium's properties. This guarantees that (as we claimed in Sec. 7.3.1) *the dispersion relation can be obtained by*

analyzing a plane, monochromatic wave in a homogeneous, time-independent medium and then letting the medium's properties, in the dispersion relation, vary slowly with $\mathbf{x}$ *and* t.

Each next-order ("subleading") term in the wave equation will entail just one of the wave operator's derivatives acting on a slowly-varying quantity (A or a medium property or ω or $\mathbf{k}$), and all the other derivatives acting on $e^{i\varphi}$. The subleading terms that interest us, for the moment, are those where the one derivative acts on A thereby propagating it. The subleading terms, therefore, can be deduced from the leading-order terms (7.37a) by replacing just one $i\omega A e^{i\varphi} = -A(e^{i\varphi})_{,t}$ by $-A_{,t}e^{i\varphi}$, and replacing just one $ik_j A e^{i\varphi} = A(e^{i\varphi})_{,j}$ by $A_{,j}e^{i\varphi}$. A little thought then reveals that the equation for the vanishing of the subleading terms must take the form [deducible from the leading terms (7.37a)]

$$-2i\omega\frac{\partial A}{\partial t} - 2i\Omega(\mathbf{k},\mathbf{x},t)\frac{\partial\Omega(\mathbf{k},\mathbf{x},t)}{\partial k_j}\frac{\partial A}{\partial x_j} = \text{terms proportional to } A\ . \tag{7.37b}$$

Using the dispersion relation $\omega = \Omega(\mathbf{x},t;\mathbf{k})$ and the group velocity (first Hamilton equation) $\partial\Omega/\partial k_j = V_{g\,j}$, we bring this into the "propagate A along a ray" form

$$\frac{dA}{dt} \equiv \frac{\partial A}{\partial t} + \mathbf{V}_g\cdot\boldsymbol{\nabla}A = \text{terms proportional to } A\ . \tag{7.37c}$$

Let us return to the leading order terms (7.37a) in the wave equation, i.e. to the dispersion relation $\omega = \Omega(\mathbf{x},t;k)$. For our general wave, as for the prototypical dispersionless wave of the previous two sections, the argument embodied in Eqs. (7.26) shows that *the rays are determined by Hamilton's equations* (7.25),

$$\boxed{\frac{dx_j}{dt} = \left(\frac{\partial\Omega}{\partial k_j}\right)_{\mathbf{x},t} \equiv V_{g\,j}}\ , \quad \boxed{\frac{dk_j}{dt} = -\left(\frac{\partial\Omega}{\partial x_j}\right)_{\mathbf{k},t}}\ , \quad \boxed{\frac{d\omega}{dt} = \left(\frac{\partial\Omega}{\partial t}\right)_{\mathbf{x},\mathbf{k}}}\ , \tag{7.38}$$

but using the general wave's dispersion relation $\Omega(\mathbf{k},\mathbf{x},t)$ rather than $\Omega = C(\mathbf{x},t)k$. *These Hamilton equations include propagation laws for* $\omega = -\partial\varphi/\partial t$ *and* $k_j = \partial\varphi/\partial x_j$*, from which we can deduce the propagation law (7.27) for* φ *along the rays*

$$\boxed{\frac{d\varphi}{dt} = -\omega + \mathbf{V}_g\cdot\mathbf{k}}\ . \tag{7.39}$$

For waves with dispersion, by contrast with sound in a fluid and other waves that have $\Omega = Ck$*,* φ *will not be constant along a ray.*

For our general wave, as for dispersionless waves, *the Hamilton equations for the rays can be reinterpreted as Hamilton's equations for the world lines of the waves' quanta* [Eq. (7.30) and associated discussion]. And for our general wave, as for dispersionless waves, *the medium's slow variations are incapable of creating or destroying wave quanta.* [This is a general feature of quantum theory; creation and destruction of quanta require imposed oscillations at the high frequency and short wavelength of the waves themselves, or at some submultiple of them (in the case of nonlinear creation and annihilation processes; Chap. 10).] Correspondingly, if one knows the relationship between the waves' energy density U

and their amplitude A, and thence the relationship between the waves' quantum number density $n = U/\hbar\omega$ and A, then *from the quantum conservation law* [boxed Eqs. (7.34)]

$$\boxed{\frac{\partial n}{\partial t} + \boldsymbol{\nabla}\cdot(n\mathbf{V}_g) = 0} \quad \text{or} \quad \boxed{\frac{dn}{dt} + n\boldsymbol{\nabla}\cdot\mathbf{V}_g = 0} \quad \text{or} \quad \boxed{\frac{d(nC\mathcal{A})}{dt} = 0} \tag{7.40}$$

one can deduce the propagation law for A — and the result must be the same propagation law as one obtains from the subleading terms in the eikonal approximation.

7.3.4 Examples of Geometric-Optics Wave Propagation

Spherical scalar waves.

As a simple example of these geometric-optics propagation laws, consider a scalar wave propagating radially outward from the origin at the speed of light in flat spacetime. Setting the speed of light to unity, the dispersion relation is Eq. (7.4) with $C = 1$: $\Omega = k$. It is straightforward (Ex 7.6) to integrate Hamilton's equations and learn that the rays have the simple form $\{r = t + \text{constant},\ \theta = \text{constant},\ \phi = \text{constant},\ \mathbf{k} = \omega\mathbf{e}_r\}$ in spherical polar coordinates, with $\mathbf{e}_r$ the unit radial vector. Because the wave is dispersionless, its phase φ must be conserved along a ray [Eq. (7.28)], i.e. φ must be a function of $t - r, \theta, \phi$. In order that the waves propagate radially, it is essential that $\mathbf{k} = \boldsymbol{\nabla}\varphi$ point very nearly radially; this implies that φ must be a rapidly varying function of $t - r$ and a slowly varying function of (θ, ϕ). The law of conservation of quanta in this case reduces to the propagation law $d(rA)/dt = 0$ (Ex. 7.6) so rA is also a constant along the ray; we shall call it $\mathcal{B}$. Putting this all together, we conclude that

$$\psi = \frac{\mathcal{B}(t - r, \theta, \phi)}{r} e^{i\varphi(t-r,\theta,\phi)} , \tag{7.41}$$

where the phase φ is rapidly varying in $t - r$ and slowly varying in the angles, and the amplitude is slowly varying in $t - r$ and the angles.

Flexural waves.

As another example of the geometric-optics propagation laws, consider flexural waves on a spacecraft's tapering antenna. The dispersion relation is $\Omega = k^2\sqrt{D/\Lambda}$ [Eq. (7.6)] with $D/\Lambda \propto h^2$, where h is the antenna's thickness in its direction of bend (or the antenna's diameter, if it has a circular cross section); cf. Eq. (12.34). Since Ω is independent of t, as the waves propagate from the spacecraft to the antenna's tip, their frequency ω is conserved [third of Eqs. (7.38)], which implies by the dispersion relation that $k = (D/\Lambda)^{-1/4}\omega^{1/2} \propto h^{-1/2}$, whence the wavelength decreases as $h^{1/2}$. The group velocity is $V_g = 2(D/\Lambda)^{1/4}\omega^{1/2} \propto h^{1/2}$. Since the energy per quantum $\hbar\omega$ is constant, particle conservation implies that the waves' energy must be conserved, which in this one-dimensional problem, means that the energy flux must be constant along the antenna. On physical grounds the constant energy flux must be proportional to A^2V_g, which means that the amplitude A must increase $\propto h^{-1/4}$ as the flexural waves approach the antenna's end. A qualitatively similar phenomenon is seen in the "cracking" of a bullwhip.

Light through lens, and Alfvén waves

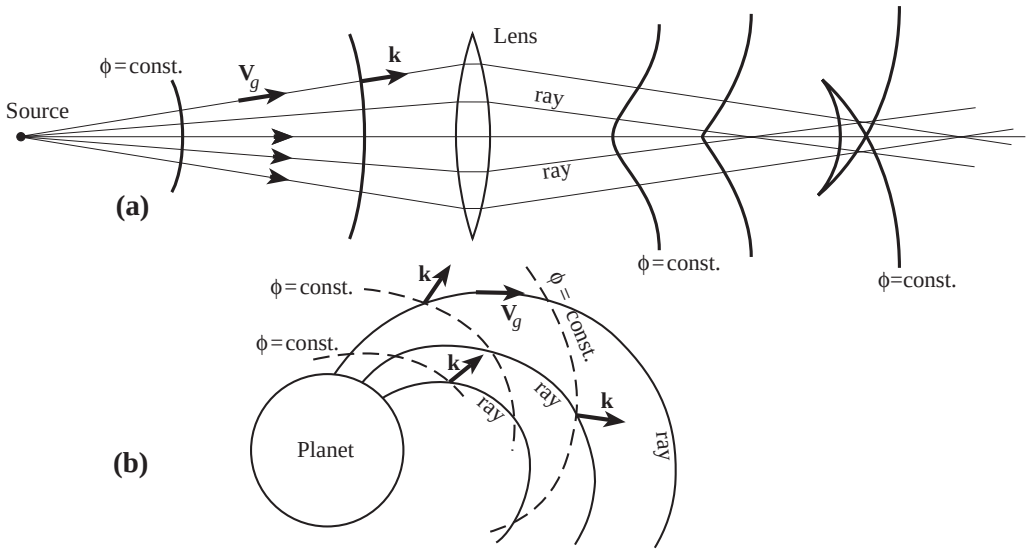


Fig. 7.3: (a) The rays and the surfaces of constant phase φ at a fixed time for light passing through a converging lens [dispersion relation $\Omega = ck/\mathfrak{n}(\mathbf{x})$, where $\mathfrak{n}$ is the index of refraction]. In this case the rays (which always point along $\mathbf{V}_g$) are parallel to the wave vector $\mathbf{k} = \boldsymbol{\nabla}\varphi$ and thus also parallel to the phase velocity $\mathbf{V}_{\rm ph}$, and the waves propagate along the rays with a speed $V_g = V_{\rm ph} = c/\mathfrak{n}$ that is independent of wavelength. The strange self-intersecting shape of the last phase front is due to caustics; see Sec. 7.5. (b) The rays and surfaces of constant phase for Alfvén waves in the magnetosphere of a planet [dispersion relation $\Omega = \mathbf{a}(\mathbf{x}) \cdot \mathbf{k}$]. In this case because $\mathbf{V}_g = \mathbf{a} \equiv \mathbf{B}/\sqrt{\mu_0\rho}$, the rays are parallel to the magnetic field lines and not parallel to the wave vector, and the waves propagate along the field lines with speeds $V_g = B/\sqrt{\mu_0\rho}$ that are independent of wavelength; cf. Fig. 7.2c. As a consequence, if some electric discharge excites Alfvén waves on the planetary surface, then they will be observable by a spacecraft when it passes magnetic field lines on which the discharge occurred. As the waves propagate, because $\mathbf{B}$ and ρ are time independent and thence $\partial\Omega/\partial t = 0$, the frequency ω and energy $\hbar\omega$ of each quantum is conserved, and conservation of quanta implies conservation of wave energy. Because the Alfvén speed generally diminishes with increasing distance from the planet, conservation of wave energy typically requires the waves' energy density and amplitude to increase as they climb upward.

Fig. 7.3 sketches two other examples: light propagating through a lens, and Alfvén waves propagating in the magnetosphere of a planet. In Sec. 7.3.6 and the exercises we shall explore a variety of other applications, but first we shall describe how the geometric-optics propagation laws can fail (Sec. 7.3.5).

****

EXERCISES

Exercise 7.6 *Derivation and Practice: Quasi-Spherical Solution to Vacuum Scalar Wave Equation*

Derive the quasi-spherical solution (7.41) of the vacuum scalar wave equation $-\partial^2\psi/\partial t^2 + \nabla^2\psi = 0$ from the geometric optics laws by the procedure sketched in the text.

****

7.3.5 Relation to Wave Packets; Breakdown of the Eikonal Approximation and Geometric Optics

The form $\psi = Ae^{i\varphi}$ of the waves in the eikonal approximation is remarkably general. At some initial moment of time, A and φ can have any form whatsoever, so long as the two-lengthscale constraints are satisfied [A, $\omega \equiv -\partial\varphi/\partial t$, $\mathbf{k} \equiv \boldsymbol{\nabla}\varphi$, and dispersion relation $\Omega(\mathbf{k}; \mathbf{x}, t)$ all vary on lengthscales long compared to $\lambda\!\!\!^{-} = 1/k$ and timescales long compared to $1/\omega$]. For example, ψ could be as nearly planar as is allowed by the inhomogeneities of the dispersion relation. At the other extreme, ψ could be a moderately narrow wave packet, confined initially to a small region of space (though not too small; its size must be large compared to its mean reduced wavelength). In either case, the evolution will be governed by the above propagation laws.

Of course, the eikonal approximation *is* an approximation. Its propagation laws make errors, though when the two-lengthscale constraints are well satisfied, the errors will be small for sufficiently short propagation times. Wave packets provide an important example. *Dispersion (different group velocities for different wave vectors) causes wave packets to spread (widen; disperse) as they propagate; see Ex. 7.2. This spreading is not included in the geometric optics propagation laws;* it is a fundamentally wave-based phenomenon and is lost when one goes to the particle-motion regime. In the limit that the wave packet becomes very large compared to its wavelength or that the packet propagates for only a short time, the spreading is small (Ex. 7.2). This is the geometric-optics regime, and geometric optics ignores the spreading.

Many other wave phenomena are missed by geometric optics. Examples are *diffraction*, e.g. at a geometric-optics caustic (Secs. 7.5 and 8.6), *nonlinear wave-wave coupling* (Chaps. 10 and 23, and Sec. 16.3), and *parametric amplification* of waves by rapid time variations of the medium (Sec. 10.6)—which shows up in quantum mechanics as *particle production* (i.e., a breakdown of the law of conservation of quanta). In Chap. 28, we shall study such particle production in inflationary models of the early universe.

7.3.6 Fermat's Principle

The Hamilton equations of optics allow us to solve for the paths of rays in media that vary both spatially and temporally. When the medium is time independent, the rays $\mathbf{x}(t)$ can be computed from a variational principle named after Pierre de Fermat. This *Fermat's principle* is the optical analogue of Maupertuis' principle of least action in classical mechanics. In classical mechanics, this principle states that, when a particle moves from one point to another through a time-independent potential (so its energy, the Hamiltonian, is conserved), then the path $\mathbf{q}(t)$ that it follows is one that extremizes the action

$$J = \int \mathbf{p} \cdot d\mathbf{q} \,, \tag{7.42}$$

(where $\mathbf{q}$, $\mathbf{p}$ are the particle's generalized coordinates and momentum), subject to the constraint that the paths have a fixed starting point, a fixed endpoint, and constant energy. The proof [which can be found, e.g., in Sec. 8.6 of Goldstein, Safko and Poole (2002)] carries over directly to optics when we replace the Hamiltonian by Ω, $\mathbf{q}$ by $\mathbf{x}$, and $\mathbf{p}$ by $\mathbf{k}$. The resulting Fermat principle, stated with some care, has the following form:

Consider waves whose Hamiltonian $\Omega(\mathbf{k},\mathbf{x})$ is independent of time. Choose an initial location $\mathbf{x}_{\rm initial}$ and a final location $\mathbf{x}_{\rm final}$ in space, and ask what are the rays $\mathbf{x}(t)$ that connect these two points. The rays (usually only one) are those paths that satisfy the variational principle

$$\boxed{\delta \int \mathbf{k}\cdot d\mathbf{x} = 0}\,. \tag{7.43}$$

In this variational principle, $\mathbf{k}$ *must be expressed in terms of the trial path $\mathbf{x}(t)$ using Hamilton's equation $dx^j/dt = -\partial\Omega/\partial k_j$; the rate that the trial path is traversed (i.e., the magnitude of the group velocity) must be adjusted so as to keep Ω constant along the trial path* (which means that the total time taken to go from $\mathbf{x}_{\rm initial}$ to $\mathbf{x}_{\rm final}$ can differ from one trial path to another); and, of course, the trial paths must all begin at $\mathbf{x}_{\rm initial}$ and end at $\mathbf{x}_{\rm final}$.

