Part II

OPTICS

Optics

Prior to the opening up of the electromagnetic spectrum and the development of quantum mechanics, the study of optics was only concerned with visible light. Reflection and refraction were first described by the Greek philosophers and further studied by the medieval scholastics and used 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 developing the telescope and the compound microscope. Naturally, the discovery of Snell's law and the observation of diffractive phenomena, stimulated serious speculation about the physical nature of light. The corpuscular and wave theories were propounded by Newton and Huygens, respectively. The corpuscular theory initially held sway, but the studies of interference by Young and the derivation of a wave equation for electromagnetic disturbances by Maxwell seemed to settle the matter in favor of the undulatory theory, only for the debate to be resurrected 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.

Physics students are all too familiar with this potted history and may consequently regard optics as an ancient precursor to modern physics that has been completely subsumed by quantum mechanics. However, this is not the case. Optics has developed dramatically and independently from quantum mechanics in recent decades, and is now a major branch of classical physics. 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 III of this book), fluids (Part IV), plasmas (Part V) and the geometry of spacetime (Part VI). 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 white light hologram to the laser scanner at a supermarket checkout, from laser printers to CD's and DVD'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 principles that we shall discuss.

The treatment of optics in this text differs from that found in traditional texts in that we shall assume familiarity with basic classical 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 Chapter 6, we shall discuss optical (wave-propagation) phenomena in the *geometric optics* approximation. This approximation is accurate whenever the wavelength and 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 in such a way that optical rays become the classical trajectories of particles, e.g. photons, and how, in general, ray systems develop singularities or caustics where the geometric optics approximation breaks down and we must revert to the wave description.

In Chapter 7 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," in which the wavefronts are approximately spherical or planar, 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 notions 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 Chapter 8. In this chapter we will also introduce the notion of coherence, which is a quantitative measure of the similarity of the combining waves and their capacity to interfere constructively.

The final chapter on optics, Chapter 9, 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 Part IV and plasma waves in Part V). For light (the focus of Chapter 9), they have become especially important; the nonlinear effects that arise when laser light is shone through certain crystals are beginning to have a strong impact on technology and on fundamental scientific research. We shall explore several examples.

Chapter 6

Geometric Optics

Version 0206.2, 06 Nov 02 *Please send comments, suggestions, and errata via email to kip@tapir.caltech.edu and rdb@caltech.edu, or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125*

6.1 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. 6.2), we shall begin our study of geometric optics in Sec. 6.3. There we shall derive the geometric-optic propagation equations with the aid of the *eikonal* approximation, and we shall elucicate the connection to Hamilton-Jacobi theory. 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. 6.4 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 using the problem of magnetically focusing a beam of charged particles and shall show how matrix methods can be used to describe the particle (i.e. ray) trajectories. In Sec. 6.5, 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 "geometrical" (or "adiabatic" or "Berry") phase. Finally, In Sec. 6.6, we shall discuss the formation of images in geometric optics, illustrating our treatment with gravitational lenses. We shall pay special attention to the behavior of images at caustics, and its relationship to catastrophe theory.

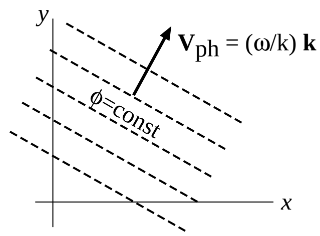

Fig. 6.1: A monochromatic plane wave in a homogeneous medium.

6.2 Waves in a Homogeneous Medium

6.2.1 Monochromatic, plane waves

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

$$\psi = Ae^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} , \tag{6.1}$$

where ψ is any oscillatory physical quantity associated with the wave. If, as is usually the case, the physical quantity is real (not complex), then we must take the real part of Eq. (6.1). In Eq. (6.1), A is the wave's complex amplitude, 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*, $\lambdabar = \lambda/2\pi$, its reduced wavelength and $\hat{\mathbf{k}} \equiv \mathbf{k}/k$ its *wave-vector direction*). The quantity in the exponential, $\phi = \mathbf{k}\cdot\mathbf{x} - \omega t$, is the wave's *phase*. Surfaces of constant phase are orthogonal to the propagation direction $\hat{\mathbf{k}}$ and move with the *phase velocity*

$$\mathbf{V}_{\rm ph} \equiv \left(\frac{\partial \mathbf{x}}{\partial t}\right)_\phi = -\frac{(\partial\phi/\partial t)_{\mathbf{x}}}{(\partial\phi/\partial\mathbf{x})_t} = \frac{\omega}{k}\hat{\mathbf{k}} ; \tag{6.2}$$

cf. Fig. 6.1. Lest there be confusion, Eq. (6.2) is short-hand notation for the Cartesian-component equation

$$V_{{\rm ph}\,j} \equiv \left(\frac{\partial x_j}{\partial t}\right)_\phi = -\frac{(\partial\phi/\partial t)_{\mathbf{x}}}{(\partial\phi/\partial x_j)_t} = \frac{\omega}{k}\hat{\mathbf{k}}_j . \tag{6.3}$$

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 $\omega = \Omega(\mathbf{k})$ is called the wave's *dispersion relation*.

Some examples of plane waves that we shall study in this book are: (i) Electromagnetic waves propagating through a dielectric medium with index of refraction n (this chapter), for which ψ could be any Cartesian component of the electric or magnetic field or vector potential and the dispersion relation is

$$\omega = \Omega(\mathbf{k}) = \frac{c}{n}k = \frac{c}{n}|\mathbf{k}| , \tag{6.4}$$

with c the speed of light in vacuum. (ii) Sound waves propagating through a solid (Part III) or fluid (liquid or vapor; Part IV), for which ψ could be the pressure or density perturbation produced by the sound wave, and the dispersion relation is the same as for electromagnetic waves (6.4), but with c now the sound speed under some fiducial condition at which (by convention) we set $n = 1$. (iii) Waves on the surface of a deep body of water (depth $\gg \lambdabar$; Part IV), for which ψ could be the height of the water above equilibrium and the dispersion relation is

$$\omega = \Omega(\mathbf{k}) = \sqrt{gk} = \sqrt{g|\mathbf{k}|} \ , \tag{6.5}$$

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

$$\omega = \Omega(\mathbf{k}) = \sqrt{\frac{EJ}{\Lambda}} k^2 = \sqrt{\frac{EJ}{\Lambda}} \mathbf{k} \cdot \mathbf{k} \ , \tag{6.6}$$

where EJ is the rod's "flexural rigidity" and Λ is its mass per unit length. (v) Alfvén waves (bending oscillations of plasma-laden magnetic field lines) in a magnetized, nonrelativistic plasma, for which ψ could be the transverse displacement of a magnetic field line and the dispersion relation is

$$\omega = \Omega(\mathbf{k}) = \mathbf{a} \cdot \mathbf{k}, \tag{6.7}$$

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

6.2.2 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:

$$\psi(\mathbf{x}, t) = \int \mathrm{A}(\mathbf{k}) e^{i\alpha(\mathbf{k})} e^{i(\mathbf{k}\cdot\mathbf{x} - \omega t)} d^3k \ , \tag{6.8}$$

where A is the modulus and α the phase of the complex amplitude $A = \mathrm{A}e^{i\alpha}$, and the integration element is $d^3k \equiv dk_x dk_y dk_z$ in terms of components of $\mathbf{k}$ on Cartesian axes x, y, z. Suppose, as is often the case, that $\mathrm{A}(\mathbf{k})$ is sharply concentrated around some specific wave vector $\mathbf{k}_o$ [see Ex. 6.2 for an example]. Then in the integral (6.8), 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 $\mathbf{k}$. This is the spacetime region in which the wave packet is concentrated, and its center is where $\boldsymbol{\nabla}_{\mathbf{k}}(\text{phasefactor}) = 0$:

$$\left(\frac{\partial \alpha}{\partial k_j} + \frac{\partial}{\partial k_j} (\mathbf{k} \cdot \mathbf{x} - \omega t) \right)_{\mathbf{k}=\mathbf{k}_o} = 0 \ . \tag{6.9}$$

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

$$\mathbf{x}_j - \left(\frac{\partial \Omega}{\partial k_j} \right)_{\mathbf{k}=\mathbf{k}_o} t = - \left(\frac{\partial \alpha}{\partial k_j} \right)_{\mathbf{k}=\mathbf{k}_o} = \text{const} \ . \tag{6.10}$$

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

$$\mathbf{V}_g = \nabla_{\mathbf{k}}\Omega \ , \quad \text{i.e.} \quad \mathbf{V}_{g\,j} = \left(\frac{\partial\Omega}{\partial k_j}\right)_{\mathbf{k}=\mathbf{k}_o} \ . \tag{6.11}$$

When, as for electromagnetic waves in a dielectric medium or sound waves in a solid or fluid, the dispersion relation has the simple form $\omega = \Omega(\mathbf{k}) \propto k = |\mathbf{k}|$, then the group and phase velocities are the same and the waves are said to be *dispersionless*. Denoting the proportionality constant by c/n as in Eq. (6.4), we have

$$\mathbf{V}_g = \mathbf{V}_{\rm ph} = \frac{c}{n}\hat{\mathbf{k}} \ . \tag{6.12}$$

If the dispersion relation has any other form, then the group and phase velocities are different, and the wave is said to exhibit *dispersion*; cf. Ex. 6.2. Examples are (see above): (iii) Waves on a deep body of water [dispersion relation (6.5)] for which

$$\mathbf{V}_g = \frac{1}{2}\mathbf{V}_{\rm ph} = \frac{1}{2}\sqrt{\frac{g}{k}}\hat{\mathbf{k}} \ . \tag{6.13}$$

(iv) Flexural waves on a rod or beam [dispersion relation (6.6)] for which

$$\mathbf{V}_g = 2\mathbf{V}_{\rm ph} = 2\sqrt{\frac{EJ}{\Lambda}}k^2\hat{\mathbf{k}} \ . \tag{6.14}$$

(v) Alfvén waves in a magnetized plasma [dispersion relation (6.7)] for which

$$\mathbf{V}_g = \mathbf{a} \ , \quad \mathbf{V}_{\rm ph} = (\mathbf{a}\cdot\hat{\mathbf{k}})\hat{\mathbf{k}} \ . \tag{6.15}$$

Notice that 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; see Fig. 6.2

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* $\mathbf{V}_{\rm ph}$. Similarly, 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 $\tilde{E} = \hbar\Omega(\mathbf{k}_o)$ and $\mathbf{p} = \hbar\mathbf{k}_o$.