Notice that, once a ray has been identified via this action principle, it has $\mathbf{k} = \boldsymbol{\nabla}\varphi$, and therefore the extremal value of the action $\int \mathbf{k}\cdot d\mathbf{x}$ along the ray is equal to the waves' phase difference $\Delta\varphi$ between $\mathbf{x}_{\rm initial}$ and $\mathbf{x}_{\rm final}$. Correspondingly, for any trial path we can think of the action as a phase difference along that path,

$$\Delta\varphi = \int \mathbf{k}\cdot d\mathbf{x}\,, \tag{7.44a}$$

and *we can think of the action principle as one of extremal phase difference $\Delta\varphi$*. This can be reexpressed in a form closely related to ***Feynman's path-integral formulation of quantum mechanics:*** We can regard *all* the trial paths as being followed with equal probability; for each path we are to construct a probability amplitude $e^{i\Delta\varphi}$; and we must then add together these amplitudes

$$\sum_{\text{all paths}} e^{i\Delta\varphi} \tag{7.44b}$$

to get the net complex amplitude for quanta associated with the waves to travel from $\mathbf{x}_{\rm initial}$ to $\mathbf{x}_{\rm final}$. The contributions from almost all neighboring paths will interfere destructively. The only exceptions are those paths whose neighbors have the same values of $\Delta\varphi$, to first order in the path difference. These are the paths that extremize the action (7.43); i.e., they are the wave's rays, the actual paths of the quanta.

Specialization to $\Omega = C(\mathbf{x})k$

Fermat's principle takes on an especially simple form when not only is the Hamiltonian $\Omega(\mathbf{k},\mathbf{x})$ time independent, but it also has the simple dispersion-free form $\Omega = C(\mathbf{x})k$ — a form valid for propagation of light through a time-independent dielectric, and sound waves through a time-independent, inhomogeneous fluid, and electromagnetic or gravitational waves through a time-independent, Newtonian gravitational field (Sec. 7.6). In this $\Omega = C(\mathbf{x})k$ case, the Hamiltonian dictates that for each trial path, $\mathbf{k}$ is parallel to $d\mathbf{x}$, and therefore $\mathbf{k}\cdot d\mathbf{x} = k ds$, where s is distance along the path. Using the dispersion relation

$k = \Omega/C$ and noting that Hamilton's equation $dx^j/dt = \partial\Omega/\partial k_j$ implies $ds/dt = C$ for the rate of traversal of the trial path, we see that $\mathbf{k}\cdot d\mathbf{x} = k ds = \Omega dt$. Since the trial paths are constrained to have Ω constant, *Fermat's principle (7.43) becomes a principle of extremal time: The rays between* $\mathbf{x}_{\rm initial}$ *and* $\mathbf{x}_{\rm final}$ *are those paths along which*

$$\boxed{\int dt = \int \frac{ds}{C(\mathbf{x})} = \int \frac{\mathfrak{n}(\mathbf{x})}{c} ds} \tag{7.45}$$

is extremal. In the last expression we have adopted the convention used for light in a dielectric medium, that $C(\mathbf{x}) = c/\mathfrak{n}(\mathbf{x})$, where c is the speed of light in vacuum and $\mathfrak{n}$ is the medium's index of refraction. Since c is constant, *the rays are paths of extremal optical path length* $\int \mathfrak{n}(\mathbf{x}) ds$.

We can use Fermat's principle to demonstrate that, if the medium contains no opaque objects, then there will always be at least one ray connecting any two points. This is because there is a lower bound on the optical path between any two points given by $\mathfrak{n}_{\rm min} L$, where $\mathfrak{n}_{\rm min}$ is the lowest value of the refractive index anywhere in the medium and L is the distance between the two points. This means that for some path the optical path length must be a minimum, and that path is then a ray connecting the two points.

From the principle of extremal time, we can derive the Euler-Lagrange differential equation for the ray. For ease of derivation, we write the action principle in the form

$$\delta \int \mathfrak{n}(\mathbf{x}) \sqrt{\frac{d\mathbf{x}}{d\mathbf{s}} \cdot \frac{d\mathbf{x}}{d\mathbf{s}}}\, ds, \tag{7.46}$$

where the quantity in the square root is identically one. Performing a variation in the usual manner then gives

$$\boxed{\frac{d}{ds}\left(\mathfrak{n}\frac{d\mathbf{x}}{ds}\right) = \boldsymbol{\nabla}\mathfrak{n}\,, \quad \text{i.e. } \frac{d}{ds}\left(\frac{1}{C}\frac{d\mathbf{x}}{ds}\right) = \boldsymbol{\nabla}\left(\frac{1}{C}\right)}\,. \tag{7.47}$$

This is equivalent to Hamilton's equations for the ray, as one can readily verify using the Hamiltonian $\Omega = kc/\mathfrak{n}$ (Ex. 7.7).

Equation (7.47) is a second-order differential equation requiring two boundary conditions to define a solution. We can either choose these to be the location of the start of the ray and its starting direction, or the start and end of the ray. A simple case arises *when the medium is stratified*, i.e. when $\mathfrak{n} = \mathfrak{n}(z)$, where (x, y, z) are Cartesian coordinates. Projecting Eq. (7.47) perpendicular to $\mathbf{e}_z$, we discover that $\mathfrak{n}dy/ds$ and $\mathfrak{n}dx/ds$ are constant, which implies

$$\boxed{\mathfrak{n}\sin\theta = \text{constant}}\,, \tag{7.48}$$

where θ *is the angle between the ray and* $\mathbf{e}_z$. This is Snell's law of refraction. Snell's law is just a mathematical statement that the rays are normal to surfaces (wavefronts) on which the eikonal (phase) φ is constant (cf. Fig. 7.4).

****

EXERCISES

Exercise 7.7 *Derivation: Hamilton's Equations for Dispersionless Waves; Fermat's Principle*

Show that Hamilton's equations for the standard dispersionless dispersion relation (7.4) imply the same ray equation (7.47) as we derived using Fermat's principle.

Exercise 7.8 *Example: Self-Focusing Optical Fibers*

Optical fibers in which the refractive index varies with radius are commonly used to transport optical signals. Provided that the diameter of the fiber is many wavelengths, we can use geometric optics. Let the refractive index be

$$\mathfrak{n} = \mathfrak{n}_0(1 - \alpha^2 r^2)^{1/2} \ , \tag{7.49a}$$

where n_0 and α are constants and r is radial distance from the fiber's axis.

(a) Consider a ray that leaves the axis of the fiber along a direction that makes a small angle θ to the axis. Solve the ray transport equation (7.47) to show that the radius of the ray is given by

$$r = \frac{\sin\theta}{\alpha}\left|\sin\left(\frac{\alpha z}{\cos\theta}\right)\right| \ , \tag{7.49b}$$

where z measures distance along the fiber.

(b) Next consider the propagation time T for a light pulse propagating along a long length L of fiber. Show that

$$T = \frac{\mathfrak{n}_0 L}{c}[1 + O(\theta^4)] \ , \tag{7.49c}$$

and comment on the implications of this result for the use of fiber optics for communication.

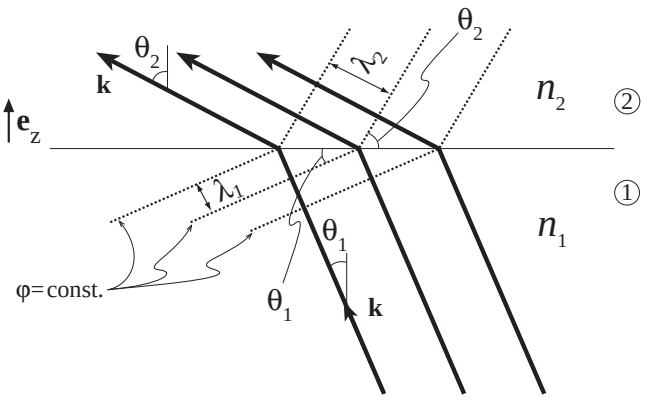


Fig. 7.4: Illustration of Snell's law of refraction at the interface between two media where the refractive indices are $\mathfrak{n}_1$, and $\mathfrak{n}_2$ (assumed less than $\mathfrak{n}_1$). As the wavefronts must be continuous across the interface, simple geometry tells us that $\lambda_1/\sin\theta_1 = \lambda_2/\sin\theta_2$. This and the fact that the wavelengths are inversely proportional to the refractive index, $\lambda_j \propto 1/\mathfrak{n}_j$, imply that $\mathfrak{n}_1 \sin\theta_1 = \mathfrak{n}_2 \sin\theta_2$, in agreement with Eq. (7.48).

Exercise 7.9 **** Example: Geometric Optics for the Schrödinger Equation*
Consider the non-relativistic Schrödinger equation for a particle moving in a time-dependent, 3-dimensional potential well:

$$-\frac{\hbar}{i}\frac{\partial\psi}{\partial t} = \left[\frac{1}{2m}\left(\frac{\hbar}{i}\nabla\right)^2 + V(\mathbf{x},t)\right]\psi\ . \tag{7.50}$$

(a) Seek a geometric optics solution to this equation with the form $\psi = Ae^{iS/\hbar}$, where A and V are assumed to vary on a lengthscale $\mathcal{L}$ and timescale $\mathcal{T}$ long compared to those, $1/k$ and $1/\omega$, on which S varies. Show that the leading order terms in the two-lengthscale expansion of the Schrödinger equation give the Hamilton-Jacobi equation

$$\frac{\partial S}{\partial t} + \frac{1}{2m}(\boldsymbol{\nabla}S)^2 + V = 0\ . \tag{7.51a}$$

Our notation $\varphi \equiv S/\hbar$ for the phase φ of the wave function ψ is motivated by the fact that the geometric-optics limit of quantum mechanics is classical mechanics, and the function $S = \hbar\varphi$ becomes, in that limit, "Hamilton's principal function," which obeys the Hamilton-Jacobi equation.[5] [Hint: Use a formal parameter σ to keep track of orders (Box 7.2), and argue that terms proportional to $\hbar^n$ are of order σ^n. This means there must be factors of σ in the Schrödinger equation (7.50) itself.]

(b) From Eq. (7.51a) derive the equation of motion for the rays (which of course is identical to the equation of motion for a wave packet and therefore is also the equation of motion for a classical particle):

$$\frac{d\mathbf{x}}{dt} = \frac{\mathbf{p}}{m}\ , \quad \frac{d\mathbf{p}}{dt} = -\boldsymbol{\nabla}V\ , \tag{7.51b}$$

where $\mathbf{p} = \boldsymbol{\nabla}S$.

(c) Derive the propagation equation for the wave amplitude A and show that it implies

$$\frac{d|A|^2}{dt} + |A|^2\frac{\boldsymbol{\nabla}\cdot\mathbf{p}}{m} = 0\ . \tag{7.51c}$$

Interpret this equation quantum mechanically.

****

7.4 Paraxial Optics

It is quite common in optics to be concerned with a bundle of rays that are almost parallel, i.e. for which the angle the rays make with some reference ray can be treated as small. This approximation is called *paraxial optics*, and it permits one to linearize the geometric optics

[5] See, e.g., Chap. 10 of Goldstein, Safko and Poole.

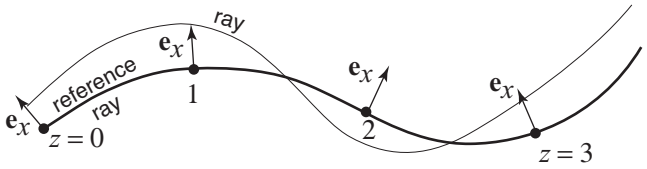


Fig. 7.5: A reference ray (thick curve), parameterized by distance z along it, and with a parallel-transported transverse unit basis vector $\mathbf{e}_x$ that delineates a slowly changing transverse x axis. Other rays, such as the thin curve, are identified by their transverse distances $x(z)$ and $y(z)$, along $\mathbf{e}_x$ and $\mathbf{e}_y$ (not shown).

equations and use matrix methods to trace their rays. The resulting matrix formalism underlies the first order theory of simple optical instruments, e.g. the telescope and the microscope.

We shall develop the paraxial optics formalism for waves whose dispersion relation has the simple, time-independent, nondispersive form $\Omega = kc/\mathfrak{n}(\mathbf{x})$. This applies to light in a dielectric medium — the usual application. As we shall see below, it also applies to charged particles in a storage ring or electron microscope (Sec. 7.4.2) and to light being lensed by a weak gravitational field (Sec. 7.6).

We start by linearizing the ray propagation equation (7.47). Let z measure distance along a reference ray. Let the two dimensional vector $\mathbf{x}(z)$ be the transverse displacement of some other ray from this reference ray, and denote by $(x, y) = (x_1, x_2)$ the Cartesian components of $\mathbf{x}$, with the transverse Cartesian basis vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ transported parallely along the reference ray[6] (Fig. 7.5).

Under paraxial conditions, $|\mathbf{x}|$ is small compared to the z-lengthscales of the propagation, so we can Taylor expand the refractive index $\mathfrak{n}(\mathbf{x}, z)$ in (x_1, x_2):

$$\mathfrak{n}(\mathbf{x}, z) = \mathfrak{n}(0, z) + x_i \mathfrak{n}_{,i}(0, z) + \frac{1}{2} x_i x_j \mathfrak{n}_{,ij}(0, z) + \dots \ . \tag{7.52a}$$

Here the subscript commas denote partial derivatives with respect to the transverse coordinates, $\mathfrak{n}_{,i} \equiv \partial \mathfrak{n}/\partial x_i$. The linearized form of the ray propagation equation (7.47) is then given by

$$\frac{d}{dz}\left(\mathfrak{n}(0, z)\frac{dx_i}{dz}\right) = \mathfrak{n}_{,i}(0, z) + x_j \mathfrak{n}_{,ij}(0, z) \ . \tag{7.52b}$$

Unless otherwise stated, we restrict ourselves to aligned optical systems in which there is a particular choice of reference ray called the *optic axis*, for which the term $\mathfrak{n}_{,i}(0, z)$ on the right hand side of Eq. (7.52b) vanishes, and we choose the reference ray to be the optic axis. Then Eq. (7.52b) is a linear, homogeneous, second-order equation for $\mathbf{x}(z)$,

$$\boxed{(d/dz)(\mathfrak{n} dx_i/dz) = x_j \mathfrak{n}_{,ij}} \ . \tag{7.53}$$

Here $\mathfrak{n}$ and $\mathfrak{n}_{,ij}$ are evaluated on the reference ray. It is helpful to regard z as "time" and think of Eq. (7.53) as an equation for the two dimensional motion of a particle (the ray) in a

[6]By parallel transport, we mean this: the basis vector $\mathbf{e}_x$ is carried a short distance along the reference ray, keeping it parallel to itself. Then, if the reference ray has bent a bit, $\mathbf{e}_x$ is projected into the new plane that is transverse to the ray.

quadratic potential well. We can solve Eq. (7.53) given starting values $\mathbf{x}(z'), \dot{\mathbf{x}}(z')$ where the dot denotes differentiation with respect to z, and z' is the starting location. The solution at some later point z is linearly related to the starting values. We can capitalize on this linearity by treating $\{\mathbf{x}(z), \dot{\mathbf{x}}(z)\}$ as a 4 dimensional vector $V_i(z)$, with

$$\boxed{V_1 = x, \quad V_2 = \dot{x}, \quad V_3 = y, \quad V_4 = \dot{y}}\,, \tag{7.54a}$$

and embodying the linear transformation [linear solution of Eq. (7.53)] from location z' to location z in a *transfer matrix* $J_{ab}(z, z')$:

$$\boxed{V_a(z) = J_{ab}(z, z') \cdot V_b(z')}\,. \tag{7.54b}$$

The transfer matrix contains full information about the change of position and direction of all rays that propagate from z' to z. As is always the case for linear systems, the transfer matrix for propagation over a large interval, from z' to z, can be written as the product of the matrices for two subintervals, from z' to z'' and from z'' to z:

$$\boxed{J_{ac}(z, z') = J_{ab}(z, z'')J_{bc}(z'', z')}\,. \tag{7.54c}$$

7.4.1 Axisymmetric, Paraxial Systems; Lenses, Mirrors, Telescope, Microscope and Optical Cavity

If the index of refraction is everywhere axisymmetric, so $\mathfrak{n} = \mathfrak{n}(\sqrt{x^2 + y^2}, z)$, then there is *no coupling between the motions of rays along the x and y directions, and the equations of motion along x are identical to those along y.* In other words, $J_{11} = J_{33}$, $J_{12} = J_{34}$, $J_{21} = J_{43}$, and $J_{22} = J_{44}$ are the only nonzero components of the transfer matrix. This reduces the dimensionality of the propagation problem from 4 dimensions to 2: V_a can be regarded as either $\{x(z), \dot{x}(z)\}$ or $\{y(z), \dot{y}(z)\}$, and in both cases the 2×2 transfer matrix J_{ab} is the same.

Let us illustrate the paraxial formalism by deriving the transfer matrices of a few simple, axisymmetric optical elements. In our derivations it is helpful conceptually to focus on rays that move in the x-z plane, i.e. that have $y = \dot{y} = 0$. We shall write the 2-dimensional V_i as a column vector

$$\boxed{V_a = \begin{pmatrix} x \\ \dot{x} \end{pmatrix}}\,. \tag{7.55a}$$

The simplest case is a straight section of length d extending from z' to $z = z' + d$. The components of V will change according to

$$\begin{aligned} x &= x' + \dot{x}' d\,, \\ \dot{x} &= \dot{x}'\,, \end{aligned}$$

so

$$\boxed{J_{ab} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \text{ for straight section of length } d}\,, \tag{7.55b}$$

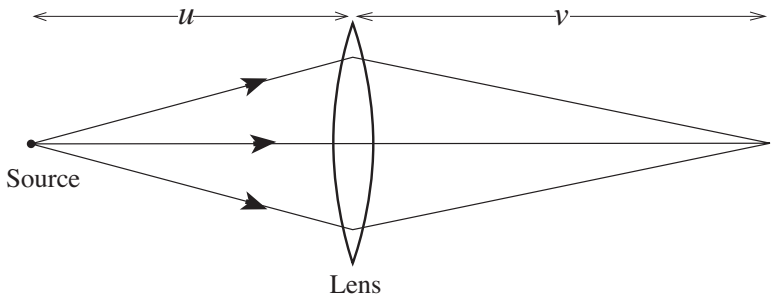


Fig. 7.6: Simple converging lens used to illustrate the use of transfer matrices. The total transfer matrix is formed by taking the product of the straight section transfer matrix with the lens matrix and another straight section matrix.

where $x' = x(z')$ etc. Next, consider a thin lens with focal length f. The usual convention in optics is to *give f a positive sign when the lens is converging and a negative sign when diverging.* A thin lens gives a deflection to the ray that is linearly proportional to its displacement from the optic axis, but does not change its transverse location. Correspondingly, the transfer matrix in crossing the lens (ignoring its thickness) is:

$$\boxed{J_{ab} = \begin{pmatrix} 1 & 0 \\ -f^{-1} & 1 \end{pmatrix} \text{ for thin lens with focal length } f} \,. \tag{7.55c}$$

Similarly, a spherical mirror with radius of curvature R (again *adopting a positive sign for a converging mirror and a negative sign for a diverging mirror*) has a transfer matrix

$$\boxed{J_{ab} = \begin{pmatrix} 1 & 0 \\ -2R^{-1} & 1 \end{pmatrix} \text{ for spherical mirror with radius of curvature } R} \,. \tag{7.55d}$$

As a simple illustration, we consider rays that leave a point source which is located a distance u in front of a converging lens of focal length f, and we solve for the ray positions a distance v behind the lens (Fig. 7.6). The total transfer matrix is the transfer matrix (7.55b) for a straight section, multiplied by the product of the lens transfer matrix (7.55c) and a second straight-section transfer matrix:

$$J_{ab} = \begin{pmatrix} 1 & v \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -f^{-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 - vf^{-1} & u + v - uvf^{-1} \\ -f^{-1} & 1 - uf^{-1} \end{pmatrix} . \tag{7.56}$$

When the 1-2 element (upper right entry) of this transfer matrix vanishes, the position of the ray after traversing the optical system is independent of the starting direction. In other words, *rays from the point source form a point image. When this happens, the planes containing the source and the image are said be conjugate.* The condition for this to occur is

$$\boxed{\frac{1}{u} + \frac{1}{v} = \frac{1}{f}} \,. \tag{7.57}$$

This is the standard thin-lens equation. *The linear magnification of the image* is given by $\mathcal{M} = J_{11} = 1 - v/f$, i.e.

$$\boxed{\mathcal{M} = -\frac{v}{u}}\,, \tag{7.58}$$

where the negative sign means that the image is inverted. Note that, if a ray is reversed in direction, it remains a ray, but with the source and image planes interchanged; u and v are exchanged, Eq. (7.57) is unaffected, and the magnification (7.58) is inverted: $\mathcal{M} \to 1/\mathcal{M}$.

****

EXERCISES

Exercise 7.10 *Problem: Matrix Optics for a Simple Refracting Telescope*

Consider a simple refracting telescope (Fig. 7.7) that comprises two converging lenses, the *objective* and the *eyepiece*. This telescope takes parallel rays of light from distant stars, which make an angle $\theta \ll 1$ with the optic axis, and converts them into parallel rays making a much larger angle $\mathcal{M}\theta$. Here $\mathcal{M}$ is the magnification with $\mathcal{M}$ negative, $|\mathcal{M}| \gg 1$ and $|\mathcal{M}\theta| \ll 1$. (The parallel output rays are then focused by the lens of a human's eye, to a point on the eye's retina.)

(a) Use matrix methods to investigate how the output rays depend on the separation of the two lenses and hence find the condition that the output rays are parallel when the input rays are parallel.

(b) How does the magnification $\mathcal{M}$ depend on the ratio of the focal lengths of the two lenses?

(c) If, instead of looking through the telescope with one's eye, one wants to record the stars' image on a photographic plate or CCD, how should the optics be changed?

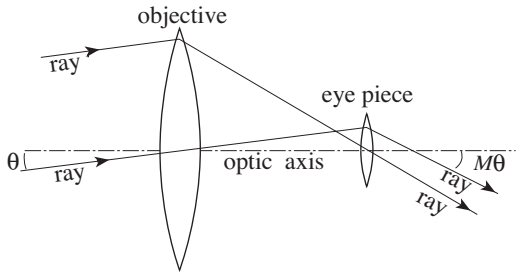


Fig. 7.7: Simple refracting telescope. By convention $\theta > 0$ and $\mathcal{M}\theta < 0$, so the image is inverted.

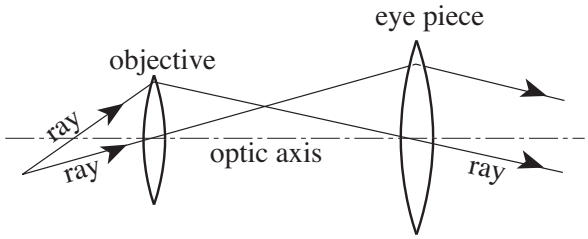


Fig. 7.8: Simple microscope.

Exercise 7.11 *Problem: Matrix Optics for a Simple Microscope*

A microscope takes light rays from a point on a microscopic object, very near the optic axis, and transforms them into parallel light rays that will be focused by a human eye's lens onto the eye's retina. Use matrix methods to explore the operation of such a microscope. A single lens (magnifying glass) could do the same job (rays from a point converted to parallel rays). Why does a microscope need two lenses? What focal lengths and lens separations are appropriate for the eye to resolve a bacterium, 100μm in size?

Exercise 7.12 *Example: Optical Cavity – Rays Bouncing Between Two Mirrors*

Consider two spherical mirrors, each with radius of curvature R, separated by distance d so as to form an "optical cavity," as shown in Fig. 7.9. A laser beam bounces back and forth between the two mirrors. The center of the beam travels along a geometric-optics ray. (We shall study such beams, including their diffractive behavior, in Sec. 8.5.5.)

(a) Show, using matrix methods, that the central ray hits one of the mirrors (either one) at successive locations $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \ldots$ (where $\mathbf{x} \equiv (x, y)$ is a 2 dimensional vector in the plane perpendicular to the optic axis), which satisfy the difference equation

$$\mathbf{x}_{k+2} - 2b\mathbf{x}_{k+1} + \mathbf{x}_k = 0 \tag{7.59a}$$

where

$$b = 1 - \frac{4d}{R} + \frac{2d^2}{R^2}. \tag{7.59b}$$

Explain why this is a difference-equation analogue of the simple-harmonic-oscillator equation.

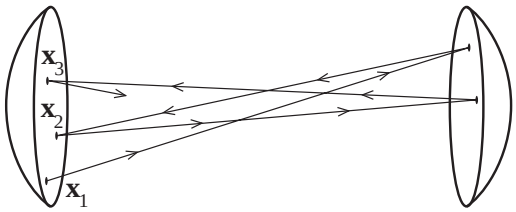


Fig. 7.9: An optical cavity formed by two mirrors, and a light beam bouncing back and forth inside it.

(b) Show that this difference equation has the general solution

$$\mathbf{x}_k = \mathbf{A}\cos(k\cos^{-1} b) + \mathbf{B}\sin(k\cos^{-1} b)\ . \tag{7.59c}$$

Obviously, $\mathbf{A}$ is the transverse position $\mathbf{x}_0$ of the ray at its 0'th bounce. The ray's 0'th position $\mathbf{x}_0$ and its 0'th direction of motion $\dot{\mathbf{x}}_0$ together determine $\mathbf{B}$.

(c) Show that if $0 \le d \le 2R$, the mirror system is "stable". In other words, all rays oscillate about the optic axis. Similarly, show that if $d > 2R$, the mirror system is unstable and the rays diverge from the optic axis.

(d) For an appropriate choice of initial conditions $\mathbf{x}_0$ and $\dot{\mathbf{x}}_0$, the laser beam's successive spots on the mirror lie on a circle centered on the optic axis. When operated in this manner, the cavity is called a *Harriet delay line*. How must d/R be chosen so that the spots have an angular step size θ? (There are two possible choices.)

****

7.4.2 Converging Magnetic Lens for Charged Particle Beam

Since geometric optics is the same as particle dynamics, matrix equations can be used to describe paraxial motions of electrons or ions in a storage ring. (Note, however, that the Hamiltonian for such particles is dispersive, since the Hamiltonian does not depend linearly on the particle momentum, and so for our simple matrix formalism to be valid, we must confine attention to a *mono-energetic beam of particles.*)

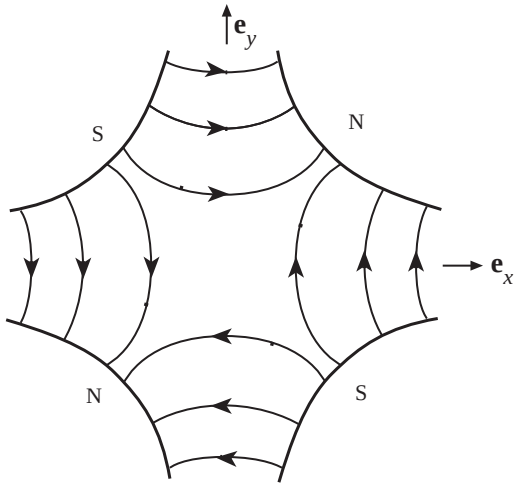


Fig. 7.10: Quadrupolar Magnetic Lens. The magnetic field lines lie in a plane perpendicular to the optic axis. Positively charged particles moving along $\mathbf{e}_z$ are converged when $y = 0$ and diverged when $x = 0$.

The simplest practical lens for charged particles is a quadrupolar magnet. Quadrupolar magnetic fields are used to guide particles around storage rings. If we orient our axes appropriately, the magnet's magnetic field can be expressed in the form

$$\mathbf{B} = \frac{B_0}{r_0}(y\mathbf{e}_x + x\mathbf{e}_y) \quad \text{independent of } z \text{ within the lens} \tag{7.60}$$

(Fig. 7.10). Particles traversing this magnetic field will be subjected to a Lorentz force which will curve their trajectories. In the paraxial approximation, a particle's coordinates will satisfy the two differential equations

$$\ddot{x} = -\frac{x}{\lambda^2}\,, \quad \ddot{y} = \frac{y}{\lambda^2}\,, \tag{7.61a}$$

where the dots (as above) mean $d/dz = v^{-1}d/dt$ and

$$\lambda = \left(\frac{pr_0}{qB_0}\right)^{1/2}\,, \tag{7.61b}$$

with q the particle's charge (assumed positive) and p its momentum. The motions in the x and y directions are decoupled. It is convenient in this case to work with two 2-dimensional vectors, $\{V_{x1}, V_{x2}\} \equiv \{x, \dot{x}\}$ and $\{V_{y1}, V_{y2}\} = \{y, \dot{y}\}$. From the elementary solutions to the equations of motion (7.61a), we infer that the transfer matrices from the magnet's entrance to its exit are $J_{x\,ab}, J_{y\,ab}$, where

$$J_{x\,ab} = \begin{pmatrix} \cos\phi & \lambda\sin\phi \\ -\lambda^{-1}\sin\phi & \cos\phi \end{pmatrix}\,, \tag{7.62a}$$

$$J_{y\,ab} = \begin{pmatrix} \cosh\phi & \lambda\sinh\phi \\ \lambda^{-1}\sinh\phi & \cosh\phi \end{pmatrix}\,, \tag{7.62b}$$

and

$$\phi = L/\lambda\,, \tag{7.62c}$$

with L the distance from entrance to exit (i.e. the lens thickness).