****

EXERCISES

Exercise 6.1 *Practice: Group and Phase Velocities*

Derive the group and phase velocities (6.12)–(6.15) from the dispersion relations (6.4)–(6.7).

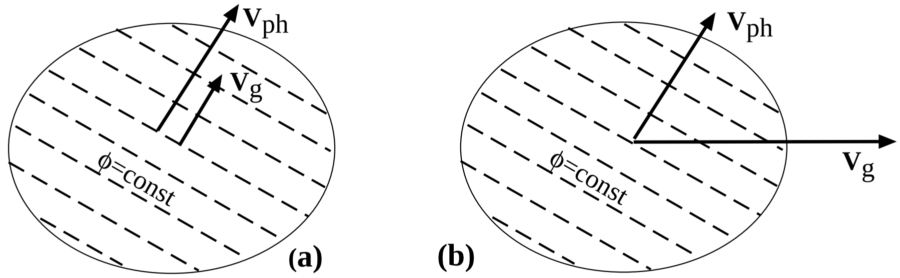

Fig. 6.2: (a) A Wave packet of waves on a deep body of water. The packet is localized in the spatial region bounded by the thin ellipse. Its center moves with the group velocity $\mathbf{V}_g$, and its surfaces of constant phase (the wave's oscillations) move twice as fast and in the same direction, $\mathbf{V}_{\rm ph} = 2\mathbf{V}_g$. 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. (6.8) and associated discussion]. (b) An Alfvén wave packet. Its center moves with a group velocity $\mathbf{V}_g$ that points along the homogeneous magnetic field [Eq. (6.15)], 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. (6.15)], which implies that the wave's oscillations remain fixed inside the packet as the packet moves.

Exercise 6.2 *Example: Gaussian Wave-Packet and its Spreading*

Consider a one-dimensional wave packet, $\psi(x,t) = \int \mathrm{A}(k) e^{i\alpha(k)} e^{i(kx-\omega t)} dk$ with dispersion relation $\omega = \Omega(k)$. For concreteness, let $\mathrm{A}(k)$ be a narrow Gaussian peaked around k_o: $\mathrm{A} \propto \exp[-\kappa^2/2(\Delta k)^2]$, where $\kappa = k - k_o$. Expand α as $\alpha(k) = \alpha_o - x_o\kappa$ assuming, for simplicity, that the quadratic term is 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 $\Omega = \omega_o + V_g\kappa + (dV_g/dk)\kappa^2/2$.

(a) Show that the wave packet is given by

$$\psi \propto \exp[i(\alpha_o + k_o x - \omega_o t)] \int_{-\infty}^{+\infty} \exp[i\kappa(x - x_o - V_g t)] \exp\left[-\frac{\kappa^2}{2}\left(\frac{1}{(\Delta k)^2} + i\frac{dV_g}{dk}t\right)\right] d\kappa \; . \tag{6.16}$$

The term in front of the integral describes the phase evolution of the waves inside the packet; cf. Fig. 6.2.

(b) Evaluate the integral analytically (with the help of Mathematica or Maple, if you wish). Show, from your answer, that the modulus of ψ is given by

$$|\psi| \propto \exp\left[-\frac{(x - x_o - V_g t)^2}{2L^2}\right] \; , \quad \text{where } L = \frac{1}{2\Delta k}\sqrt{1 + \left(\frac{dV_g}{dk}\Delta k\, t\right)^2} \tag{6.17}$$

is the packet's half width.

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

(d) Equation (6.17) shows that the wave packet spreads 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 initial widths can a water wave on the ocean spread by less than a factor 2 while traveling from Hawaii to California?

****

6.3 Waves in an Inhomogeneous, Time-Varying Medium: The Eikonal Approximation

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,

$$\mathcal{L} \gg \lambda\!\!\!^{-} = 1/k \; , \quad \mathcal{T} \gg 1/\omega \; , \tag{6.18}$$

then the waves can be locally 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 *eikonal approximation* or *geometric optics approximation*. When the waves are those of elementary quantum mechanics, it is called the *WKB approximation*. We shall derive the eikonal approximation's wave-propagation laws via the two-lengthscale expansion in Sec. 6.3.4; but first we shall motivate, describe and discuss those *geometric optics* laws (Sec. 6.3.1), give examples of them (Sec. 6.3.2), and describe their relationship to wave packets and enumerate phenomena, such as wave packet spreading, that they miss (Sec. 6.3.3).

6.3.1 Principal conclusions of the Eikonal approximation

Motivated by the mathematical form, Eq. (6.1) of plane waves in a homogeneous, time-independent medium, we write the waves in the *eikonal-approximation* form

$$\psi(\mathbf{x}, t) = A(\mathbf{x}, t) e^{i\phi(\mathbf{x}, t)} \; . \tag{6.19}$$

Here A is the waves' (slowly varying) complex amplitude and ϕ is their (rapidly varying) phase, and we *define* the wave vector (field) and angular frequency (field) by

$$\mathbf{k}(\mathbf{x}, t) \equiv \boldsymbol{\nabla}\phi \; , \quad \omega(\mathbf{x}, t) \equiv -\partial\phi/\partial t \; . \tag{6.20}$$

The eikonal approximation includes, in addition to $\mathcal{L} \gg 1/k$ and $\mathcal{T} \gg 1/\omega$, also the requirement that A, $\mathbf{k}$ and ω vary slowly, i.e., vary on lengthscales and timescales long compared to $\lambda\!\!\!^{-} = 1/k$ and $1/\omega$. This requirement guarantees that the waves are locally planar

($\phi \simeq \mathbf{k} \cdot x - \omega t + \text{constant}$), and because they are locally planar, their frequency and wave vector (6.20) must be related by essentially the same dispersion relation as for a homogeneous medium:

$$\omega = \Omega(\mathbf{k}; \mathbf{x}, t) \; . \tag{6.21}$$

The dependence of Ω on x and t is induced by the slow variations of the medium. For example, in the case of Alfvén waves, the plasma density ρ and magnetic field $\mathbf{B}$ will vary slowly in space and time, producing corresponding slow variations of the Alfven velocity $\mathbf{a}(\mathbf{x}, t) = \mathbf{B}(\mathbf{x}, t)/\sqrt{4\pi\rho(\mathbf{x}, t)}$ and thence slow variations of the dispersion relation $\Omega(\mathbf{k}; \mathbf{x}, t) = \mathbf{a}(\mathbf{x}, t) \cdot \mathbf{k}$ [Eq. (6.7)].

Although the waves described by Eq. (6.19) are classical and our analysis will be classical, their propagation laws in the eikonal approximation can be described most nicely in quantum mechanical language.[1] Quantum mechanics insists that associated with any wave, in the geometric optics regime, there are quanta: the wave's associated 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. When we multiply the wave's $\mathbf{k}$, and ω by Planck's constant, we obtain the particles' momentum $\mathbf{p} = \hbar\mathbf{k}$ and energy $\tilde{E}$. Although the 19th century theory of classical waves was unaware of these quanta, once quantum mechanics had been formulated the quanta became a powerful conceptual tool for thinking about classical waves.

As for wave packets in a homogeneous medium, so also for the wave (6.19) in the inhomogeneous, time-varying case, *the wave's quanta, their energy, their amplitude, and any wave packets in which they may be concentrated, all propagate with the group velocity* $\mathbf{V}_g$. We embody this group velocity in the following equation of motion for the quanta:

$$\frac{dx_j}{dt} = V_{g\,j} = \left(\frac{\partial\Omega}{\partial k_j}\right)_{\mathbf{x},t} = \left(\frac{\partial(\hbar\Omega)}{\partial p_j}\right)_{\mathbf{x},t} \; . \tag{6.22}$$

Here $p_j = \hbar k_j$ is the momentum of the quantum, and this wave-theory expression for the group velocity [cf. Eq. (6.11)] can be regarded as one of Hamilton's equations for a quantum's motion, with $H(\mathbf{p}; \mathbf{x}, t) \equiv \hbar\Omega(\mathbf{k}; \mathbf{x}, t)$ playing the role of the quantum's Hamiltonian. The world lines with velocity $\mathbf{V}_g$ along which a quantum moves are regarded classically as *rays* along which the waves propagate (see Fig. 6.3 for two examples). The second of Hamilton's equations

$$\frac{dp_j}{dt} = -\left(\frac{\partial(\hbar\Omega)}{\partial x_j}\right)_{\mathbf{p},t} \; , \quad \text{i.e.} \quad \frac{dk_j}{dt} = -\left(\frac{\partial\Omega}{\partial x_j}\right)_{\mathbf{k},t} \tag{6.23}$$

describes how the quantum's momentum $\mathbf{p}$ and the wave's wave vector $\mathbf{k} = \mathbf{p}/\hbar$ vary along the ray. Here, as in Eq. (6.22) d/dt is the time derivative along the ray and can be written as

$$\frac{d}{dt} = \frac{\partial}{\partial t} + V_{g\,j}\frac{\partial}{\partial x_j} \; . \tag{6.24}$$

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

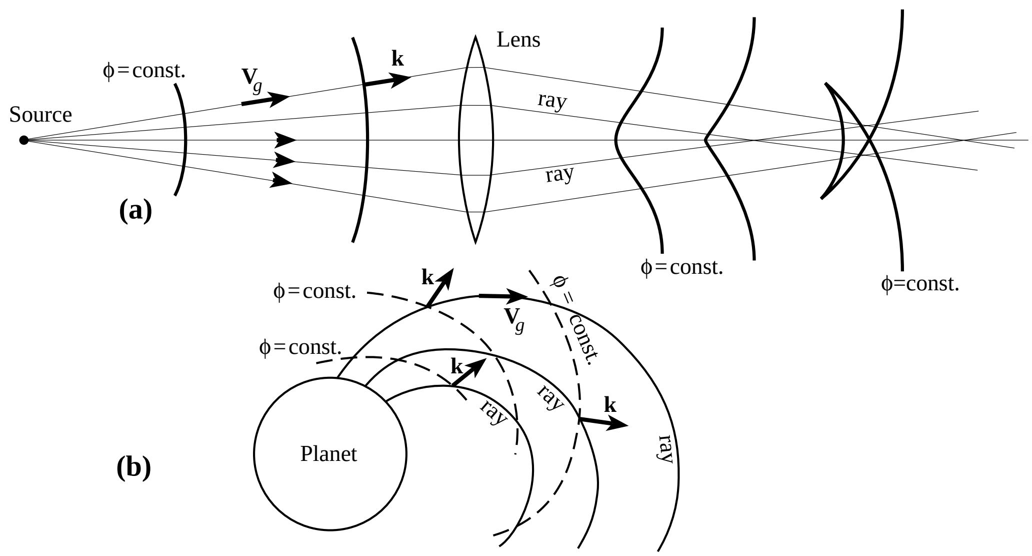

Fig. 6.3: (a) The rays and the surfaces of constant phase ϕ at a fixed time for light passing through a converging lens [dispersion relation $\Omega = ck/n(\mathbf{x})$]. In this case the rays (which always point along $\mathbf{V}_g$) are parallel to the wave vector $\mathbf{k} = \boldsymbol{\nabla}\phi$ and thus also parallel to the phase velocity $\mathbf{V}_{\rm ph}$, and the waves propagate along the rays with a speed $V_g = V_{\rm ph} = c/n$ that is independent of wavelength. (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 $\mathbf{V}_g = \mathbf{a} \equiv \mathbf{B}/\sqrt{4\pi\rho}$, 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 $V_g = B/\sqrt{4\pi\rho}$ that are independent of wavelength; cf. Fig. 6.2 (b). 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 and energy of each quantum is conserved, and conservation of quanta implies conservation of wave energy. Because the Alfvén speed generally diminishes with distance from the planet, conservation of wave energy typically requires the waves' energy density and amplitude to increase as they climb upward.

The third of Hamilton's equations

$$\frac{d\tilde{E}}{dt} = \left(\frac{\partial(\hbar\Omega)}{\partial t}\right)_{\mathbf{x},\mathbf{p}} , \quad \text{i.e.} \quad \frac{d\omega}{dt} = \left(\frac{\partial\Omega}{\partial t}\right)_{\mathbf{x},\mathbf{k}} \tag{6.25}$$

describes how the quantum's energy $\tilde{E} = \hbar\omega$ and the wave's angular frequency ω vary along the ray.

Turn next to the manner in which the field $\psi = Ae^{i\phi}$ is transported along the ray. The transport of the phase ϕ is determined by the transport of $\omega = -\partial\phi/\partial t$ [Eq. (6.25)] or equally well by the transport of $\mathbf{k} = \boldsymbol{\nabla}\phi$ [Eq. (6.23)]; having computed $\omega(\mathbf{x}, t)$ or $\mathbf{k}(\mathbf{x}, t)$ by integrating the Hamilton equations along the ray, one can then compute ϕ via a simple time or spatial integral. The transport equation for the amplitude A can be deduced from the law of conservation of quanta. [That the number of quanta in the wave is conserved, even if the wave's medium is varying in time (albeit slowly, $\mathcal{T} \gg 1/\omega$), can be proved in three different ways: (i) it is a direct consequence of the classical wave's Eikonal approximation; (ii) it is a consequence of the theory of *adiabatic invariants*, discussed in standard texts on analytical mechanics, e.g. Goldstein (1980); (iii) it is a consequence of the quantum mechanical laws for creation and annihilation of quanta.]

Denoting by $\varepsilon = T^{00}$ the waves' energy density and by $\varepsilon V_{g\,j} = T^{0j}$ their energy flux (an expression which embodies the fact that the energy is carried by the quanta, with velocity $\mathbf{V}_g$), we can write the number density and flux of quanta as

$$S^0 = \frac{\varepsilon}{\hbar\omega} \;, \quad S_j = \frac{\varepsilon V_{g\,j}}{\hbar\omega} \;, \tag{6.26}$$

where we have used the fact that each quantum carries an energy $\hbar\omega$. Then the law of conservation of quanta and the corresponding (equivalent) law for the evolution of energy are

$$\frac{\partial S^0}{\partial t} + \frac{\partial S_j}{\partial x_j} = 0 \;, \quad \frac{\partial(\varepsilon/\omega)}{\partial t} + \frac{\partial(\varepsilon V_{g\,j}/\omega)}{\partial x_j} = 0 \;. \tag{6.27}$$

The latter equation can be rewritten as a law for transporting ε/ω along a ray by using $d/dt = \partial/\partial t + \mathbf{V}_g \cdot \boldsymbol{\nabla}$ for the time derivative moving with the quanta:

$$\frac{d(\varepsilon/\omega)}{dt} + (\varepsilon/\omega)\boldsymbol{\nabla} \cdot \mathbf{V} = 0 \;. \tag{6.28}$$

If $\partial\Omega/\partial t = 0$, then ω is conserved along a ray [Eq. (6.25)] and particle conservation implies energy conservation. If $\partial\Omega/\partial t \neq 0$, then energy is exchanged between the medium and the wave but ε/ω is still conserved. If one knows how to express the energy density ε in terms of the waves' amplitude A (an expression that will depend on the physical nature of the waves), then by inserting that expression into Eq. (6.27) one obtains a law for the propagation of the waves' amplitude along the ray.

6.3.2 Examples of Propagation Laws

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. (6.4) with $c = n = 1$: $\Omega = k$. It is straightforward [Exercise 6.3] 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. The Hamilton equation $d\omega/dt = 0$ [Eq. (6.25)] says that ω is conserved along a ray, so it must be a function of retarded time, $t - r$. Integrating $\partial\phi/\partial t = -\omega(t-r)$, or equally well $\boldsymbol{\nabla}\phi = \mathbf{k} = \omega(t-r)\mathbf{e}_r$, we infer that the phase ϕ must also be a function of retarded time. The law of conservation of quanta in this case reduces to the propagation law $d(rA)/dt = 0$ [Exercise 6.3] so rA is also a constant along the ray and is therefore a function of retarded time. Putting this all together, we conclude that

$$\psi = \frac{\mathcal{A}(t-r)}{r} e^{i\phi(t-r)} \;. \tag{6.29}$$

This is not only the general solution to the scalar wave equation $-\partial^2\psi/\partial t^2 + \nabla^2\psi = 0$ in the geometric optics approximation; it is an exact solution without any approximation at all.

As another example of the geometric-optics propagation laws, consider flexural waves on a spacecraft's tapering antenna. The dispersion relation is $\Omega = k^2\sqrt{EJ/\Lambda}$ [Eq. (6.6)] with $EJ/\Lambda \propto d^2$, where d is the antenna diameter (cf. Chaps. 10 and 11). Since Ω is independent of t, as the waves propagate from the spacecraft to the antenna's tip, their frequency ω is conserved [Eq. (6.25)], which implies by the dispersion relation that $k = (EJ/\Lambda)^{-1/4}\omega^{1/2} \propto d^{-1/2}$, whence the wavelength decreases as $d^{1/2}$. The group velocity is $V_g = 2(EJ/\Lambda)^{1/4}\omega^{1/2} \propto d^{1/2}$. 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 $\propto d^{-1/4}$ as the flexural waves approach the antenna's end. A qualitatively similar phenomenon is seen in the "cracking" of a bullwhip.

Fig. 6.3 sketches two other examples: light propagating through a lens, and Alfvén waves propagating in the magnetosphere of a planet. Below we shall explore a variety of other applications, but first we shall sketch derivations of the propagation laws (Sec. 6.3.4) and their relationship to wave packets and how they can fail (Sec. 6.3.3).

6.3.3 Relation to Wave Packets; Breakdown of the Eikonal Approximation

The form $\psi = Ae^{i\phi}$ 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, $\omega \equiv -\partial\phi/\partial t$, $\mathbf{k} \equiv \boldsymbol{\nabla}\phi$, and dispersion relation $\Omega(\mathbf{k}; \mathbf{x}, t)$ all vary on lengthscales $\mathcal{L}$ long compared to $\lambdabar = 1/k$ and timescales $\mathcal{T}$ long compared to $1/\omega$]. For example, ψ could be as nearly a plane wave 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) as they propagate; see Ex. 6.2. This spreading is not correctly reproduced by the geometric optics propagation laws, because 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. 6.2). This is the geometric-optics regime, and geometric optics ignores the spreading.

Many other wave phenomena are missed by geometric optics. Examples are diffraction (Chap. 7), nonlinear wave-wave coupling (chap. 9 and Parts III and IV), and parametric amplification of waves by rapid time variations of the medium—which shows up in quantum mechanics as particle production (i.e., a breakdown of the law of conservation of quanta). In Part VI, we shall study such particle production in inflationary models of the early universe.

6.3.4 Derivation of Propagation Laws

One can best get the flavor of the origin of the geometric-optics propagation laws by focusing, initially, on a specific, simple type of wave. We choose one whose wave equation is

$$-\frac{\partial}{\partial t}\left(\frac{n^2}{c^2}\frac{\partial \psi}{\partial t}\right) + \nabla^2\psi = 0 \ , \tag{6.30}$$

where the index of refraction varies slowly in space and time, $n = n(\mathbf{x}, t)$, and c is a constant, fiducial velocity. As we shall see, this wave equation gives rise to the dispersionless dispersion relation $\Omega(\mathbf{k}) = (c/n)k$ [Eq. (6.4)]. As was discussed in the last section, this dispersion relation holds for electromagnetic waves in a dielectric medium or sound waves in a solid or fluid, but in these cases the wave equation (6.30) must be augmented by terms proportional to $\boldsymbol{\nabla} \ln n$. As we shall see in Sec. 6.4 below and in Part VI, Eq: (6.30) holds true without $\boldsymbol{\nabla} \ln n$ modifications for electromagnetic or gravitational waves propagating through a slowly changing Newtonian gravitational field $\Phi(\mathbf{x}, t)$, in which case c is the speed of light in vacuum and $n = 1 - 2\Phi(\mathbf{x}, t)/c^2$.