The matrices $J_{x\,ab}, J_{y\,ab}$ can be decomposed as follows

$$J_{x\,ab} = \begin{pmatrix} 1 & \lambda\tan\phi/2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\sin\phi/\lambda & 1 \end{pmatrix}\begin{pmatrix} 1 & \lambda\tan\phi/2 \\ 0 & 1 \end{pmatrix} \tag{7.62d}$$

$$J_{y\,ab} = \begin{pmatrix} 1 & \lambda\tanh\phi/2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \sinh\phi/\lambda & 1 \end{pmatrix}\begin{pmatrix} 1 & \lambda\tanh\phi/2 \\ 0 & 1 \end{pmatrix} \tag{7.62e}$$

Comparing with Eqs. (7.55b), (7.55c), we see that *the action of a single magnet is equivalent to the action of a straight section, followed by a thin lens, followed by another straight section.* Unfortunately, if the lens is focusing in the x direction, it must be de-focusing in the y direction and *vice versa*. However, *we can construct a lens that is focusing along both directions by combining two magnets that have opposite polarity but the same focusing strength* $\phi = L/\lambda$:

Consider first the particles' motion in the x direction. Let

$$f_+ = \lambda/\sin\phi \quad \text{and} \quad f_- = -\lambda/\sinh\phi \tag{7.63}$$

be the equivalent focal length of the first converging lens and the second diverging lens. If we separate the magnets by a distance s, this must be added to the two effective lengths of the two magnets to give an equivalent separation, $d = \lambda\tan(\phi/2) + s + \lambda\tanh(\phi/2)$ for the two equivalent thin lenses. The combined transfer matrix for the two thin lenses separated by this distance d is then

$$\begin{pmatrix} 1 & 0 \\ -f_-^{-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -f_+^{-1} & 1 \end{pmatrix} = \begin{pmatrix} 1 - df_+^{-1} & d \\ -f_*^{-1} & 1 - df_-^{-1} \end{pmatrix} , \tag{7.64a}$$

where

$$\frac{1}{f_*} = \frac{1}{f_-} + \frac{1}{f_+} - \frac{d}{f_-f_+} = \frac{\sin\phi}{\lambda} - \frac{\sinh\phi}{\lambda} + \frac{d\sin\phi\sinh\phi}{\lambda^2} . \tag{7.64b}$$

If we assume that $\phi \ll 1$ and $s \ll L$, then we can expand as a Taylor series in ϕ to obtain

$$f_* \simeq \frac{3\lambda}{2\phi^3} = \frac{3\lambda^4}{2L^3} . \tag{7.65}$$

The effective focal length f_* of the combined magnets is positive and so the lens has a net focusing effect. From the symmetry of Eq. (7.64b) under interchange of f_+ and f_-, it should be clear that f_* is independent of the order in which the magnets are encountered. Therefore, if we were to repeat the calculation for the motion in the y direction, we would get the same focusing effect. (The diagonal elements of the transfer matrix are interchanged, but as they are both close to unity, this is a rather small difference.)

The combination of two quadrupole lenses of opposite polarity can therefore imitate the action of a converging lens. Combinations of magnets like this are used to collimate particle beams in storage rings, particle accelerators, and electron microscopes.

7.5 Catastrophe Optics — Multiple Images; Formation of Caustics and their Properties

Many simple optical instruments are carefully made so as to form point images from point sources. However, naturally occuring optical systems, and indeed precision optical instruments, when examined in fine detail, bring light to a focus not at a point, but instead on a 2-dimensional (2D) surface in 3D space, called a *caustic*.[7] Caustics are often seen in everyday life. For example, when when sunlight is refracted by the wavy water on the surface of a swimming pool, its rays form a complex pattern of 2D caustics which intersect the pool's 2D bottom in bright lines (Fig. 7.11a). And when sunlight passes through a glass of water (Fig. 7.11b), it produces 2D caustics that meet in a line cusp, which intersect a sheet of paper in caustic lines and a point cusp. What may be surprising is that caustics formed under quite general conditions can be classified into a rather small number of types, called *catastrophes*,

including the "fold catastrophes" in Fig. 7.11a, and the "cusp catastrophes" in Fig. 7.11b, and three others. Exercise 7.13 gives some insight into this.

A simple example that is easy to understand analytically occurs when light from a distant source passes through an imperfect but axisymmetric converging lens; see Fig. 7.12. Here the caustic is the envelope $\mathcal{C}$ of refracted rays (more precisely, the axially symmetric surface formed by rotating the curves $\mathcal{C}$ around the optic axis). When the light from the lens hits a screen, it will form a bright circle where the caustic $\mathcal{C}$ intersects the screen, analogous to the bright Caustic lines in Fig. 7.11.

An observer at any point $\mathcal{A}$ outside $\mathcal{C}$ will see a single image of the source in the direction of the single ray that passes through $\mathcal{A}$; whereas an observer at any point $\mathcal{B}$ inside $\mathcal{C}$ will see three images. If the observer at $\mathcal{B}$ moves downward toward the caustic, then she will see two of the images (those along the dotted rays $\mathcal{R}_1$ and $\mathcal{R}_2$) approach each other, merge, and then vanish, leaving just one ray (the solid line through $\mathcal{B}$) and its image as she moves beyond $\mathcal{C}$.

This behavior can be understood with the aid of Fermat's theorem and the construction in the inset at the right side of Fig. 7.12. Because of axisymmetry, the rays all lie in planes of constant azimuthal angle ϕ, so we shall focus solely on such a plane — the plane shown in the figure. We choose a point $\mathcal{B}$ and characterize it by its longitudinal distance d from the lens, and its cylindrical radius ϖ. We construct a set of *virtual rays*[8] that hit the lens orthogonally, then bend (at constant ϕ) and travel straight through $\mathcal{B}$ (the solid lines in the inset). We compute the total phase $\varphi(b; d, \varpi)$ along each of these virtual rays when it arrives

[7] See, for example, Berry & Upstill (1980).

[8] A virtual ray is a path that will become a real ray if it satisfies Fermat's principle.

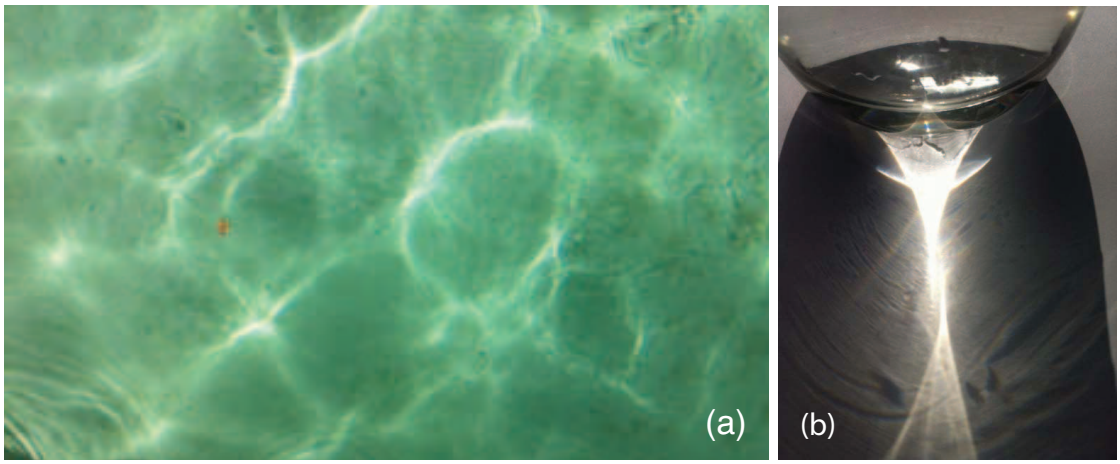


Fig. 7.11: Photographs of caustics. (a) Rich pattern of caustics on the bottom of a swimming pool, produced when sunlight passes through the pool's wavy surface; away from points of intersection, these are fold catastrophes. (b) Two caustics (cusp catastrophes) on a sheet of paper, produced when sunlight passes through a water glass. (If the glass were perfectly circularly symmetric instead of flawed, the cusps would meet at their points.) In both cases the caustic lines shown are cross sections through 2D caustic surfaces, which extend upward from the pool bottom or the sheet of paper toward the sun.

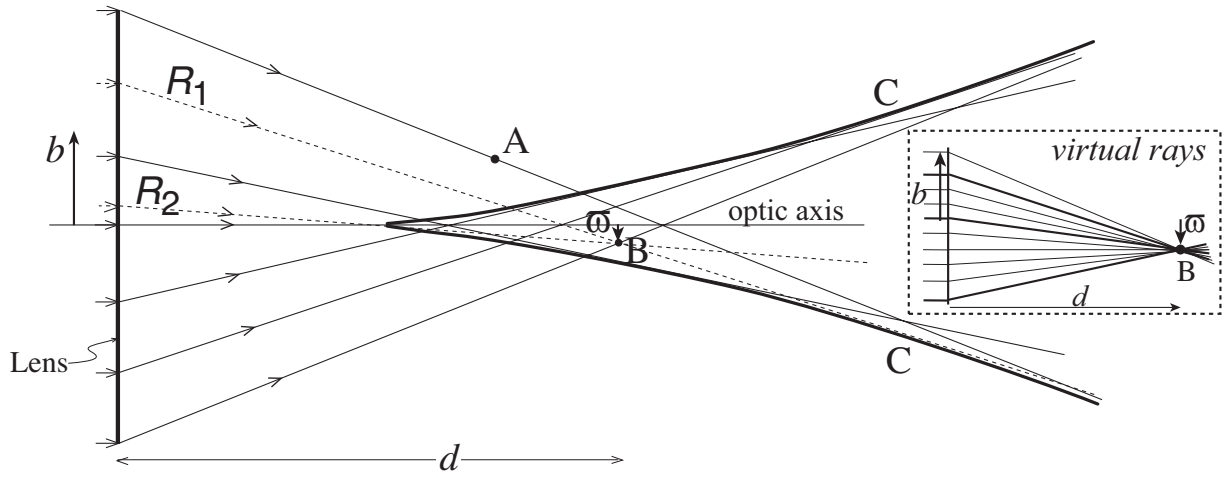


Fig. 7.12: The formation of caustics by a thin lens that is constrained to be circularly symmetric. Light from a distant source is refracted at the lens, and in the geometric optics limit, its rays (the arrowed lines, mostly solid but two dotted) bend toward the optic axis as shown. If the lens were perfect, the rays would all intersect at a point, the lens's focus. Because it is imperfect, they pass through the optic axis at different distances from the lens. The caustic $\mathcal{C}$ is the envelope of these rays. In the inset labeled "virtual rays", we show a construction used to explain the caustic. In this example, the caustic terminates in a cusp point, which becomes structurally unstable (ceases to be a cusp) if we remove the constraint that the lens be axisymmetric.

at $\mathcal{B}$. (That total φ will consist of a phase shift $\delta\varphi(b)$ produced by the lens, plus the phase change $k\ell(b;d,\varpi)$ due to propagation a distance ℓ from the lens to the point of observation $\mathcal{B}$.) By Fermat's principle, those virtual rays which extremize the total $\varphi(b;d,\varpi)$, when b is varied and d and ϖ are held fixed, are the true rays through $\mathcal{B}$.

To understand why, mathematically, there are three real rays [three extrema of $\varphi(b;d,\varpi)$] when $\mathcal{B}$ is inside the caustic $\mathcal{C}$ and only one when outside, we examine the behavior of φ as we cross this caustic, i.e. as $\mathcal{B}$'s radius ϖ passes through $\varpi_c(d)$, the radius of the caustic at longitudinal location d. From the big diagram in Fig. 7.12, it is clear that, as $\varpi \to \varpi_c$, the two disappearing images (along the dotted rays) approach one another and then vanish. Algebraically, this means that the curves $\varphi(b;d,\varpi)$ with fixed observation point (d,ϖ) must have the form shown in Fig. 7.13: as ϖ passes through ϖ_c, a maximum and a minimum smoothly merge through a point of inflexion and then vanish. It is clear that close enough to the caustic, $\varphi(b;d,\varpi)$, has the form of a cubic:

$$\varphi(b;d,\varpi) = \frac{1}{3}A(b-b_c)^3 - B(\varpi-\varpi_c)(b-b_c)\ , \tag{7.66}$$

where the factor $1/3$ is just a convention and b_c is some constant. For any given lens, we can compute the coefficients A, B accurately through a careful Taylor expansion about the caustic. However, their precise form does not interest us here. Invoking Fermat's principle and differentiating Eq. (7.66) with respect to b, we see that for $\varpi < \varpi_c$, there are two true rays and, passing through $b-b_c = \pm[B(\varpi-\varpi_c)/A)^{1/2}$, and two corresponding images; while for $\varpi > \varpi_c$ there are no rays or images. [In reality there is another ray, the solid-line ray

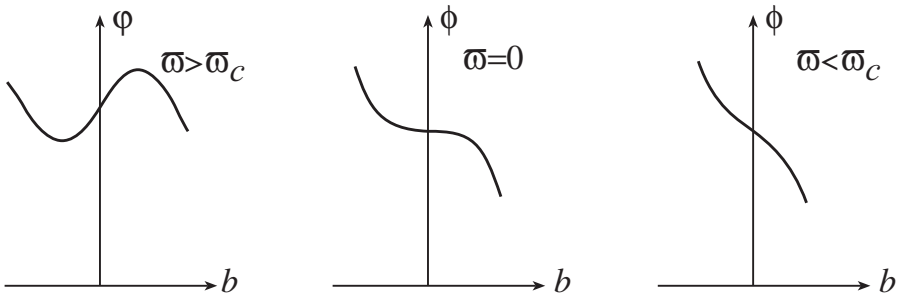


Fig. 7.13: Optical phase φ for three different near-caustic observer locations ϖ , with d fixed. The true rays are refracted at impact parameters b corresponding to extrema of the phase. There are two such real rays when the observer lies inside the caustic surface ($\varpi < \varpi_c$), and no such real rays when she lies outside ($\varpi > \varpi_c$). There is a actually one more real ray that is missed here because it is at b far from the near-caustic values. For the point $\mathcal{B}$ in Fig. 7.12, the real rays captured by this simple analysis are the dotted ones, and the one missed is the solid one through $\mathcal{B}$.

passing through $\mathcal{B}$ in Fig. 7.13. It has been lost from this analysis because we restricted our mathematics to the vicinity of the disappearing rays when we approximated φ as a cubic in b.)

Remarkably, as $\mathcal{B}$ approaches the caustic, its two merging images become brighter and brighter before suddenly disappearing. We can understand this by recalling that, in geometric optics, a wave's amplitude varies $\propto 1/(\text{area of a bundle of rays})^{1/2}$, so its energy flux varies $\propto 1/(\text{bundle's area})$. Now consider a ray bundle that hits the lens in the region b to $b+db$ and ϕ to $\phi+d\phi$, and then arrives at $\mathcal{B}$ in the range ϖ to $\varpi+d\varpi$, and (because ϕ is conserved along a ray) ϕ to $\phi+d\phi$. The ratio of the cross sectional areas of this bundle of rays is (area at lens)/(area at $\mathcal{B}$) $= (bd\phi db)/\varpi d\phi d\varpi = (b/\varpi)(db/d\varpi) \equiv \mathcal{M}$; and the observed energy flux is magnified by precisely this factor. As $\varpi \to \varpi_c$, the real rays have $b \to b_c$, and $db/d\varpi \propto 1/\sqrt{\varpi - \varpi_c}$. Therefore, *each of the merging images is magnified in energy flux, as it nears the caustic, by the factor*

$$\boxed{\mathcal{M} \propto \frac{db}{d\varpi} \propto \frac{1}{\sqrt{\varpi_c - \varpi}}}\,. \tag{7.67}$$

The scaling law $\mathcal{M} \propto 1/\sqrt{(\text{distance from caustic})}$ *is a general property of fold caustics* — equally true for this imperfect lens, for caustics in reflection at a spherical mirror, for the caustics formed by the rippled surface of the water in a swimming pool (Fig, 7.11a), for the fold portion (not the cusps) of the caustics formed by a water glass (Fig. 7.11), and even for caustics formed by gravitational lenses (next section). This is just one example of several scaling laws which apply to caustics.