We begin our derivation by rewriting the geometric-optics expression $\psi = Ae^{i\phi}$ for the wave [Eq. (6.19)] in a slightly different form:

$$\psi = (A + \epsilon B + \ldots)e^{i\phi/\epsilon} \ . \tag{6.31}$$

Here ϵ is a bookkeeping device that tells us how the terms it multiplies scale with $1/k\mathcal{L}$ and $1/\omega\mathcal{T}$. In particular, because the phase ϕ has the form $\mathbf{k}\cdot\mathbf{x} - \omega t$ when the medium is homogeneous, it must scale linearly with k and ω; hence the factor $1/\epsilon$ in the exponential. Similarly, since the amplitude A is constant when the medium is homogeneous, it must be left unchanged when k and ω are doubled; hence the absence of any factor ϵ multiplying A. The quantity B is the first small correction to A caused by finiteness of the wavelength, so it is halved when k and ω are doubled; hence its factor ϵ. We set the numerical value of ϵ to unity, and thereby we can delete it from all equations—after we have used it to make sure that our bookkeeping on the orders of terms has been carried out properly. The use of this bookkeeping device is common in two-lengthscale computations.

Equation (6.31) is called the *eikonal approximation*, and the phase ϕ is called the *eikonal* after the Greek word for image.

By inserting the eikonal approximation (6.31) into the wave equation (6.30) and then collecting terms in powers of ϵ, we obtain the following: The $O(\epsilon^{-2})$ terms [i.e. the terms that are of order $(k\mathcal{L})^2$ and $(\omega\mathcal{T})^2$ and thus are huge] are

$$\epsilon^{-2}\left(\frac{n^2\omega^2}{c^2} - k^2\right)Ae^{i\phi/\epsilon} = 0 \ ; \tag{6.32}$$

and the much smaller, $O(\epsilon^{-1})$ terms are

$$\epsilon^{-1}\left[\left(\frac{n^2\omega^2}{c^2} - k^2\right)B + i\left(\frac{\partial(n^2\omega)}{c^2\partial t}A + \frac{2n^2\omega}{c^2}\frac{\partial A}{\partial t} + A\boldsymbol{\nabla}\cdot\mathbf{k} + 2(\mathbf{k}\cdot\boldsymbol{\nabla})A\right)\right]e^{i\phi/\epsilon} = 0 \ . \tag{6.33}$$

The next-order terms, i.e. those of $O(\epsilon^0)$, turn out to govern the evolution of the "post-geometric-optics correction factor" B, but they are rarely of much use because, when B is important, the eikonal approximation tends to break down severely.

The leading-order Eq. (6.32) can be satisfied only if the frequency and wave vector are related by the standard dispersionless dispersion relation

$$\omega = \Omega(\mathbf{k}, \mathbf{x}, t) \equiv \frac{c}{n(\mathbf{x}, t)} k \tag{6.34}$$

[Eq. (6.4)]. In the next-order Eq. (6.33), the term proportional to B vanishes by virtue of the dispersion relation, and the rest can be regarded as a differential equation for the waves' amplitude:

$$\frac{dA}{dt} \equiv \frac{\partial A}{\partial t} + \mathbf{V}_g \cdot \boldsymbol{\nabla} A = -\frac{1}{2}\left(\frac{c}{nk}\boldsymbol{\nabla}\cdot\mathbf{k} + \frac{\partial \ln(\omega n^2)}{\partial t}\right) A \, . \tag{6.35}$$

Eq. (6.35) implies that, in the geometric optics limit, the value of A at any location $\mathcal{P}$ in spacetime is influenced only by A at earlier locations along the ray that passes through $\mathcal{P}$, and not by the value of A at any other locations. In this sense, the amplitude propagates along the wave's rays, as was discussed above. When n is time independent and hence so is ω [Eq. (6.34)], the last term vanishes and the $\boldsymbol{\nabla}\cdot\mathbf{k}$ term enforces a "1/r" falloff of the amplitude A, thereby guaranteeding energy conservation; cf. Eq. (6.40) below and associated discussion.

With this simple case as a model, we now go back and study a scalar wave equation with the much more general form

$$\frac{\partial^2\psi}{\partial t^2} + \Omega^2(-i\boldsymbol{\nabla}, \mathbf{x}, t)\psi = F(-i\boldsymbol{\nabla}, i\partial/\partial t, \mathbf{x}, t)\psi \, . \tag{6.36}$$

Here Ω^2 is a polynomial of order N in $-i\boldsymbol{\nabla}$ with coefficients that vary slowly in space $\mathbf{x}$ and time t, and F is a polynomial of order $N-1$ or smaller in $-i\boldsymbol{\nabla}$ and is first order in $i\partial/\partial t$. (The generalization to other wave equations, e.g. vectorial or tensorial ones, is straightforward.) When one inserts the eikonal approximation (6.31) into the general wave Eq. (6.36), one obtains at leading order in ϵ the dispersion relation $\omega = \Omega(\mathbf{k}, \mathbf{x}, t)$, where Ω is the function appearing in Eq. (6.36). The next order terms have the form

$$-2i\omega\frac{\partial A}{\partial t} - 2i\Omega(\mathbf{k}, \mathbf{x}, t)\frac{\partial\Omega(\mathbf{k}, \mathbf{x}, t)}{\partial k_j}\frac{\partial A}{\partial x_j} = \text{terms proportional to } A \, . \tag{6.37}$$

Using the dispersion relation $\omega = \Omega(\mathbf{x}, t, \mathbf{k})$, we bring this into the "propagate A along a ray" form

$$\frac{dA}{dt} \equiv \frac{\partial A}{\partial t} + \mathbf{V}_g \cdot \boldsymbol{\nabla} A = \text{terms proportional to } A \, , \tag{6.38}$$

where the group velocity as always is

$$V_{g\,j} = \partial\Omega/\partial k_j \, . \tag{6.39}$$

The laws of quantum mechanics guarantee that the terms on the right side of A's propagation law (6.38) have just the right form to enforce conservation of quanta, Eq. (6.27). We shall verify this below for the special case of sound waves in a time-varying fluid, but first we must derive Hamilton's equations for the rays.

We begin our derivation of Hamilton's equations by inserting the definitions $\omega = -\partial\phi/\partial t$ and $\mathbf{k} = \boldsymbol{\nabla}\phi$ into the general dispersion relation $\omega = \Omega(\mathbf{k}, \mathbf{x}, t)$ for an arbitrary wave, thereby obtaining

$$\frac{\partial \phi}{\partial t} + \Omega(\boldsymbol{\nabla}\phi, \mathbf{x}, t) = 0\ . \tag{6.40}$$

This equation is known in optics as the *eikonal equation*. It is formally the same as the Hamilton-Jacobi equation of classical mechanics[2] if we identify $\hbar\Omega$ with the Hamiltonian and $\hbar\phi$ with Hamilton's principal function. This suggests that we follow the same procedure as is used to derive Hamilton's equations of motion. We take the gradient of Eq. (6.40) to obtain

$$\frac{\partial^2 \phi}{\partial t \partial x_j} + \frac{\partial \Omega}{\partial k_l}\frac{\partial^2 \phi}{\partial x_l \partial x_j} + \frac{\partial \Omega}{\partial x_j} = 0\ , \tag{6.41}$$

where the partial derivatives of Ω are with respect to its arguments $(\mathbf{k}, \mathbf{x}, t)$; we then use $\partial\phi/\partial x_j = k_j$ and $\partial\Omega/\partial k_l = V_{g\,l}$ to write this as $dk_j/dt = -\partial\Omega/\partial x_j$. This is the second of Hamilton's equations (6.23), and it tells us how the wave vector changes along a ray. The equation for the ray's tangent vector, $dx_j/dt = \partial\Omega/\partial k_j$, is the first of Hamilton's equations, (6.22). The third Hamilton equation, $d\omega/dt = \partial\Omega/\partial t$ [Eq. (6.25)] is obtained by taking the time derivative of the eikonal equation (6.40).

Turn, finally, to the evolution law for energy, Eq. (6.27) and its relationship to conservation of particles and to the propagation law for amplitude. We shall not attempt a general derivation but instead shall focus on a simple example, sound waves in a fluid. Suppose that the fluid is gradually warmed, keeping its density ρ fixed. The warming raises the fluid's pressure and lowers its index of refraction for sound waves, so n is independent of $\mathbf{x}$ but slowly decreasing in t. Sound waves in the fluid will then have oscillatory velocities given by $\mathbf{v} = \boldsymbol{\nabla}\psi$, where $\psi(\mathbf{x}, t)$ obeys the simple wave equation (6.30); see Chap. 15. As we show in Ex. 6.7 and in Chap. 15, the waves' energy density and flux are given by

$$\varepsilon = \frac{\rho}{2}\left\langle \left(\frac{n^2}{c^2}\right)\left(\frac{\partial \psi}{\partial t}\right)^2 + (\boldsymbol{\nabla}\psi)^2 \right\rangle\ , \tag{6.42}$$

$$\mathbf{F} = \varepsilon \mathbf{V}_g = -\rho \left\langle \left(\frac{\partial \psi}{\partial t}\right)\boldsymbol{\nabla}\psi \right\rangle\ , \tag{6.43}$$

where the average $\langle \ldots \rangle$ is over a wave period. Now let us examine the evolution of the waves' energy. Differentiating Eq. (6.42) with respect to time, taking the divergence of Eq. (6.43), adding them together, and using the wave equation (6.30), we obtain

$$\frac{\partial \varepsilon}{\partial t} + \boldsymbol{\nabla}\cdot F = -\left(\frac{\rho n^2}{c^2}\right)\frac{\partial \ln n}{\partial t}\left\langle \left(\frac{\partial \psi}{\partial t}\right)^2 \right\rangle\ , \tag{6.44}$$

which can be simplified further by using the equipartition of energy between the two contributions to the energy density in Eq. (6.42):

$$\frac{\partial \varepsilon}{\partial t} + \boldsymbol{\nabla}\cdot F = -\varepsilon\frac{\partial \ln n}{\partial t}\ . \tag{6.45}$$

[2]See, for example, Goldstein (1980).

This equation shows that the wave energy is not conserved; if it were, the right hand side would vanish. The medium does work on the waves as it gradually decreases their index of refraction. By inserting the eikonal approximation $\psi = \Re(Ae^{i\phi})$ (where $\Re$ means take the real part) into expressions (6.42) and (6.43) for ϵ and $\mathbf{F}$, and putting these in turn into the energy evolution law (6.45), we can obtain an evolution law for the wave amplitude A. Not surprisingly, this evolution law is identical to the law (6.35) of propagation along the rays, which we obtained directly from the wave equation in the eikonal approximation.

Although the fluid's wave energy is not conserved, the number of quanta (phonons) in the wave *is* conserved: The density and flux of quanta are $\varepsilon/\hbar\omega$ and $\mathbf{F}/\hbar\omega = \varepsilon\mathbf{V}_g/\hbar\omega$, and their evolution law [just a rewrite of energy evolution (6.45)] is

$$\frac{\partial}{\partial t}\left(\frac{\varepsilon}{\hbar\omega}\right) + \nabla\cdot\left(\frac{\varepsilon\mathbf{V}_g}{\hbar\omega}\right) = -\frac{\varepsilon}{\hbar\omega}\frac{d\ln(\omega n)}{dt} = 0\ . \tag{6.46}$$

That the second expression vanishes, i.e. that ωn is constant along a ray, follows from Hamilton's equations (6.22)–(6.25) with $\Omega = [c/n(t)]k$; it also should be clear from a simple physical argument: The product ωn is equal to ck by the dispersion relation, and therefore it is proportional to the wave number k. Now, imagine a standing wave, inside our spatially homogeneous medium, made from two waves of the same frequency and wave number that travel in opposite directions. It should be clear that the wavelength of this standing wave cannot change with time, and therefore its reciprocal, the wave number k, cannot change, and therefore $n\omega = ck$ cannot change.

6.3.5 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 Fermat. This *Fermat's principle* is the optical analogue of Maupertuis' principle of least action in classical mechanics.[3] 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

$$J = \int \mathbf{p}\cdot d\mathbf{q}\ , \tag{6.47}$$

(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 in any text on analytical dynamics, 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

[3] Goldstein (1980).

these two points. The rays (usually only one) are those paths that satisfy the variational principle

$$\delta \int \mathbf{k} \cdot d\mathbf{x} = 0 \; . \tag{6.48}$$

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 $\mathbf{k} = \boldsymbol{\nabla}\phi$, 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 $\Delta\phi$ 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 and we can think of the action principle as one of extremal phase difference $\Delta\phi$. 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 $e^{i\Delta\phi}$; and we must then add together these amplitudes. The contributions from almost all neighboring paths will interfere destructively. The only exceptions are those paths whose neighbors have the same values of $\Delta\phi$, to first order in the path difference. These are the paths that extremize the action (6.48); i.e., they are the wave's rays.

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 $\Omega = kc/n(\mathbf{x})$ — 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. In this case, the Hamiltonian dictates that for each trial path, $\mathbf{k}$ is parallel to $d\mathbf{x}$, and therefore $\mathbf{k} \cdot d\mathbf{x} = kds$, where s is distance along the path. Using the dispersion relation $k = n\Omega/c$ and noting that Hamilton's equation $dx^j/dt = \partial\Omega/\partial k_j$ implies $ds/dt = c/n$ for the rate of traversal of the trial path, we see that $\mathbf{k} \cdot d\mathbf{x} = kds = \Omega dt$. Since the trial paths are constrained to have Ω constant, Fermat's principle (6.48) becomes a *principle of extremal time*: The rays between $\mathbf{x}_{\rm initial}$ and $\mathbf{x}_{\rm final}$ are those paths along which

$$\int dt = \int \frac{n(\mathbf{x})}{c} ds \tag{6.49}$$

is extremal—or, equivalently, since c is a constant, they are the paths of extremal *optical path length* $\int n(\mathbf{x}) ds$.

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 $n_{\rm min} L$, where $n_{\rm min}$ 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 equa-

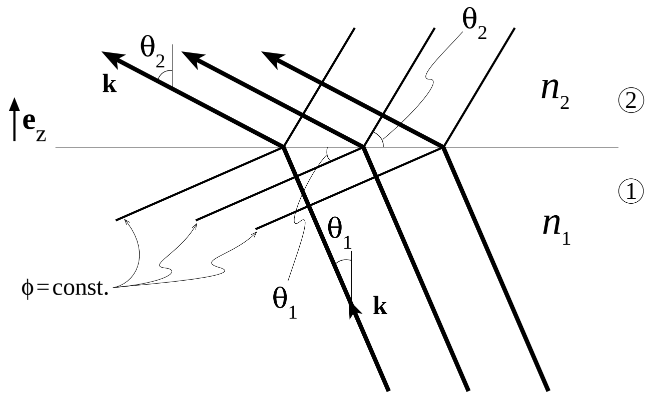

Fig. 6.4: Illustration of Snell's law of refraction at the interface between two media where the refractive indices are n_1, n_2. As the wavefronts must be continuous across the interface, and the wavelengths are inversely proportional to the refractive index, we have from simple trigonometry that $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

tion for the ray. For ease of derivation, we write the action principle in the form

$$\delta \int n(\mathbf{x}) \left(\frac{d\mathbf{x}}{d\mathbf{s}} \cdot \frac{d\mathbf{x}}{d\mathbf{s}} \right)^{1/2} ds, \tag{6.50}$$

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

$$\frac{d}{ds}\left(n \frac{d\mathbf{x}}{ds} \right) = \boldsymbol{\nabla} n \; . \tag{6.51}$$

This is equivalent to Hamilton's equations for the ray, as one can readily verify using the Hamiltonian $\Omega = kn/c$ [Ex. 6.3].

Equation (6.51) 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 $n = n(z)$, where (x, y, z) are Cartesian coordinates. Projecting Eq. (6.51) perpendicular to $\mathbf{e}_z$, we discover that ndy/ds and ndx/ds are constant, which implies

$$n \sin\theta = \text{constant} \tag{6.52}$$

where θ is the angle between 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 ϕ is constant (cf. Fig. 6.4).

****

EXERCISES

Exercise 6.3 *Derivation and Practice: Spherical Solution to Vacuum Scalar Wave Equation*

Derive the spherical solution (6.29) 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. Use the propagation law (6.35) for the amplitude, which [as is briefly discussed after Eq. (6.45)] follows from the law of conservation of quanta.

Exercise 6.4 *Problem: Gravitational Waves From a Spinning, Deformed Neutron Star*

Gravitational waves, propagating through the external, nearly Newtonian gravitational field $\Phi = -GM/r$ of their source, obey the wave equation (6.30) with $n = 1 - 2\Phi/c^2 = 1 + 2GM/c^2 r$. Here M is the source's mass, and the wave field ψ is a dimensionless "strain (distortion) of space" $h_+(\mathbf{x}, t)$ which we shall study in Chapter 26. For a spinning, deformed neutron star residing at the origin of spherical polar coordinates (r, θ, φ) with its spin along the polar axis, a particular geometric-optics solution of this wave equation has amplitude and phase with the following forms:

$$A = \frac{\mathcal{A}(1+\cos^2\theta)e^{2i\varphi}}{r[1+(r_*-t)^2/\tau^2]} , \quad \phi = \omega_o \tau e^{(r_*-t)/\tau} , \quad \text{where } r_* \equiv r + 2M \ln\left(\frac{r}{2M} - 1\right) . \tag{6.53}$$

Here and throughout this problem, for simplicity of notation, we adopt units in which $G = c = 1$ (cf. Chapter 24). The quantity $\mathcal{A}$ is a constant characterizing the strength of the waves, ω_o is some constant frequency, and $\tau \gg 1/\omega_o$ is a long timescale on which the waves' frequency and amplitude change. We restrict attention to times $t > r_*$, so the amplitude is slowly dying and (as you shall see) the frequency is slowly decreasing—due to gradual spindown of the star.

(a) What are $\omega(\mathbf{x}, t)$ and $\mathbf{k}(\mathbf{x}, t)$ for these gravitational waves?

(b) Verify that these ω and $\mathbf{k}$ satisfy the dispersion relation (6.4).

(c) For this simple dispersion relation, there is no dispersion; the group and phase velocities are the same. Explain why this means that the phase must be constant along the rays, $d\phi/dt = 0$. From this fact, deduce that the rays are given by $\{t - r_*, \theta, \phi\}$ = constant. Explain why this means means that $t - r_*$ can be regarded as the *retarded time* for these waves.

(d) Verify that the waves' amplitude satisfies the propagation law (6.35).

Exercise 6.5 *Derivation: Hamilton's Equations for Dispersionless Waves*

Show that Hamilton's equations for the standard dispersionless dispersion relation (6.4) imply the same ray equation (6.51) as we derived using Fermat's principle.

Exercise 6.6 *Problem: Propagation of Sound waves in a Wind*

Consider sound waves propagating in an isothermal atmosphere with constant sound speed c in which there is a horizontal wind shear. Let the (horizontal) wind velocity $\mathbf{u} = u_x\mathbf{e}_x$ increase linearly with height z above the ground according to $u_x = Sz$, where S is the constant shearing rate. Just consider rays in the $x - z$ plane.

(a) Give an expression for the dispersion relation $\omega = \Omega(\mathbf{k}, \mathbf{x}, t)$. [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

$$\left| \frac{\omega}{k_x} - u_x(z) \right| < c \, . \tag{6.54}$$

(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.)

Exercise 6.7 *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

$$n = n_0(1 - \alpha^2 r^2)^{1/2} \tag{6.55}$$

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 (6.51) to show that the radius of the ray is given by

$$r = \frac{\sin\theta}{\alpha} \left| \sin\left(\frac{\alpha z}{\cos\theta}\right) \right| \tag{6.56}$$

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

$$T = \frac{n_0 L}{c}[1 + O(\theta^4)] \tag{6.57}$$

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

Exercise 6.8 *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.

$$-\frac{\hbar}{i}\frac{\partial \psi}{\partial t} = \left[\frac{1}{2m}\left(\frac{\hbar}{i}\nabla\right)^2 + V(\mathbf{x}, t) \right] \psi \, . \tag{6.58}$$

(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 (leading order in a bookkeeping parameter ϵ) give the Hamilton-Jacobi equation

$$\frac{\partial S}{\partial t} + \frac{1}{2m}(\boldsymbol{\nabla} S)^2 + V = 0 \; . \tag{6.59}$$

Our notation $\phi \equiv S/\hbar$ 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 $S = \hbar\phi$ becomes, in that limit, "Hamilton's principal function," which obeys the Hamilton-Jacobi equation.[4]

(b) From this equation 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):

$$\frac{d\mathbf{x}}{dt} = \frac{\mathbf{p}}{m} \; , \quad \frac{d\mathbf{p}}{dt} = -\boldsymbol{\nabla} V \; , \tag{6.60}$$

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

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

$$\frac{d|A|^2}{dt} + |A|^2 \frac{\boldsymbol{\nabla}\cdot\mathbf{p}}{m} = 0 \tag{6.61}$$

Interpret this equation quantum mechanically.