The theory of optical caustics is a special case of a more general formalism called *catastrophe theory*, which deals with boundaries at which smooth behavior ceases and sudden jumps occur. Examples of such catastrophes, in addition to optical caustics, are phase transitions in statistical physics (e.g. Ex. 7.14 below), bifurcation of equilibria in elasticity theory (Sec. 11.8), sudden changes of behavior in nonlinear dynamical systems, and various sudden

changes that occur in biological systems and in human enterprises. For discussions see, e.g., Saunders (1980),

In catastrophe theory, it is shown that there are only a few types of catastrophe and they have many generic properties. The key mathematical requirement is that the behavior of a catastrophe should (by definition of "catastrophe") be *structurally stable*. This means that, if we make small changes in the physical conditions, its scaling laws etc. are robust.

The caustic that we have just considered is the most elementary example of a catastrophe and is called the *fold*. The next simplest catastrophe, known as the *cusp*, is the curve in 3D where two fold catastrophes meet. (The point cusp displayed in Fig. 7.13, is actually structurally unstable as a consequence of the assumed strict axisymmetry. However if b and ϖ were 1D Cartesian coordinates rather than cylindrical radii, then Fig. 7.13 would still correctly depict the rays' and caustic's geometry, and its cusp catastrophe would be structurally stable; it would be the cusp in the picture, extending into the picture to form two intersecting surfaces.)

In total there are five elementary catastrophes, when one deals with a two-parameter family of time-independent optical rays in three dimensions—for example, the rays from a distant point source of light that pass through a phase-shifting surface or thin layer, e.g. an imperfect lens or the wavy surface of a swimming pool. If the incoming light hits the surface orthogonally (for simplicity) and $\delta\varphi(a,b)$ is the phase shift produced by the surface at Cartesian-coordinate-location (a,b), and if the observation point is at Cartesian-coordinate-location (x,y,z) a distance $\ell(a,b;x,y,z)$ from point (a,b) on the surface, then the phase of the light arriving along a straight-line virtual ray from (a,b) to (x,y,z) is $\varphi(a,b;x,y,z) = \delta\varphi(a,b) + k\ell(a,b;x,y,z)$. In the vicinity of a structurally stable caustic, a thereom due to Réné Thom says that that there are only five possible independent functional forms for this $\varphi(a,b;x,y,z)$ (see Ex. 7.13) and they lead to the five caustic structures (catastrophes) shown in Fig. 7.14. If one adds a fourth dimension (e.g., time), then there are seven elementary catastrophes.

We conclude with a very important remark. If we have a point source of light, geometric optics predicts that the magnification will diverge to infinity as any caustic is approached, e.g. for a fold caustic, as $\mathcal{M} \propto 1/\sqrt{\text{distance from caustic}}$ [Eq. (7.67)]. Two factors prevent the magnification from becoming truly infinite. The first is that a point source is only an idealization, and if we allow the source to have finite size, different parts will produce caustics at slightly different locations. The second is that geometric optics, on which our analysis was based, pretends that the wavelength of light is vanishingly small. In actuality, the wavelength is always nonzero, and near a caustic its finiteness leads to diffraction effects, which limit the magnification to a finite value (Sec. 8.6). Diffraction is the subject of the next chapter.

****

EXERCISES

Exercise 7.13 ***Example: The Five Catastrophes in Optics*
Consider any two-parameter family of virtual light rays arriving, with phase φ, at an observation point near a caustic in 3-dimensional space. Denote by $(\xi_1,\xi_2) = (a,b)$ the ray

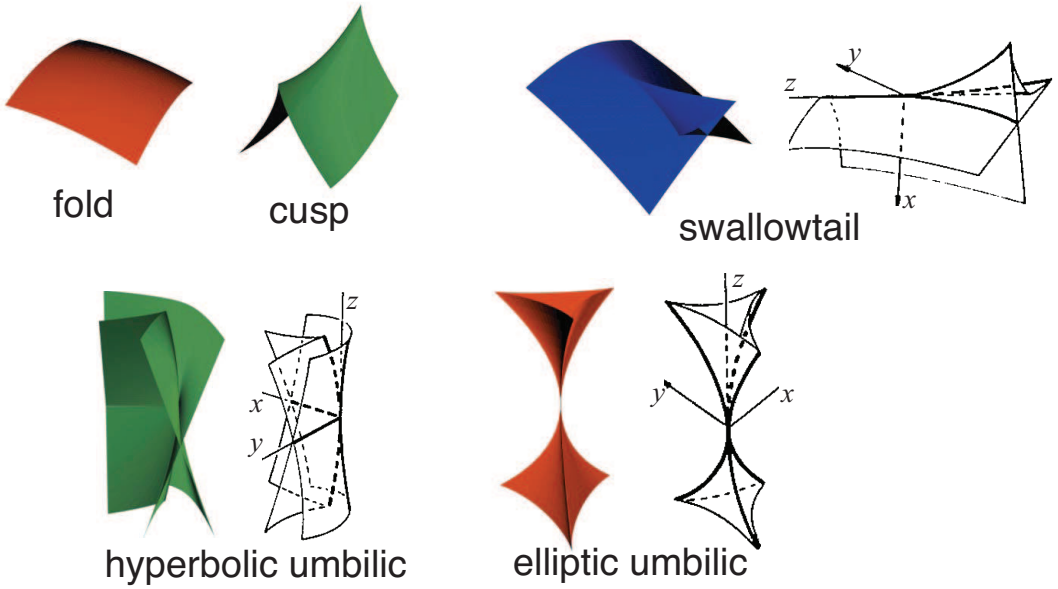


Fig. 7.14: The five elementary catastrophes (caustic structures) that are possible for a 2-parameter set of light rays in 3-dimensional space. For the three most complicated catastrophes, we show two depictions. When the light hits a screen, all bright caustic lines on the screen are made up of two-dimensional cross sections of these caustics. All drawings are adapted from Berry and Upstill (1980) or from an unpublished poster by Michael Berry: `http://www.phy.bris.ac.uk/people/berry_mv/pictures/poster1.pdf` .

parameters (called *state variables* in catastrophe theory), and by $(x_1, x_2, x_3) = (x, y, z)$ the Cartesian coordinates of the observation point, so the arriving light's phase is $\varphi(a, b, ; x, y, z)$. In this exercise we shall explore the five possible functional forms of this phase near a caustic, as identified by Thom's theorem, and their consequences for the caustic's (catastrophe's) structure..

(a) ***The Fold Catastrophe.*** Very near a fold catastrophe (caustic), with an appropriate parametrization and choice of coordinates (including absorbing some constants into a, b, z, y, z), the light's *phase* function is

$$\varphi = \frac{1}{3}b^3 - xb \; . \tag{7.68}$$

Notice that this is identical to the phase near the caustic of the lens treated in the text (Fig. 7.13): it is the same as Eq. (7.66) with $A = B = 1$, $b_c = \varpi_c = 0$ and $\varpi = -x$. In this case, the phase is independent of a, y, z. (i) Following the procedure used in the text, infer from Fermat's principle that the rays (if any) that pass through location x have $b = \pm\sqrt{x}$. This is the *ray map* in the language of optics; it maps the observational location x onto the values of b that label the real rays passing through x. (ii) Depict the ray map by plotting b upward and x horizontally, restricting attention to real b and x since by definition both must be real. This *equilibrium surface* (as it is called in catastrophe theory) folds over at $x = 0$, which is why this catastrophe is called the

fold. (iii) As this plot and its equation $b = \pm\sqrt{x}$ show, there are two rays (two values of b) for $x > 0$ and none for $x < 0$. Therefore, $x = 0$ — the location of the fold — is the caustic for all y, z, which means that the caustic is a plane as depicted in Fig. 7.14.

(b) ***The Cusp Catastrophe.*** For the cusp catastrophe, the phase function is

$$\varphi = \frac{1}{4}b^4 - \frac{1}{2}xb^2 - yb \; . \tag{7.69}$$

Because φ is independent of z, the caustic will be translation invariant in the z direction. (i) From Fermat's principle infer that the ray map (equilibrium surface) is the set of solutions $b(x, y)$ to the equation $b^3 - xb - y = 0$. Because this is a cubic in b and rays can only be created and disappear in pairs, there will be some regions of space through which pass three rays, and some, one. (ii) Plot the equilibrium surface $b(x, y)$ (using a computer). From the plot, identify visually the region with three rays and the region with one, and the caustic that separates them. (iii) Show analytically that the caustic is the cusp $y = \pm 2(x/3)^{3/2}$. Verify visually that this agrees with your plot in (ii), and with the shape depicted in Fig. 7.14.

(c) ***The Swallowtail Catastrophe.*** For the swallowtail catastrophe, the phase function is

$$\varphi = \frac{1}{5}b^5 - \frac{1}{3}xb^3 - \frac{1}{2}yb^2 - zb \; . \tag{7.70}$$

(i) From Fermat's principle infer that the ray map is the set of solutions $b(x, y, z)$ to the equation $b^4 - xb^2 - yb - z = 0$. Because this is a quartic, there can be regions of space with 4, 2, and 0 rays passing through each event. It is not practical to plot the "equilibrium surface" $b(x, y, z)$, since that would be a four dimensional plot. (ii) Show that the caustic, at which the number of rays changes, is given by $z = b^4 - xb^2 - yb$ with $b(x, y)$ the solution to $4b^3 - 2xb - y = 0$. Draw this caustic surface with a computer and verify that it agrees with the swallowtail shown in Fig. 7.14.

(d) ***The Hyperbolic Umbilic Catastrophe.*** For the hyperbolic umbilic catastrophe, the phase function depends on two ray parameters $\xi_1 = a$, $\xi_2 = b$, and on all three spatial coordinates x, y, z:

$$\varphi = \frac{1}{3}(a^3 + b^3) - zab - xa - yb \; . \tag{7.71}$$

(i) Explain why Fermat's principle dictates that the ray map be given by $\partial\varphi/\partial\xi_j = 0$, which is two equations. Show that the resulting ray map $\{a(x, y, z), b(x, y, z)\}$ is given by the solution to $a^2 - zb - x = 0$ and $b^2 - za - y = 0$. (ii) Explain why the caustic — the points in space at which the number of rays changes by two — is given by eliminating the parameters ξ_i from the following three equations: $\partial\varphi/\partial\xi_i = 0$ and $\det[\partial^2\varphi/\partial\xi_i\partial\xi_j] = 0$. Show that these three equations are $a^2 - zb - x = 0$, $b^2 - za - y = 0$, and $4ab - z^2 = 0$, with a and b to be eliminated. Plot the resulting $z(x, y)$ and show that it agrees with the hyperbolic umbilic shape depicted in Fig. 7.14

(e) ***The Elliptic Umbilic Catastrophe.*** For the elliptic umbilic catastrophe, the phase function is:

$$\varphi = \frac{1}{3}a^3 - ab^2 - z(a^2+b^2) - xa - yb\ . \tag{7.72}$$

Following the procedure developed in part (d), show that the caustic is the function $z(x,y)$ obtained by eliminating a and b from the three equations $a^2 - b^2 - x - 2za = 0$, $2ab + y + 2zb = 0$, and $b^2 + a^2 - z^2 = 0$. Plot the resulting $z(x,y)$ and show that it agrees with the elliptic umbilic shape depicted in Fig. 7.14

Exercise 7.14 ***Example: Van der Waals Catastrophe*

The van der Waals equation of state $(P + a/v^2)(v - b) = k_B T$ for H_2O relates the pressure P and specific volume (volume per molecule) v to the temperature T; see Sec. 5.7. Figure 5.8 makes it clear that, at some temperatures and pressures T and P, there are three allowed volumes $v(T,P)$, one describing liquid water, one water vapor, and the third an unstable phase that cannot exist in Nature. At other T, P, there is only one allowed v. The transition between three v's and one occurs along some curve in the T, P plane—a catastrophe curve.

(a) This curve must correspond to one of the elementary catastrophes explored in the previous exercise. Based on the number of solutions for $v(T,P)$, which catastrophe must it be?

(b) Change variables, in the van der Waals equation of state, to $p = P/P_c - 1$, $t = T/T_c - 1$, $\rho = V_c/V - 1$, where $T_c = 8a/27bk_B$, $P_c = a/27b^2$, $v_c = 3b$ are the temperature, pressure, and specific volume at the critical temperature (the temperature below which liquid water and water vapor can exist as separate phases; Sec. 5.7). Show that this change of variables brings the van der Waals equation of state into the form

$$\rho^3 - x\rho - y = 0\ , \quad \text{where} \quad x = -(p/3 + 8t/3)\ , \quad y = 2p/3 - 8t/3\ . \tag{7.73}$$

(c) This equation $\rho^3 - x\rho - y$ is the equilibrium surface associated with the catastrophe-theory potential $\varphi(\rho; x, y) = \frac{1}{4}\rho^4 - \frac{1}{2}x\rho^2 - y\rho$ [Eq. (7.69)]. Correspondingly, the catastrophe [the boundary between three solutions $v(T,P)$ and one] has the universal cusp form $x = \pm 2(y/3)^{2/3}$. Plot this curve in the temperature-pressure plane.

Note: We were guaranteed by catastrophe theory that the catastrophe curve would have this cusp form near its cusp point. However, it is a surprise and quite unusual that, for the van der Waals case, the cusp shape $x = \pm 2(y/3)^{2/3}$ is not confined to the vicinity of the cusp point, but remains accurate far from the cusp point.

****

7.6 T2 Gravitational Lenses; Their Multiple Images and Caustics

Albert Einstein's general relativity theory predicts that light rays should be deflected by the gravitational field of the Sun (Ex. 27.3; Sec. 27.2.3). Newton's law of gravity combined with his corpuscular theory of light also predicts this deflection, but through an angle half as great as relativity predicts. A famous measurement, during a 1919 solar eclipse, confirmed the relativistic prediction, thereby making Einstein world famous.

The deflection of light by gravitational fields allows a distant galaxy to behave like a crude lens and, in particular, to produce multiple images of a more distant quasar. Many examples of this phenomenon have been observed. The optics of these gravitational lenses provides an excellent illustration of the use of Fermat's principle.[9] We shall explore these issues in this section.

7.6.1 T2 Refractive-Index Model of Gravitational Lensing

The action of a gravitational lens can only be understood properly using general relativity. However, when the gravitational field is weak, there exists an equivalent Newtonian model that is adequate for our purposes. In this model, curved spacetime behaves as if it were spatially flat and endowed with a refractive index given by

$$\boxed{\mathfrak{n} = 1 - \frac{2\Phi}{c^2}}\,, \tag{7.74}$$

where Φ is the Newtonian gravitational potential, normalized to vanish far from the source of the gravitational field and chosen to have a negative sign (so, e.g., the field at a distance r from a point mass M is $\Phi = -GM/r$). Time is treated in the Newtonian manner in this model. In Sec. 27.2.3, we will use a general relativistic version of Fermat's principle to show that for static gravitational fields this index-of-refraction model gives the same predictions as general relativity, up to fractional corrections of order $|\Phi|/c^2$, which are $\lesssim 10^{-5}$ for the lensing examples in this chapter.