Exercise 6.9 *Example: Energy Density and Flux, and Adiabatic Invariant, for a Dispersionless Wave*

(a) Show that the standard dispersionless scalar wave equation (6.30) follows from the variational principle

$$\delta \int dt d\mathbf{x} \left\{ \rho \left[\frac{n^2}{2c^2} \left(\frac{\partial \psi}{\partial t} \right)^2 - \frac{1}{2} \left(\boldsymbol{\nabla} \psi \right)^2 \right] \right\} = 0 \; , \tag{6.62}$$

where ρ is a constant (the mass density in the case of sound waves traveling through a fluid).

(b) The quantity in square brackets is known as the *Lagrangian density*, $\mathcal{L}$. For any scalar-field Lagrangian $\mathcal{L}(\psi, \boldsymbol{\nabla}\psi, \mathbf{x}, t)$, there is a *canonical*, relativistic procedure for constructing a stress-energy tensor:

$$T_\mu{}^\nu = -\frac{\partial \mathcal{L}}{\partial \psi_{,\nu}} \psi_{,\mu} + \delta_\mu{}^\nu \mathcal{L} \; . \tag{6.63}$$

[4]See, e.g., Chap. 10 of Goldstein (1980).

Show that, if $\mathcal{L}$ has no explicit time dependence (e.g., for the Lagrangian of Eq. (6.62) if $n = n(\mathbf{x})$ does 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 n is independent of x), then the field's momentum is conserved, $T^{j\nu}{}_{,\nu} = 0$.

(c) Show that expression (6.63) for the field's energy density $\varepsilon = T^{00}$ and its energy flux $F_i = T^{0i}$ agrees with Eqs. (6.42) and (6.43).

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

$$J \equiv \int_0^{2\pi/\omega} \int \Pi(\partial\psi/\partial t) d^3x \, dt \; , \tag{6.64}$$

which is the continuum analog of Eq. (6.47). Note that 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 (6.47), 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 (6.64) is also an adiabatic invariant. This means that the wave action (and thence also the number of quanta in the waves) is conserved when the medium [in our case the index of refraction $n(\mathbf{x})$] 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 of this adiabatic invariant, Eq. (6.47) in detail when we analyze charged particle motion in a magnetic field in Chap. 19.]

****

6.4 Paraxial Optics

It is quite common in optics to be concerned with a bundle of rays that are almost parallel. This implies that the angle that the rays make with some reference ray can be treated as small—an approximation that underlies the first order theory of simple optical instruments like the telescope and the microscope. This approximation is called *paraxial optics*, and it permits one to linearize the geometric optics equations and use matrix methods to trace their rays.

We shall develop the paraxial optics formalism for waves whose dispersion relation $\omega = \Omega$ has the simple, time-independent, nondispersive form $\Omega = kc/n(\mathbf{x})$. Recall that 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.

We start by linearizing the ray propagation equation (6.51). 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. Under paraxial conditions, $|\mathbf{x}|$ is small compared to the z-lengthscales of the propagation. Now, let us Taylor expand the refractive index, $n(\mathbf{x}, z)$.

$$n(\mathbf{x}, z) = n(0, z) + x_i n_{,i}(0, z) + \frac{1}{2} x_i x_j n_{,ij}(0, z) + \dots \ , \tag{6.65}$$

where the subscript commas denote partial derivatives with respect to the transverse coordinates, $n_{,i} \equiv \partial n / \partial x_i$. The linearized form of Eq. (6.51) is then given by

$$\frac{d}{dz}\left(n(0, z)\frac{dx_i}{dz}\right) = n_{,i}(0, z) + x_j n_{,ij}(0, z) \ . \tag{6.66}$$

It is helpful to regard z as "time" and think of Eq. (6.66) as an equation for the two dimensional simple harmonic motion of a particle (the ray) in a quadratic potential well.

We are usually concerned with aligned optical systems in which there is a particular choice of reference ray called the *optic axis*, for which the term $n_{,i}(0, z)$ on the right hand side of Eq. (6.66) vanishes. If we choose the reference ray to be the optic axis, then Eq. (6.66) is a linear, homogeneous, second-order equation for $\mathbf{x}(z)$,

$$(d/dz)(n dx_i/dz) = x_j n_{,ij} \tag{6.67}$$

which we can solve 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 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 $V_1 = x$, $V_2 = \dot{x}$, $V_3 = y$, $V_4 = \dot{y}$—and embodying the linear transformation from location z' to location z in a *transfer matrix* $J_{ab}(z, z')$:

$$V_a(z) = J_{ab}(z, z') \cdot V_b(z'). \tag{6.68}$$

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:

$$J_{ac}(z, z') = J_{ab}(z, z'') J_{bc}(z'', z'). \tag{6.69}$$

6.4.1 Axisymmetric, Paraxial Systems

If the index of refraction is everywhere axisymmetric, so $n = n(\sqrt{x^2 + y^2}, z)$, 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.

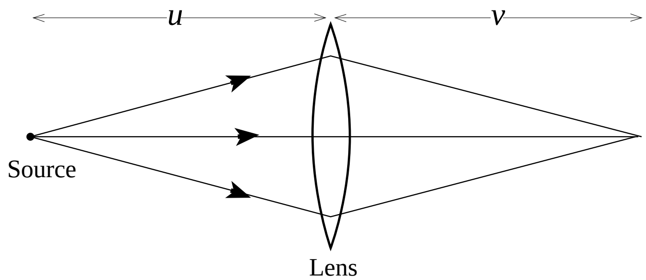

Fig. 6.5: Simple converging lens used to illustrate the use of transfer matrices.The total transfer matrix is formed by taking the product of the straight section transfer matrix with the lens matrix and another straight section matrix.

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

$$V_a = \begin{pmatrix} x \\ \dot{x} \end{pmatrix} \tag{6.70}$$

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

$$x = x' + \dot{x}'d \tag{6.71}$$

$$\dot{x} = \dot{x}' \tag{6.72}$$

so

$$J_{ab} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}, \tag{6.73}$$

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:

$$J_{ab} = \begin{pmatrix} 1 & 0 \\ -f^{-1} & 1 \end{pmatrix}. \tag{6.74}$$

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

$$J_{ab} = \begin{pmatrix} 1 & 0 \\ 2R^{-1} & 1 \end{pmatrix}. \tag{6.75}$$

As a simple illustration let us consider rays that leave a point source which is located a distance u in front of a converging lens of focal length f and solve for the ray positions a distance v behind the lens (Fig. 6.5). The total transfer matrix is the product of the transfer matrix for a straight section, Eq. (6.73) with the product of the lens transfer matrix and a second straight-section transfer matrix:

$$J_{ab} = \begin{pmatrix} 1 & v \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -f^{-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 - vf^{-1} & u + v - uvf^{-1} \\ -f^{-1} & 1 - uf^{-1} \end{pmatrix} \tag{6.76}$$

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

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \tag{6.77}$$

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

$$M = -\frac{v}{u} \ , \tag{6.78}$$

where the negative sign indicates that the image is inverted. Note that the system does not change with time, so we could have interchanged the source and the image planes.

6.4.2 Converging Magnetic Lens

Since geometric optics is the same as particle dynamics, these matrix equations can be used for paraxial motions of electrons and 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.) Quadrupolar magnetic fields are used to guide the particles around the storage ring. Since these magnetic fields are not axisymmetric, to analyze them we must deal with a four-dimensional vector $\mathbf{V}$.

The simplest, practical magnetic lens is quadrupolar. If we orient our axes appropriately, the magnetic field can be expressed in the form

$$\mathbf{B} = \frac{B_0}{r_0}(y\mathbf{e}_x + x\mathbf{e}_y) \ . \tag{6.79}$$

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

$$\ddot{x} = -\frac{x}{\lambda^2} \ , \quad \ddot{y} = \frac{y}{\lambda^2} \ , \tag{6.80}$$

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

$$\lambda = \left(\frac{pr_0}{qB_0}\right)^{1/2} \tag{6.81}$$

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

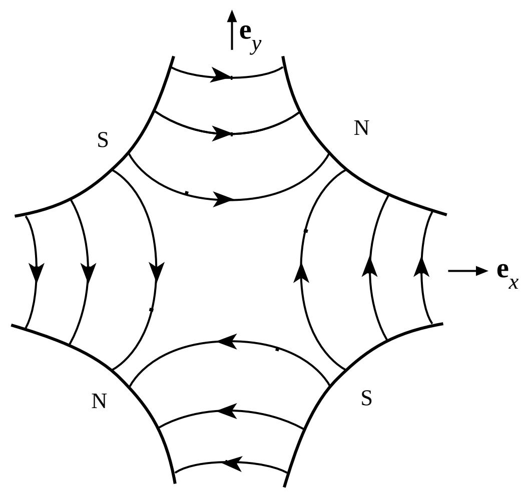

Fig. 6.6: Quadrupolar Magnetic Lens. The magnetic field lines lie in a plane perpendicular to the optic axis. Positively charged particles moving along $\mathbf{e}_z$ are converged when $y = 0$ and diverged when $x = 0$.

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 (6.80), we infer that the transfer matrices from the magnet's entrance to its exit are $J_{x\,ab}$, $J_{y\,ab}$, where

$$J_{x\,ab} = \begin{pmatrix} \cos\phi & \lambda\sin\phi \\ -\lambda^{-1}\sin\phi & \cos\phi \end{pmatrix} \tag{6.82}$$

$$J_{y\,ab} = \begin{pmatrix} \cosh\phi & \lambda\sinh\phi \\ \lambda^{-1}\sinh\phi & \cosh\phi \end{pmatrix} \tag{6.83}$$

and $\phi = L/\lambda$ with L the distance from entrance to exit.