There is a second Newtonian model that gives the same predictions as this index-of-refraction model, to within its errors, $\sim |\Phi|/c^2$. We deduce it by rewriting the ray equation (7.47) in terms of Newtonian time t using $ds/dt = C = c/\mathfrak{n}$:

$$\frac{d^2\mathbf{x}}{dt^2} = \mathfrak{n}c^2\boldsymbol{\nabla}\mathfrak{n} = -\boldsymbol{\nabla}2\Phi\ . \tag{7.75}$$

Here in the last expression we have used Eq. (7.74) and have replaced the multiplicative factor $\mathfrak{n}$, in the second expression, by unity, thereby making a fractional error $\sim |\Phi|/c^2$ of the same magnitude as the errors in our index-of-refraction model. Equation (7.75) says that *the photons that travel along rays feel a Newtonian gravitational potential that is twice as large as the potential felt by low-speed particles; and the photons, moving at a speed that is c aside from fractional changes of order $|\Phi|/c^2$, respond to that doubled Newtonian field in the same way as any Newtonian particle would.*

[9] Schneider, Ehlers & Falco (1999).

7.6.2 T2 Lensing by a Point Mass

We shall explore the predictions of this model, beginning with gravitational lensing by a point mass M, with $\Phi = -GM/r$; Fig. 7.15. In this case, there is an obvious optic axis: the line between the earth $\oplus$ and the lens M; the index of refraction is axisymmetric around that axis, so the ray from the source Q to earth $\oplus$ will lie in a plane of constant aziumthal angle ϕ (and $\phi + \pi$)—the plane shown in Fig. 7.15. Moreover, the ray will bend near the source, sharply as seen on the lengthscales used in the figure. Therefore, the point mass's gravity acts like an axisymmetric phase-shifting surface, i.e. an imperfect axisymmetric lens, in the language of the Sec. 7.5.

Computing the ray (photon trajectory) from Eq. (7.75) with $2\Phi = -2GM/r$ is equivalent, of course, to computing the deflection of a charged particle passing by an oppositely charged, far more massive particle. That computation gives a ray deflection angle

$$\alpha = \frac{4GM}{bc^2} = \frac{-4\Phi(r=b)}{c^2} \ , \tag{7.76a}$$

where b is the ray's impact parameter; see Ex. 7.15. For a ray from a distant star, passing close to the limb of the sun, this deflection is $\alpha = 1.75$ arc seconds.

In Fig. 7.15, we identify the location of the point source Q by its angle θ' to the optic axis, as seen from Earth in the absence of the lensing mass. The lensing pushes the image away from the optic axis, to a larger angle θ. Elementary geometry plus Eq. (7.76a) reveals the following relationship between θ and θ' (Ex. 7.15):

$$\theta' = \theta - \frac{{\theta_E}^2}{\theta} \ , \quad \text{where} \quad \theta_E \equiv \sqrt{\frac{4Mu}{v(u+v)}} \tag{7.76b}$$

is called the Einstein (angular) radius. (Here u and v are the distances shown in Fig. 7.15.) This is a quadratic equation for θ in terms of θ' and θ_E, and it has two solutions:

$$\theta_\pm = \frac{\theta'}{2} \pm \sqrt{{\theta_E}^2 + \left(\frac{\theta'}{2}\right)^2} \ . \tag{7.76c}$$

The ray shown in Fig. 7.15 produces the image at $\theta = \theta_+$. A ray, not shown there, which passes below the lensing mass M and gets deflected upward, produces a second image, at the negative (upward sloped) angle $\theta = \theta_-$. If the source of gravity were extended, rather than being a point mass, there would be a third image, along a ray passing through the source.

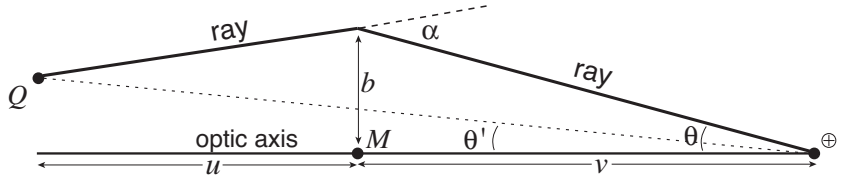


Fig. 7.15: Geometry for gravitational lensing of a point source Q by a point mass M.

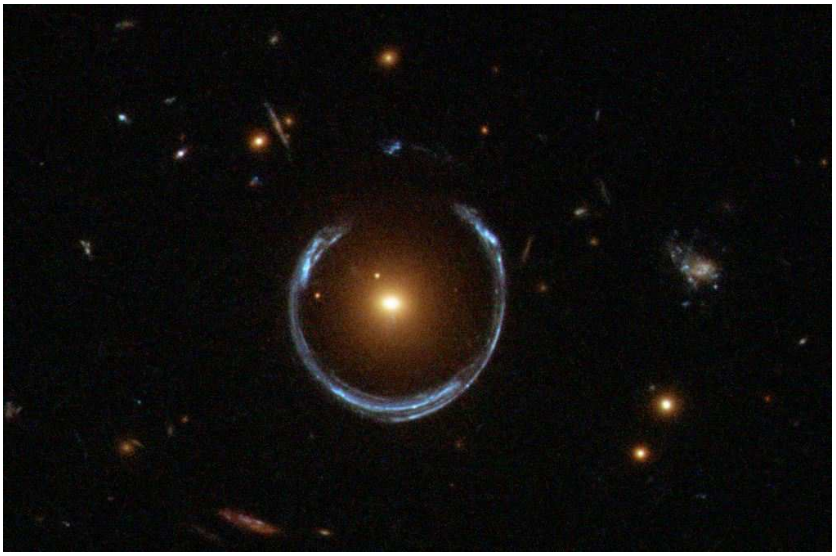


Fig. 7.16: The Einstein ring LRG 3-757, also called "the cosmic horseshoe", photographed by the Hubble Space Telescope. The ring light comes from a very distant galaxy, with cosmological redshift $z \simeq 2.4$. The lensing galaxy is the bright yellow spot at the center of the ring.

As the source Q is moved downward toward the optic axis so $\theta' \to 0$, the two images move toward $\pm\theta_E$, i.e. opposite edges of a circle with angular radius Θ_E; and by axisymmetry, the images are then suddenly transformed into a thin ring of light, the so-called Einstein ring. An astronomical photograph of such an Einstein ring is shown in Fig. 7.16. When the source is moved onward to negative θ', the ring breaks back up into two images, again on opposite sides of the former ring, one just outside the ring and the other just inside it.

This behavior is very different from that at a generic caustic (Sec. 7.5), where the images would coalesce then disappear. This behavior is not structurally stable; if the axisymmetry is broken, then the behavior will change significantly. Nevertheless, for finite-sized sources of light and generic sources of gravity (e.g. compact stars or galaxies; next subsection), there can be structurally stable Einstein rings, as Fig. 7.16 demonstrates.

Suppose that the light source Q is finite in size, as always is the case in Nature. Denote by $\Delta\Omega'$ the solid angle it would subtend as seen from earth in the absence of the lens, and by $\Delta\Omega$ its solid angle in the lens's presence. The total energy flux received at earth from the image is $dE/dtdA = (\int I_\nu d\nu)\Delta\Omega$, where $I_\nu = dE/dtdAd\nu d\Omega$ is the spectral intensity traveling along the ray. Kinetic theory tells us that I_ν/ν^3 is conserved along the ray (end of Sec. 3.6), and because gravity is weak, so is ν and thence so is $\int I_\nu d\nu = I$. In the presence of the lens, then, the received flux is $dE/dtdA = I\Delta\Omega$, and in the absence of the lens, it would be $dE/dtdA = I\Delta\Omega'$. The ratio of the energy fluxes is the magnification,

$$\boxed{\mathcal{M} = \Delta\Omega/\Delta\Omega'} \,. \tag{7.76d}$$

This is a very general result, as scrutiny of the derivation shows: it holds true, in the geomtric optics approximation, whenever the frequencies of photons are the same when they reach an observer, as when they left their source.

We can compute this magnification most easily by focusing not on the lensing of the source's total surface, but rather the lensing of an elementary angular element on the surface with azimuthal extent $d\phi$ and poloidal extent $d\theta$. By axisymmetry around the optic axis, $d\phi$ is unaffected by the lens, so the magnification is (Ex. 7.15)

$$\mathcal{M} = \frac{d\Omega}{d\Omega'} = \frac{\theta\, d\phi\, d\theta}{\theta'\, d\phi\, d\theta'} = \frac{\theta}{\theta'}\frac{d\theta}{d\theta'} = \frac{1}{1-(\theta_E/\theta)^4}\ . \tag{7.76e}$$

Here we have used Eq. (7.76b) to express θ' in terms of θ. The image at $\theta_+ > \theta_E$ has positive magnification, which means it has the same parity as the source; the image at $\theta_- < \theta_E$ has negative magnification, which means opposite parity to the source.

Notice that, as the source is moved toward and onto the optic axis, so $\theta \to \theta_E$, the magnification diverges. This is a signal that our geometric optics approximation is breaking down; cf. the last paragraph (before the exercises) of Sec. 7.5. From Eqs. (7.76c) and (7.76e), it can be shown (Ex. 7.15) that the ratio of magnifications of the two images is

$$R = \frac{\mathcal{M}_+}{\mathcal{M}_-} = -\left(\frac{\theta_+}{\theta_-}\right)^2\ . \tag{7.76f}$$

****

EXERCISES

Exercise 7.15 *Derivation: Point-Mass Gravitational Lens*

For the point-mass gravitational lens geometry shown in Fig. 7.15, derive Eqs. (7.76). [Incidentally, Einstein performed a calculation similar to this, in one of his unpublished notebooks, prior to understanding general relativity.]

****

7.6.3 T2 Lensing of a Quasar by a Galaxy

When a distant quasar is lensed by a more nearby galaxy as in Fig. 7.17, the images' rays typically pass through the galaxy, not around it. We can see this by the following rough estimate: The virial theorem (which says that the kinetic energy of a swarm of stars is half the magnitude of its potential energy) tells us that the gravitational potential at a ray's point of closest approach to the galaxy center is $|\Phi| = GM/b \sim 2 \times \frac{3}{2}\sigma^2$. Here σ^2 is the stars' mean square velocity in one dimension so $\frac{3}{2}\sigma^2$ is their mean kinetic energy per unit mass. Correspondingly, the deflection angle is $\alpha = 4|\Phi|/c^2 \sim 12\sigma^2/c^2$. For typical galaxies, $\sigma \sim 300$ km s^{-1} and this formula gives $\alpha \sim 1-2$ arc sec. Now, the distances u, v are roughly ten billion light years $\sim 10^{26}$m and so the transverse displacement of the ray due to the galaxy is $\sim v\alpha/2 \sim 3 \times 10^{20}$m $\sim 30,000$ light years (about the distance of the earth from the center of our Milky Way galaxy), which is well within the galaxy. This means that light

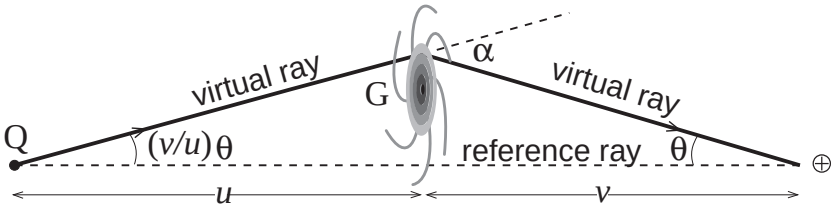


Fig. 7.17: Geometry for gravitational lensing of a quasar Q by a large galaxy G.

from a quasar lying behind the galaxy can pass through the outer regions of either side of the galaxy. We should then see at least two distinct images of the quasar, separated by an angular distance $\sim \frac{1}{2}\alpha \sim 1$ arc sec.

We shall now analyze the lensing quantitatively for the geometry of Fig. 7.17.[10] As for a point source of gravity, so also for a lensing galaxy, because the distances from galaxy to earth and from galaxy to quasar are so large, the galaxy's gravity will act like a phase shifting screen. We shall analyze its influence using Fermat's principle, in the same way as we did in Sec. 7.5 for an imperfect lens, which also behaved like a phase shifting screen.

First we trace a ray backward from the observer, in the absence of the intervening galaxy, to the quasar. We call this the reference ray. Next, we reinstate the galaxy and consider a *virtual ray*[8] that extends (backward) at an angle $\boldsymbol{\theta}$ (a 2-dimensional vector on the sky) to the reference ray in a straight line from the earth to the galaxy, where it is deflected along a straight line to the quasar. The optical phase for light propagating along this virtual ray will exceed that along the reference ray by an amount $\Delta\varphi$ called the *phase delay*. There are two contributions to $\Delta\varphi$: First, the geometrical length of the path is longer than the reference ray by an amount $(u+v)v\theta^2/2u$ (cf. Fig. 7.17), and thus the travel time is longer by an amount $(u+v)v\theta^2/2uc$. Second, the light is delayed as it passes through the potential well by a time $\int(\mathfrak{n}-1)ds/c = -2\int\Phi ds/c^3$, where ds is an element of length along the path. We can express this second delay as $2\Phi_2/c^3$. Here

$$\boxed{\Phi_2 = \int \Phi ds} \tag{7.77}$$

is a *two-dimensional (2D) Newtonian potential* and can be computed from the 2D Poisson equation

$$\boxed{\nabla^2\Phi_2 = 4\pi G\Sigma\ , \quad \text{where } \Sigma = \int \rho ds} \tag{7.78a}$$

is the surface density of mass in the galaxy integrated along the line of sight.

[10] In our treatment of this lensing example, we shall ignore complications that arise from the universe's expansion and its large scale gravitational fields. However, as we discuss in Chap. 28, the universe is spatially flat and so the relation between angles and lengths is Euclidean, as in our refractive-index model. We can also imagine making the lens observations after the universe has expanded to its present age everywhere, and stopping the expansion then so as to measure the distances u, v. If we use this definition of distances, it turns out that we can ignore cosmological effects in analyzing the optics. Our index-of-refraction model is then accurate, as in non-cosmological contexts, up to fractional errors $\sim |\Phi|/c^2$.

Therefore, the phase delay $\Delta\varphi$ is given by

$$\Delta\varphi = \omega\left(\frac{(u+v)v}{2uc}\theta^2 - \frac{2\Phi_2(\boldsymbol{\theta})}{c^3}\right) . \tag{7.78b}$$

We can now invoke Fermat's principle. *Of all possible virtual rays, parametrized by the angular coordinate $\boldsymbol{\theta}$ seen from earth , the only ones that are real rays are those for which the phase difference is stationary, i.e. those for which*

$$\frac{\partial\Delta\varphi}{\partial\theta_j} = 0 , \tag{7.78c}$$

where θ_j (with $j = x, y$) are the Cartesian components of $\boldsymbol{\theta}$. Differentiating Eq. (7.78b) we obtain *a 2D vector equation for the angular location of the images as viewed from Earth:*

$$\boxed{\theta_j = \frac{2u}{(u+v)vc^2}\frac{\partial\Phi_2}{\partial\theta_j}} . \tag{7.78d}$$

Note that $\partial\Phi_2/\partial\theta_j$ is a function of θ_j, so if $\Phi_2(\theta_j)$ is known, this becomes a (usually) nonlinear equation to be solved for the vector θ_j. Referring to Fig. 7.17, and using simple geometry, we can identify the *deflection angle for the image's real ray:*

$$\boldsymbol{\alpha}_j = \frac{2}{vc^2}\frac{\partial\Phi_2}{\partial\theta_j} . \tag{7.78e}$$

We can understand quite a lot about the properties of the images by inspecting a *contour plot of the phase delay function $\Delta\varphi(\boldsymbol{\theta})$* (Fig. 7.18). Recall [sentences preceding Eq. (7.44a)] that the images appear at the stationary points of $\Delta\varphi$. When the galaxy is very light or quite distant from the line of sight, then there is a single minimum in the phase delay near $\theta_j = 0$ (the reference ray); Figs. 7.18a,b . However, a massive galaxy along the line of sight to the quasar can create two or even four additional stationary points and therefore three or five images (Figs. 7.18c,d). Note that with a transparent galaxy, the additional images are created in pairs. Note, in addition, that the stationary points are not necessarily minima. They can be minima labeled by L in the figure, maxima labeled by H, or saddle points labeled by S. This is inconsistent with Fermat's original statement of his principle ("minimum phase delay"), but there are images at all the stationary points nevertheless.

We can compute the magnifications of the quasar's images by the obvious generalization of the way we did so for a point-mass lens [Eq. (7.76d)]. We let the quasar have a finite but small size, and we focus on a small element of its surface that would subtend a solid angle $d\Omega'$ as seen from earth, if the lensing galaxy were absent. We then compute the solid angle $d\Omega$ subtended by the image of this element. Because the photon frequencies are negligibly affected by their travel, the magnification of the received energy flux is $\mathcal{M} = d\Omega/d\Omega'$.

The foundation for computing the ratio $d\Omega/d\Omega'$ is the mapping of an angular vector $\delta\boldsymbol{\theta}'$ in the quasar's surface into the corresponding vector $\delta\boldsymbol{\theta}$ in the image. We find this mapping by imagining displacing an idealized point-source quasar by a small angle $\delta\boldsymbol{\theta}'$ as seen from Earth in the absence of the lens. This is equivalent to moving the lens by a small angle $-\delta\boldsymbol{\theta}'$

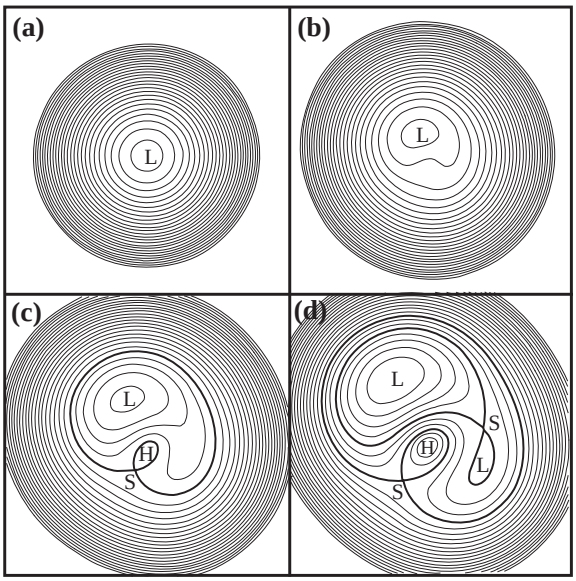


Fig. 7.18: Contour plots of the phase delay $\Delta\varphi(\boldsymbol{\theta})$ for four different gravitational lenses. (a) In the absence of a lens $\Phi_2 = 0$, the phase delay (7.78b) has a single minimum corresponding to a single undeflected image, in the direction of the reference ray $\theta_j = 0$. (b) When a small galaxy with a shallow potential Φ_2 is interposed, it pushes the phase delay $\Delta\varphi$ up in its vicinity [Eq. (7.78b) with negative Φ_2], so the minimum and hence the image are deflected slightly away from the galaxy's center. (c) When a galaxy with a deeper potential well is present, the delay surface will be raised so much near the galaxy's center that additional stationary points will be created, at H and S, and two more images will be produced. (d) If the potential well deepens even more, five images can be produced. In all four plots the local extrema of $\Delta\varphi$ are denoted L for a low point (local minimum), H for a high point (local maximum) and S for saddle point.

as seen from Earth. Equation (7.78d) says that *the image will be displaced by a small angle* $\delta\boldsymbol{\theta}$ satisfying the equation

$$\delta\theta_i - \delta\theta'_i = \frac{2u}{(u+v)vc^2}\frac{\partial^2\Phi_2}{\partial\theta_i\partial\theta_j}\delta\theta_j \ . \tag{7.79a}$$

By combining with Eq. (7.78b), we can rewrite this as

$$\delta\theta'_i = H_{ij}\delta\theta_j \ , \tag{7.79b}$$

where the matrix $[H_{ij}]$ is

$$\boxed{H_{ij} = \left(\frac{uc/\omega}{(u+v)v}\right)\frac{\partial^2\Delta\varphi}{\partial\theta_i\partial\theta_j} = \delta_{ij} - \frac{2u}{(u+v)vc^2}\Phi_{2,ij}} \ . \tag{7.79c}$$

This is the vector mapping we need.

It is a standard result in geometry (fairly easily derived) that, because the lensing maps any $\delta\theta'_i$ into $\delta\theta_i = H^{-1}_{ij}\delta\theta'_j$ [inverse of the mapping (7.79b)], it maps any infinitesimal solid

angle $d\Omega'$ into $d\Omega = \det[H_{ij}^{-1}]d\Omega' = d\Omega'/\det[H_{ij}]$. Therefore, the flux magnification is

$$\boxed{\mathcal{M} = \frac{d\Omega}{d\Omega'} = \frac{1}{\det\left[H_{ij}\right]} = \frac{(u+v)v}{uc/\omega}\frac{1}{\det\left[\partial^2\Delta\varphi/\partial\theta_i\partial\theta_j\right]}]}. \tag{7.80}$$

The curvature of the phase delay surface (embodied in $\det\left[\partial^2\Delta\varphi/\partial\theta_i\partial\theta_j\right]$) is therefore a quantitative measure of the magnification. Small curvature implies large magnification of the images and *vice versa*. Caustics are located where the curvature goes to zero, i.e. where $\det\left[\partial^2\Delta\varphi/\partial\theta_i\partial\theta_j\right] = 0$ (cf. Ex. 7.13d). Furthermore, images associated with saddle points in the phase delay surface have opposite parity to the source. Those associated with maxima and minima have the same parity as the source, although the maxima are rotated on the sky by 180°. These effects have been seen in observed gravitational lenses.

There is an additional, immediate contact to the observations and this is that *the phase delay function $\Delta\varphi$ at the stationary points is equal to ω times the extra time it takes a signal to arrive along that ray.* In order of magnitude, the time delay difference will be $\sim v\alpha^2/8c \sim 1$ month [cf. Eq. (7.78b)]. Now, many quasars are intrinsically variable, and if we monitor the variation in two or more images, then we should be able to measure the time delay between the two images; cf. Fig. 7.19. This, in turn, allows us to measure the distance to the quasar and, consequently, provides a measurement of the size of the universe.

****

EXERCISES

Exercise 7.16 *Challenge: Catastrophe Optics of an Elliptical Gravitational Lens*

Consider an elliptical gravitational lens where the potential at the lens plane varies as

$$\Phi_2(\boldsymbol{\theta}) = (1 + A\theta_1^2 + 2B\theta_1\theta_2 + C\theta_2^2)^q; \qquad 0 < q < 1/2.$$

Determine the generic form of the caustic surfaces and the types of catastrophe encountered. Note that it is in the spirit of catastrophe theory *not* to compute exact expressions but to determine scaling laws and to understand the qualitative behavior of the images.

****

7.7 Polarization

In our geometric optics analyses thus far, we have either dealt with a scalar wave (e.g., a sound wave) or simply supposed that individual components of vector or tensor waves can be treated as scalars. For most purposes, this is indeed the case, and we shall continue to use this simplification in the following chapters. However, there are some important

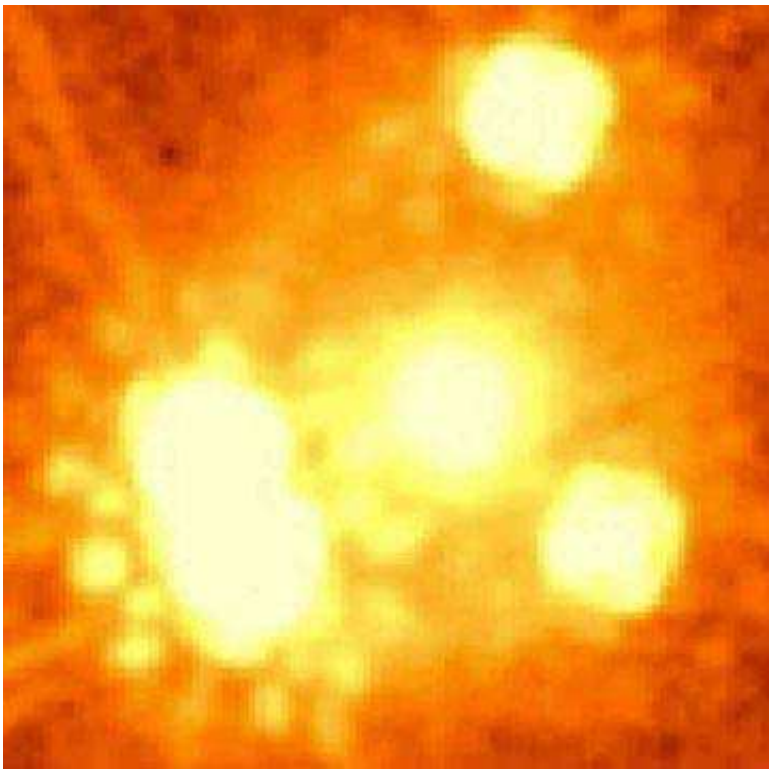


Fig. 7.19: Gravitational lens in which a distant quasar, Q1115+080, is quadruply imaged by an intervening galaxy. (There is also a fifth, unseen image.) Two of the images are observed to be much brighter and closer to each other than the other two, because the quasar is located quite close to a fold caustic surface. When the brightness of the quasar fluctuates, the four images are observed to fluctuate in the order predicted by modeling the gravitational potential of the lensing galaxy, and these fluctuations can be used to estimate the size of the universe.

wave properties that are unique to vector (or tensor) waves. Most of these come under the heading of *polarization* effects. In Secs. 27.3 and 27.4, we shall study polarization effects for (tensorial) gravitational waves. Here and in Chaps. 10 and 21–23, we shall examine them for electromagnetic waves.

An electromagnetic wave's two polarizations are powerful tools for technology, engineering, and experimental physics. However, we shall forgo any discussion of this in the present chapter, and instead shall focus solely on the geometric-optics propagation law for polarization (Sec. 7.7.1), and an intriguing aspect of it called the *geometric phase* (Sec. 7.7.2).

7.7.1 Polarization Vector and its Geometric-Optics Propagation Law

A plane electromagnetic wave *in vacuo* has its electric and magnetic fields **E** and **B** perpendicular to its propagation direction $\hat{\mathbf{k}}$ and perpendicular to each other. In a medium, **E** and **B** may or may not remain perpendicular to $\hat{\mathbf{k}}$, depending on the medium's properties. For example, an Alfvén wave has its vibrating magnetic field *perpendicular* to the background magnetic field, which can make an arbitrary angle with respect to $\hat{\mathbf{k}}$. By contrast, in the

simplest case of an *isotropic* dielectric medium, where the dispersion relation has our standard dispersion-free form $\Omega = (c/\mathfrak{n})k$, the group and phase velocities are parallel to $\hat{\mathbf{k}}$, and $\mathbf{E}$ and $\mathbf{B}$ turn out to be perpendicular to $\hat{\mathbf{k}}$ and to each other—as in vacuum. In this section, we shall confine attention to this simple situation, and to linearly polarized waves, for which $\mathbf{E}$ oscillates linearly back and forth along a polarization direction $\hat{\mathbf{f}}$ that is perpendicular to $\hat{\mathbf{k}}$:

$$\mathbf{E} = Ae^{i\varphi}\,\hat{\mathbf{f}}\ , \quad \hat{\mathbf{f}}\cdot\hat{\mathbf{k}} \equiv \hat{\mathbf{f}}\cdot\boldsymbol{\nabla}\varphi = 0\ . \tag{7.81}$$

In the eikonal approximation, $Ae^{i\varphi} \equiv \psi$ satisfies the geometric-optics propagation laws of Sec. 7.3, and the polarization vector $\hat{\mathbf{f}}$, like the amplitude A, will propagate along the rays. The propagation law for $\hat{\mathbf{f}}$ can be derived by applying the eikonal approximation to Maxwell's equations, but it is easier to infer that law by simple physical reasoning: (i) Since $\hat{\mathbf{f}}$ is orthogonal to $\hat{\mathbf{k}}$ for a plane wave, it must also be orthogonal to $\hat{\mathbf{k}}$ in the eikonal approximation (which, after all, treats the wave as planar on lengthscales long compared to the wavelength). (ii) If the ray is straight, then the medium, being isotropic, is unable to distinguish a slow right-handed rotation of $\hat{\mathbf{f}}$ from a slow left-handed rotation, so there will be no rotation at all: $\hat{\mathbf{f}}$ will continue always to point in the same direction, i.e. $\hat{\mathbf{f}}$ will be kept parallel to itself during transport along the ray. (iii) If the ray bends, so $d\hat{\mathbf{k}}/ds \neq 0$ (where s is distance along the ray), then $\hat{\mathbf{f}}$ will have to change as well, so as always to remain perpendicular to $\hat{\mathbf{k}}$. The direction of $\hat{\mathbf{f}}$'s change must be $\hat{\mathbf{k}}$, since the medium, being isotropic, cannot provide any other preferred direction for the change. The magnitude of the change is determined by the requirement that $\hat{\mathbf{f}}\cdot\hat{\mathbf{k}}$ remain zero all along the ray and that $\hat{\mathbf{k}}\cdot\hat{\mathbf{k}} = 1$. This immediately implies that the propagation law for $\hat{\mathbf{f}}$ is

$$\boxed{\frac{d\hat{\mathbf{f}}}{ds} = -\hat{\mathbf{k}}\left(\hat{\mathbf{f}}\cdot\frac{d\hat{\mathbf{k}}}{ds}\right)}\ . \tag{7.82}$$

This equation says that the vector $\hat{\mathbf{f}}$ is *parallel-transported along the ray.* Here "parallel transport" means: (i) Carry $\hat{\mathbf{f}}$ a short distance along the ray, keeping it parallel to itself in 3-dimensional space. Because of the bending of the ray and its tangent vector $\hat{\mathbf{k}}$, this will cause $\hat{\mathbf{f}}$ to no longer be perpendicular to $\hat{\mathbf{k}}$. (ii) Project $\hat{\mathbf{f}}$ perpendicular to $\hat{\mathbf{k}}$ by adding onto it the appropriate multiple of $\hat{\mathbf{k}}$. (The techniques of differential geometry for curved lines and surfaces, which we shall develop in Chaps. 24 and 25 in preparation for studying general relativity, give powerful mathematical tools for analyzing this parallel transport.)

7.7.2 T2 Geometric Phase

We shall use the polarization propagation law (7.82) to illustrate a quite general phenomenon known as the *geometric phase.* For further details and some history of this concept, see Berry (1990).

Consider, as a simple context for the geometric phase, a linearly polarized, monochromatic light beam that propagates in an optical fiber. Focus on the evolution of the polarization vector along the fiber's optic axis. We can imagine bending the fiber into any desired shape, and thereby controlling the shape of the ray. The ray's shape in turn will control the propagation of the polarization via Eq. (7.82).

If the fiber and ray are straight, then the propagation law (7.82) keeps $\hat{\mathbf{f}}$ constant. If the fiber and ray are circular, then the propagation law (7.82) causes $\hat{\mathbf{f}}$ to rotate in such a way as to always point along the generator of a cone, as shown in drawing (a) of Fig. 7.20. This polarization behavior, and that for any other ray shape, can be deduced with the aid of a unit sphere on which we plot the ray direction $\hat{\mathbf{k}}$ [drawing (b)]. For example, the ray directions at ray locations C and H [drawing (a)] are as shown in drawing (b). Notice, that the trajectory of $\hat{\mathbf{k}}$ around the unit sphere is a great circle.