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

$$J_{x\,ab} = \begin{pmatrix} 1 & \lambda\tan\phi/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\sin\phi/\lambda & 1 \end{pmatrix} \begin{pmatrix} 1 & \lambda\tan\phi/2 \\ 0 & 1 \end{pmatrix} \tag{6.84}$$

$$J_{y\,ab} = \begin{pmatrix} 1 & \lambda\tanh\phi/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \sinh\phi/\lambda & 1 \end{pmatrix} \begin{pmatrix} 1 & \lambda\tanh\phi/2 \\ 0 & 1 \end{pmatrix} \tag{6.85}$$

Comparing with Eqs. (6.73), (6.74), 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 the motion in the x direction first. Let $f_+ = \lambda/\sin\phi$ be the equivalent focal length of the first converging lens and $f_- = -\lambda/\sinh\phi$ that of 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

$$\begin{pmatrix} 1 & 0 \\ -f_-^{-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -f_+^{-1} & 1 \end{pmatrix} = \begin{pmatrix} 1 - df_+^{-1} & d \\ -f_*^{-1} & 1 - df_-^{-1} \end{pmatrix} \tag{6.86}$$

where

$$\frac{1}{f_*} = \frac{1}{f_-} + \frac{1}{f_+} - \frac{d}{f_- f_+} \tag{6.87}$$

$$= \frac{\sin\phi}{\lambda} - \frac{\sinh\phi}{\lambda} + \frac{d\sin\phi\sinh\phi}{\lambda^2}\ . \tag{6.88}$$

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

$$f_* \simeq \frac{3\lambda}{2\phi^3} = \frac{3\lambda^4}{2L^3}\ . \tag{6.89}$$

The effective focal length of the combined magnets, f_* is positive and so the lens has a net focussing effect. From the symmetry of Eq. (6.88) 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 fairly 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 and particle accelerators.

****

EXERCISES

Exercise 6.10 *Problem: Matrix Optics for a Simple Refracting Telescope*

Consider a simple refracting telescope that comprises two thin converging lenses and that takes parallel rays of light from distant stars which make an angle θ with the optic axis and converts them into parallel rays making an angle $-M\theta$ where $M \gg 1$ is the magnification (Fig. 6.7).

(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 M depend on the ratio of the focal lengths of the two lenses?

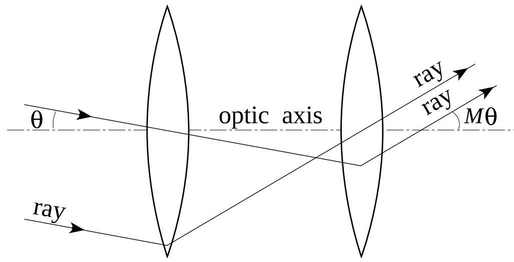

Fig. 6.7: Simple refracting telescope.

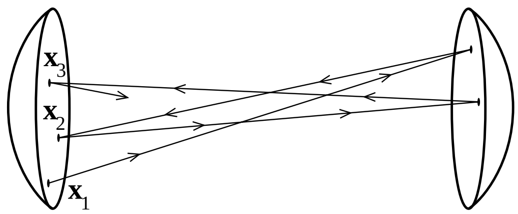

Fig. 6.8: An optical cavity formed by two mirrors, and a light beam bouncing back and forth inside it.

Exercise 6.11 *Example: 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. 6.8. A laser beam bounces back and forth between the two mirrors. The center of the beam travels along a geometric-optics ray.

(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 2D vector in the plane perpendicular to the optic axis), which satisfy the difference equation

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

where

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

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

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

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

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.)

****

6.5 Polarization and the Berry Phase

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 wave properties that are unique to vector (or tensor) waves. Most of these come under the heading of *polarization* effects. In Part VI we shall study polarization effects for (tensorial) gravitational waves. Here and in several other chapters we shall examine them for electromagnetic waves.

An electromagnetic wave *in vacuo* has its electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ perpendicular to its propagation direction $\hat{\mathbf{k}}$ and perpendicular to each other. In a medium, $\mathbf{E}$ and $\mathbf{B}$ 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 mmagnetic 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 $\Omega = (c/n)k$, 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 $\hat{\mathbf{k}}$

$$\mathbf{E} = Ae^{i\phi}\,\hat{\mathbf{f}}\ , \quad \hat{\mathbf{f}}\cdot\hat{\mathbf{k}} \equiv \hat{\mathbf{f}}\cdot\boldsymbol{\nabla}\phi = 0\ . \tag{6.90}$$

In the eikonal approximation, $Ae^{i\phi} \equiv \psi$ satisfies the geometric-optics propagation laws of Sec. 6.3, 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) 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. (ii) 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

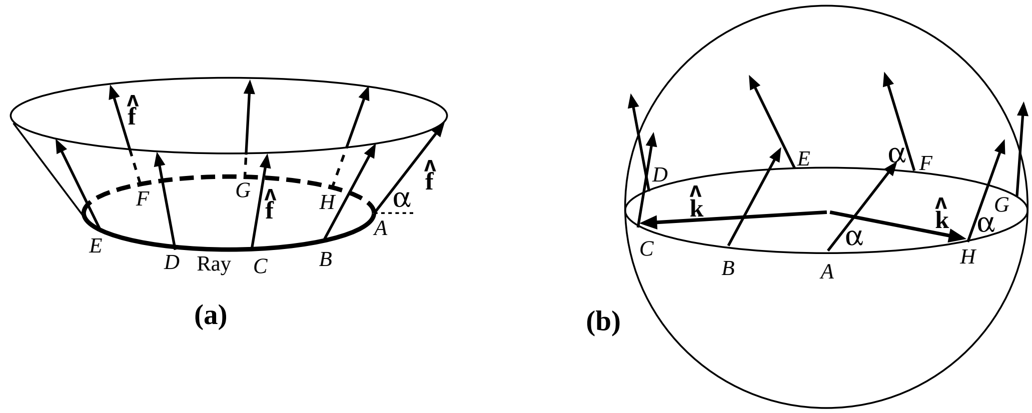

Fig. 6.9: (a) The ray along the optic axis of a circular loop of optical fiber, and the polarization vector $\hat{\mathbf{f}}$ that is transported along the ray by the geometric-optics transport law $d\hat{\mathbf{f}}/ds = -\hat{\mathbf{k}}(\hat{\mathbf{f}} \cdot d\hat{\mathbf{k}}/ds)$. (b) The polarization 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)].

implies that the propagation law for $\hat{\mathbf{f}}$ is

$$\frac{d\hat{\mathbf{f}}}{ds} = -\hat{\mathbf{k}} \left(\hat{\mathbf{f}} \cdot \frac{d\hat{\mathbf{k}}}{ds} \right) . \tag{6.91}$$

We say that the vector $\hat{\mathbf{f}}$ is *parallel-transported* along $\hat{\mathbf{k}}$. Here "parallel transport" means: (i) Carry $\hat{\mathbf{f}}$ a short distance along the trajectory keeping it parallel to itself in 3-dimensional space. 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 surfaces, which we shall develop in Part VI when studying general relativity, give powerful mathematical tools for analyzing this parallel transport.)

6.5.1 Berry Phase

We shall use the polarization propagation law (6.91) to illustrate a quite general phenomenon known as the *Berry* (or geometric or adiabatic or anholonomic) phase.[5]

Consider linearly polarized, monochromatic light beam that propagates in an optical fiber. The fiber's optic axis is the principal ray along which the light propagates. 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. (6.91).

If the fiber and ray are straight, then the propagation law (6.91) keeps $\hat{\mathbf{f}}$ constant. If the fiber and ray are circular, then the propagation law (6.91) causes $\hat{\mathbf{f}}$ to rotate in such a way as to always point along the generator of a cone as shown in Fig. 6.9 (a). This polarization behavior, and that for any other ray shape, can be deduced with the aid of a unit sphere

[5]Berry (1990).

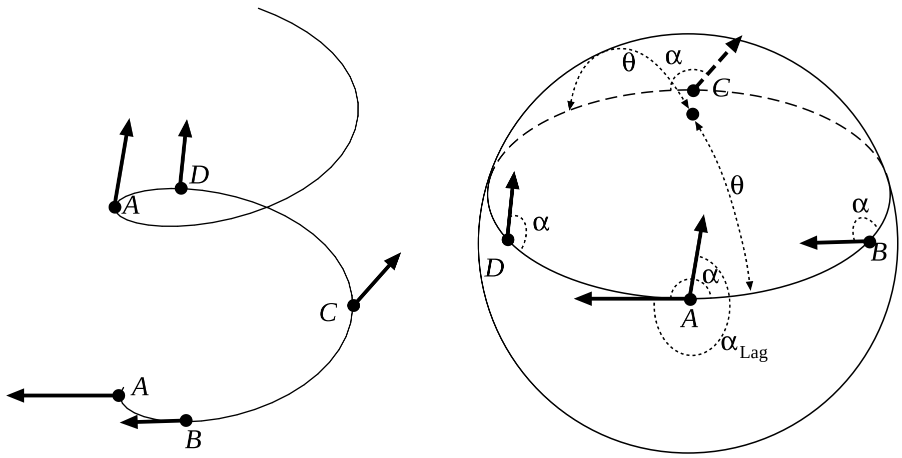

Fig. 6.10: (a) The ray along the optic axis of a helical loop of optical fiber, and the polarization vector $\hat{\mathbf{f}}$ that is transported along the ray by the geometric-optics transport law $d\hat{\mathbf{f}}/ds = -\hat{\mathbf{k}}(\hat{\mathbf{f}} \cdot d\hat{\mathbf{k}}/ds)$. The ray's propagation 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).

on which we plot the ray direction $\hat{\mathbf{k}}$ [Fig. 6.9 (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. This is because the ray in physical space is a closed circle. If, instead, the fiber and ray were bent into a helix (Fig. 6.10 below), then the trajectory on the unit sphere would be a smaller 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).

For the circular, closed ray of Fig. 6.9 (a), the parallel transport law keeps constant the angle α between $\hat{\mathbf{f}}$ and the trajectory of $\hat{\mathbf{f}}$ [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.

For the helical ray of Fig. 6.10 (a), 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 Fig. 6.10 (b) is a circle of constant polar angle θ. In this case (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

$$d\alpha/d\phi = \cos\theta \ . \tag{6.92}$$

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 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 $\theta \simeq 0$ (a nearly straight ray), the transport equation (6.92) predicts $d\alpha/d\phi = 1$: although $\hat{\mathbf{f}}$ remains 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. For an arbitrary helical pitch angle θ, the propagation equation (6.92) predicts that during one round trip α will change by an amount $2\pi\cos\alpha$ 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:

$$\alpha_{\rm Lag} = \Delta\Omega \ . \tag{6.93}$$

For the circular ray of Fig. 6.9, 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.

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 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 $\Delta\phi = \int \mathbf{k}\cdot d\mathbf{x}$ (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 6.12 *Derivation: Parallel-Transport*

Use the parallel-transport law 6.91 to derive the relation 6.92.