On the unit sphere we also plot the polarization vector $\hat{\mathbf{f}}$ — one vector at each point corresponding to a ray direction. Because $\hat{\mathbf{f}} \cdot \hat{\mathbf{k}} = 0$, the polarization vectors are always tangent to the unit sphere. Notice that each $\hat{\mathbf{f}}$ on the unit sphere is identical in length and direction to the corresponding one in the physical space of drawing (a).

The parallel transport law (7.82) keeps constant the angle α between $\hat{\mathbf{f}}$ and the trajectory of $\hat{\mathbf{k}}$, i.e. the great circle in drawing (b). Translated back to drawing (a), this constancy of α implies that the polarization vector points always along the generators of the cone whose opening angle is $\pi/2 - \alpha$, as shown.

Next let the fiber and its central axis (the ray) be helical as shown in drawing (a) of Fig. 7.21. In this case, the propagation direction $\hat{\mathbf{k}}$ rotates, always maintaining the same angle θ to the vertical direction, and correspondingly its trajectory on the unit sphere of drawing (b) is a circle of constant polar angle θ. Therefore (as one can see, e.g., with the aid of a large globe of the Earth and a pencil that one transports around a circle of latitude $90^{\rm o} - \theta$), the parallel transport law dictates that the angle α between $\hat{\mathbf{f}}$ and the circle *not* remain constant, but instead rotate at the rate

$$d\alpha/d\phi = \cos\theta \ . \tag{7.83}$$

Here ϕ is the angle (longitude on the globe) around the circle. This is the same propagation law as for the direction of swing of a Foucault Pendulum as the earth turns, and for the same

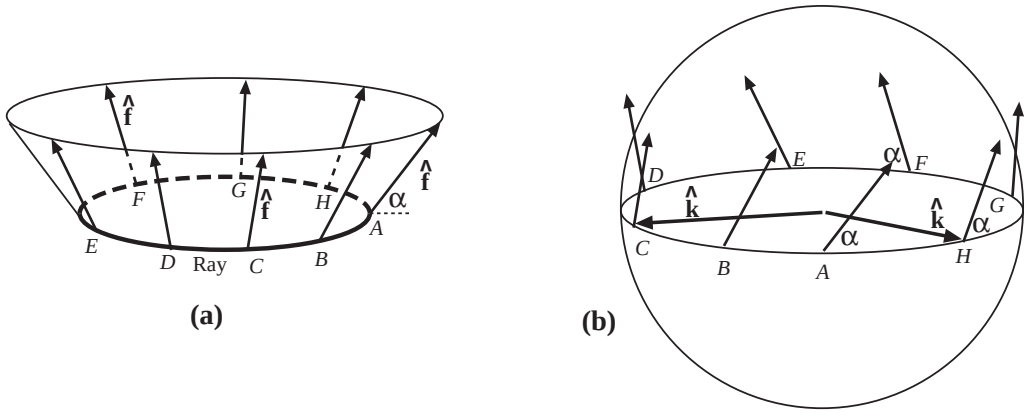


Fig. 7.20: (a) The ray along the optic axis of a circular loop of optical fiber, and the polarization vector $\hat{\mathbf{f}}$ that is transported along the ray by the geometric-optics transport law $d\hat{\mathbf{f}}/ds = -\hat{\mathbf{k}}(\hat{\mathbf{f}} \cdot d\hat{\mathbf{k}}/ds)$. (b) The polarization vector $\hat{\mathbf{f}}$ drawn on the unit sphere. The vector from the center of the sphere to each of the points A, B, ..., is the ray's propagation direction $\hat{\mathbf{k}}$, and the polarization vector (which is orthogonal to $\hat{\mathbf{k}}$ and thus tangent to the sphere) is identical to that in the physical space of the ray [drawing (a)].

reason: the gyroscopic action of the Foucault Pendulum is described by parallel transport of its plane along the earth's spherical surface.

In the case where θ is arbitrarily small (a nearly straight ray), Eq. (7.83) says $d\alpha/d\phi = 1$. This is easily understood: although $\hat{\mathbf{f}}$ remains arbitrarily close to constant, the trajectory of $\hat{\mathbf{k}}$ turns rapidly around a tiny circle about the pole of the unit sphere, so α changes rapidly—by a total amount $\Delta\alpha = 2\pi$ after one trip around the pole, $\Delta\phi = 2\pi$; whence $d\alpha/d\phi = \Delta\alpha/\Delta\phi = 1$. For any other helical pitch angle θ, Eq.(7.83) says that during one round trip α will change by an amount $2\pi\cos\theta$ that lags behind its change for a tiny circle (nearly straight ray) by the lag angle $\alpha_{\rm Lag} = 2\pi(1-\cos\theta)$, which is also the solid angle $\Delta\Omega$ enclosed by the path of $\hat{\mathbf{k}}$ on the unit sphere:

$$\alpha_{\rm Lag} = \Delta\Omega \ . \tag{7.84}$$

(For the circular ray of Fig. 7.20, the enclosed solid angle is $\Delta\Omega = 2\pi$ steradians, so the lag angle is 2π radians, which means that $\hat{\mathbf{f}}$ returns to its original value after one trip around the optical fiber, in accord with the drawings in the figure.)

Remarkably, Eq. (7.84) is true for light propagation along an optical fiber of any shape: if the light travels from one point on the fiber to another at which the tangent vector $\hat{\mathbf{k}}$ has returned to its original value, then the lag angle is given by the enclosed solid angle on the unit sphere, Eq. (7.84).

By itself, the relationship $\alpha_{\rm Lag} = \Delta\Omega$ is merely a cute phenomenon. However, it turns out to be just one example of a very general property of both classical and quantum mechanical

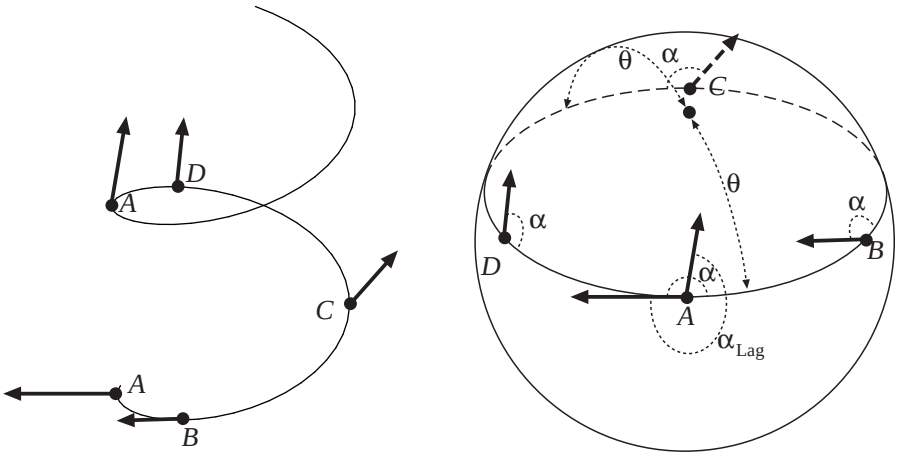


Fig. 7.21: (a) The ray along the optic axis of a helical loop of optical fiber, and the polarization vector $\hat{\mathbf{f}}$ that is transported along this ray by the geometric-optics transport law $d\hat{\mathbf{f}}/ds = -\hat{\mathbf{k}}(\hat{\mathbf{f}}\cdot d\hat{\mathbf{k}}/ds)$. The ray's propagation direction $\hat{\mathbf{k}}$ makes an angle $\theta = 73^{\rm o}$ to the vertical direction. (b) The trajectory of $\hat{\mathbf{k}}$ on the unit sphere (a circle with polar angle $\theta = 73^{\rm o}$), and the polarization vector $\hat{\mathbf{f}}$ that is parallel transported along that trajectory. The polarization vectors in drawing (a) are deduced from the parallel transport law of drawing (b). The lag angle $\alpha_{\rm lag} = 2\pi(1-\cos\theta) = 1.42\pi$ is equal to the solid angle contained inside the trajectory of $\hat{\mathbf{k}}$ (the $\theta = 73^{\rm o}$ circle).

systems, when they are forced to make slow, *adiabatic* changes described by circuits in the space of parameters that characterize them. In the more general case, one focuses on a phase lag rather than a direction-angle lag. We can easily translate our example into such a phase lag:

The apparent rotation of $\hat{\mathbf{f}}$ by the lag angle $\alpha_{\rm Lag} = \Delta\Omega$ can be regarded as an advance of the phase of one circularly polarized component of the wave by $\Delta\Omega$ and a phase retardation of the other circular polarization by the same amount. This implies that the phase of a circularly polarized wave will change, after one circuit around the fiber's helix, by an amount equal to the usual phase advance $\Delta\varphi = \int \mathbf{k}\cdot d\mathbf{x}$ (where $d\mathbf{x}$ is displacement along the fiber) *plus an extra* **geometric** *phase change* $\pm\Delta\Omega$, where the sign is given by the sense of circular polarization. This type of geometric phase change is found quite generally, when classical vector or tensor waves propagate through backgrounds that change slowly, either temporally or spatially; and the phases of the wave functions of quantum mechanical particles with spin behave similarly.

****

EXERCISES

Exercise 7.17 *Derivation: Parallel-Transport*

Use the parallel-transport law (7.82) to derive the relation (7.83).

Exercise 7.18 *Problem: Martian Rover*

A Martian Rover is equipped with a single gyroscope that is free to pivot about the direction perpendicular to the plane containing its wheels. In order to climb a steep hill on Mars without straining its motor, it must circle the summit in a decreasing spiral trajectory. Explain why there will be an error in its measurement of North after it has reached the summit. Could it be programmed to navigate correctly? Will a stochastic error build up as it traverses a rocky terrain?

****

Bibliographic Note

Modern textbooks on optics deal with the geometric optics approximation only for electromagnetic waves propagating through a dispersion-free medium. Accordingly, they typically begin with Fermat's principle, and then treat in considerable detail the paraxial approximation, and applications to optical instruments and sometimes the human eye. There is rarely any mention of the eikonal approximation or of multiple images and caustics. Examples of texts of this sort that we like are Ghatak (2010), Hecht (2002), and Bennett (2008). For a far

Box 7.3
Important Concepts in Chapter 7

- General Concepts
  - Dispersion relation – Sec. 7.2.1, Ex. 7.2, Eq. (7.37a)
  - Phase velocity and group velocity – Eqs. (7.2) and (7.9)
  - Wave Packet, its motion and spreading, and its internal waves – Sec. 7.2.2
  - Quanta associated with geometric-optics waves – Secs. 7.2.2 and 7.3.2
- General formulation of geometric optics: Sec. 7.3.3 and the following:
  - Eikonal (geometric optics) approximation – beginning of Sec. 7.3
  - Bookkeeping parameter for eikonal approximation – Box 7.2
  - Hamilton's equations for rays – Eqs. (7.25) and (7.38)
  - Connection to quantum theory – Sec. 7.3.2
  - Connection to Hamilton-Jacobi theory – Ex. 7.9, Eq. (7.26a)
  - Propagation law for amplitude (conservation of quanta) – Eqs. (7.34) and (7.40)
  - Fermat's principle, Eq. (7.43) in general; Eq. (7.45) for dispersionless waves
  - Breakdown of geometric optics – Sec. 7.3.5
- Dispersionless waves: EM waves in dielectric medium or a weak, Newtonian gravitational field; sound waves in fluid or isotropic solid
  - Lagrangian, wave equation, energy density and flux – Ex. 7.4, Eqs. (7.17)–(7.19)
  - Dispersion relation $\Omega = C(\mathbf{x}, t)k$ – Eq. (7.23)
  - Ray equation in second–order form – Eq. (7.47)
  - Fermat's principle for rays – Eq. (7.45)
  - Snell's law for rays in a stratified medium – Eq. (7.48)
  - Conservation of phase along a ray – Eq. (7.28)
  - Propagation law for amplitude – Eq. (7.34c)
  - Parallel propagation law for polarization vector – Eq. (7.82)
    - **T2** Geometric phase – Sec. 7.7.2
  - Paraxial optics – Sec. 7.4
    - Matrix formalism for rays – Secs. 7.4, 7.4.1
    - Application to charged particles in storage ring – Sec. 7.4.2
  - Multiple images, crossing of rays, coalescence of images and caustics – Sec. 7.5
    - Magnification at a caustic – Eq. (7.67)
    - Catastrophe theory – Sec. 7.5, Ex. 7.13
  - **T2** Gravitational lens – Sec. 7.6

more thorough treatise on geometric optics of scalar and electromagnetic waves in isotropic and anistropic dielectric media, see Kravtsov (2005).

We do not know of textbooks that treat the eikonal approximation in the generality of this chapter, though some should, since it has applications to all types of waves (many of which are explored later in this book). For the eikonal approximation specialized to Maxwell's equations, see Kravtsov (2005) and the classic treatise on optics by Born and Wolf (1999), which in this new edition has modern updates by a number of other authors. For the eikonal approximation specialized to the Schrödinger equation and its connection to Hamilton-Jacobi theory, see most any quantum mechanics textbook, e.g. Griffiths (2005).

Multiple image formation and caustics are omitted from most standard optics textbooks, except for a nice but out-of-date treatment in Born and Wolf (1999). Much better are the beautiful review by Berry and Upstill (1980) and the much more thorough treatments in Kravtsov (2005) and Nye (1999). For an elementary mathematical treatment of catastrophe theory, we like Saunders (1980). For a pedagogical treatise on gravitational lenses, see Schneider, Ehlers and Falco (1999), and for their applications to cosmology, see Blandford and Narayan (1992). Finally, for some history and details of the geometric phase, see Berry (1990).

Bibliography

Bennett, C. A. 2008. *Principles of Physical Optics*, New York: Wiley

Born, M. and Wolf, E. 1999. *Principles of Optics*, seventh expanded edition, Cambridge: Cambridge University Press

Berry, M. V. 1990. "Anticipations of the Geometric Phase", *Physics Today* **43**, 12, 34–40

Berry, M. V. & Upstill, C. 1980 "'Catastrophe optics: morphologies of caustics and their diffraction patterns", *Progress in Optics*, XVIII, 257–346.

Blandford, R. D. & Narayan, R. 1992. "Cosmological Applications of Gravitational Lenses", *Ann. Rev. Astr. Astrophys.*, **30**, 311–358.

Ghatak, Ajoy 2010. *Optics*, New York: McGraw-Hill

Goldstein, Herbert, Safko, John, and Poole, Charles. 2002 *Classical Mechanics* New York: Addison Wesley

Griffiths, David J. 2005. *Introduction to Quantum Mechanics*, second edition, Upper Saddle River NJ: Prentice-Hall

Hecht, E. 2002. *Optics*, fourth edition, New York: Addison Wesley

Kravtsov, Yury A. 2005. *Geometrical Optics in Engineering Physics*, Harrow UK: Alpha Science International.

Nye, J. F. 1999. *Natural Focusing and Fine Structure of Light*, Bristol U.K.: Institute of Physics Publishing.

Saunders, P.T. 1980 *Catastrophe Theory* Cambridge: Cambridge University Press

Schneider, P., Ehlers, J. and Falco, E.E. 1999 *Gravitational Lenses* Berlin: Springer-Verlag