6.6 Caustics and Catastrophes—Gravitational Lenses

Albert Einstein's General relativity theory (Part VI of this book) predicts that light rays should be deflected by the gravitational pull of the Sun. Newton's law of gravity and his corpuscular theory of light also predict such a 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 cosmologically distant galaxy to behave like a crude lense 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 and also the properties of caustics.[6]

6.6.1 Formation of Multiple Images by a Gravitational Lens

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 which is adequate for today's astronomical applications. In this model, curved spacetime behaves as if it were flat but endowed with a refractive index given by

$$n = 1 - \frac{2\Phi}{c^2} \tag{6.94}$$

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 $\Phi = -GM/r$). We will justify this index-of-refraction model in Part VI.

First, let us understand the order of magnitude of the effect. Consider a ray which passes by a point mass M with an impact parameter b. The ray trajectory is given by solving Eq. (6.67), $d/dz(nd\mathbf{x}/dz) = (\mathbf{x}\cdot\boldsymbol{\nabla})(\boldsymbol{\nabla}n)$, where $\mathbf{x}(z)$ is the ray's transverse position relative to an optic axis that passes through the point mass, and z is distance along the optic axis.

The ray will be bent through a deflection angle α, cf. Fig. 6.11. An equivalent way of expressing the motion is to say that the photons moving with speed c are subject to a Newtonian force and are accelerated kinematically with twice the Newtonian acceleration. The problem is therefore just that of computing the deflection of a charged particle passing by an oppositely charged particle. The deflection, under the impulse approximation, is given by

$$\alpha = \frac{4GM}{bc^2} = \frac{-4\Phi(r = b)}{c^2} \,, \tag{6.95}$$

where b is the ray's impact parameter. For a ray passing close to the limb of the sun, for which the potential will be indistinguishable from the spherical potential from a point, this deflection is 1.75 arc seconds.

Now let us consider a galaxy as the gravitational deflector. We can use the virial theorem to make an order of magnitude estimate of Φ and relate it to the mean square velocity of the constituent stars (measured in one dimension) σ. This quantity can be measured spectroscopically. We find that $\Phi \sim -\sigma^2$. Therefore, an order of magnitude estimate of the angle of deflection is $\alpha \sim \sigma^2/c^2$. If we do a more careful calculation for a simple model of a

[6] Blandford & Narayan (1992).

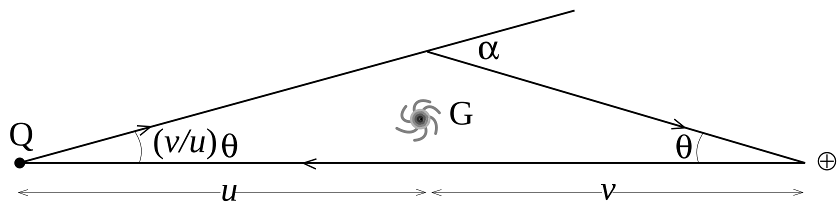

Fig. 6.11: Geometry for a gravitational lens. Light from a distant quasar, Q treated as a point source, passes by a galaxy G and is deflected through an angle α on its way to earth $\oplus$. The galaxy is a distance u from the quasar and v from earth.

galaxy in which the mass density varies inversely with the distance from the center, then we obtain

$$\alpha \sim \frac{4\pi\sigma^2}{c^2} \tag{6.96}$$

For typical galaxies, $\sigma \sim 300$ km s^{-1} and $\alpha \sim 1-2$ arc sec. The paraxial approximation therefore is fully justified. Now a cosmologically distant galaxy lies at a distance $D \sim 3 \times 10^{25}$m from earth and so the tranverse displacement of the ray due to the galaxy is $\sim D\alpha \sim 3 \times 10^{20}$m, which is still within the galaxy. This means that light from a quasar lying behind the galaxy can pass by either side of the galaxy. We should then see at least two distinct images of the quasar separated by an angular distance $\sim \alpha$.

The imaging is illustrated in Fig. 6.11. First trace a ray backward from the observer, in the absence of the intervening galaxy, to the quasar. We call this the reference ray. (We will ignore the fact that the universe is expanding and possesses a curved spacetime. This introduces unimportant corrections.) Now interpose a galaxy a distance v from the observer and a distance u from the quasar. Consider a *virtual ray* that propagates at an angle $\boldsymbol{\theta}$, a 2D vector on the sky, to the reference ray in a straight line from the earth to the galaxy where it is deflected toward the quasar. (A virtual ray is a path that will become a real ray if it satisfies Fermat's principle.) The optical phase for light propagating along this virtual ray will exceed that along the reference ray by an amount $\Delta\phi$ called the *phase delay*. There are two contributions to $\Delta\phi$: First, the geometrical length of the path is longer than the reference ray by an amount $(u+v)v\theta^2/2u$ (cf. Fig. 6.11), 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 $\int(n-1)ds/c = -2\int \Phi ds/c$, where ds is an element of length along the path. We can express this second delay as $2\Phi_2/c^3$ where $\Phi_2 = \int \Phi ds$ is the two dimensional (2D) Newtonian potential. Φ_2 can be computed from the 2D Poisson equation

$$\nabla^2\Phi_2 = 4\pi G\Sigma \tag{6.97}$$

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

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

$$\Delta\phi = \omega\left(\frac{(u+v)v}{2uc}\theta^2 - \frac{2\Phi_2(\boldsymbol{\theta})}{c^3}\right) \; . \tag{6.98}$$

We can now invoke Fermat's principle. Of all possible virtual rays, parametrized by the angular coordinate $\boldsymbol{\theta}$, the only ones that correspond to real rays are those for which the

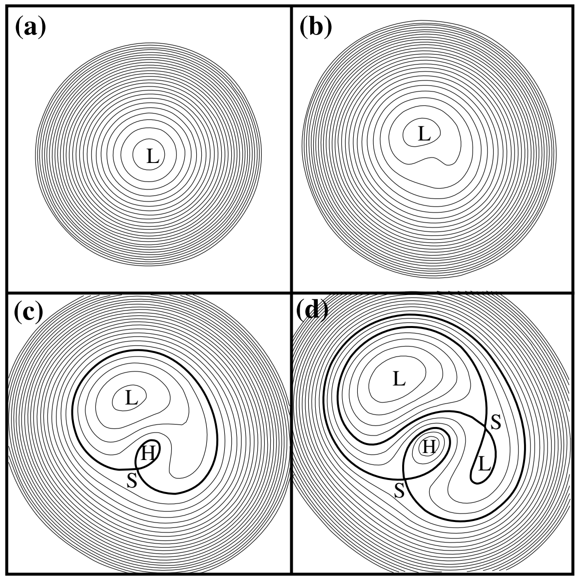

Fig. 6.12: Contour plots of the phase delay $\Delta\phi(\boldsymbol{\theta})$ for four different gravitational lenses. a) In the absence of a lens $\Phi_2 = 0$, the phase delay (6.98) has a single minimum corresponding to a single undeflected image. b) When a small galaxy with a shallow potential Φ_2 is interposed, it pushes the phase delay $\Delta\phi$ up in its vicinity [Eq. (6.98) 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 and two more images will be produced. d) If the potential well deepens even more, five images can be produced.

phase difference is stationary, i.e. those for which

$$\frac{\partial \Delta\phi}{\partial \theta_j} = 0 \,, \tag{6.99}$$

where θ_j (with $j = x, y$) are the Cartesian components of $\boldsymbol{\theta}$. Differentiating Eq. (6.98) we obtain a 2D vector equation for the location of the images.

$$\theta_j = \frac{2u}{(u+v)vc^2}\frac{\partial \Phi_2}{\partial \theta_j}. \tag{6.100}$$

Referring to Fig. 6.11, we can identify the deflection angle

$$\boldsymbol{\alpha}_j = \frac{2}{vc^2}\frac{\partial \Phi_2}{\partial \theta_j} \tag{6.101}$$

We can understand quite a lot about the properties of the images by inspecting a contour plot of the phase delay function $\Delta\phi(\boldsymbol{\theta})$ (Fig. 6.12). When the galaxy is very light or quite distant from the line of sight, then there is a single minimum in the phase delay. 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. Note that with a transparent galaxy, the additional images are created in pairs. Note in addition that the stationary points are not necessarily minima, which is inconsistent with Fermat's original statement of his principle, but there are images at the stationary points nevertheless.

Now suppose that the quasar is displaced by a small angle $\delta\boldsymbol{\theta}'$. This is equivalent to moving the lens by a small angle $-\delta\boldsymbol{\theta}'$. Equation (6.100) says that the image will be displaced by a small angle $\delta\boldsymbol{\theta}$ satisfying the equation

$$\delta\theta_i - \delta\theta_i' = \frac{2u}{(u+v)vc^2}\frac{\partial^2\Phi_2}{\partial\theta_i\partial\theta_j}\delta\theta_j \ . \tag{6.102}$$

By combining with Eq. (6.98), we can rewrite this as

$$\delta\theta_i' = H_{ij}\delta\theta_j \ , \tag{6.103}$$

where the matrix H_{ij} is

$$H_{ij} = \left(\frac{uc/\omega}{(u+v)v}\right)\frac{\partial^2\Delta\phi}{\partial\theta_i\partial\theta_j} \ . \tag{6.104}$$

Now consider a small rectangular area of source $d\theta_1' d\theta_2'$ (measured in sterradians). Its image will have area $d\theta_1 d\theta_2$. The ratio of the image area to the source area is just the magnification, M, the ratio of the flux observed from the source to that which would have been observed in the absence of the lens. However, from Eq. (6.103) we see that the magnification is just the determinant of the inverse of the matrix H_{ij}. Equivalently,

$$M = \frac{1}{||H_{ij}||} \ . \tag{6.105}$$

The curvature of the phase delay surface (embodied in $||\partial^2\Delta\phi/\partial\theta_i\partial\theta_j||$) is therefore a quantitative measure of the magnification. Small curvature implies large magnification of the images and *vice versa*. 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. 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 at the stationary points is equal to ω times the extra time it takes a signal to arrive along that ray. To order of magnitude, the time delay difference will be $\sim v\alpha^2/c \sim 1$ yr. 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. This, in turn, may allow us to measure the distance to the quasar and consequently provide an accurate unit of length for the universe. These measurements are currently being made.

6.6.2 Catastrophe Optics — Formation of Caustics

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 detail, bring light to a focus on a 2D surface in 3D space, called

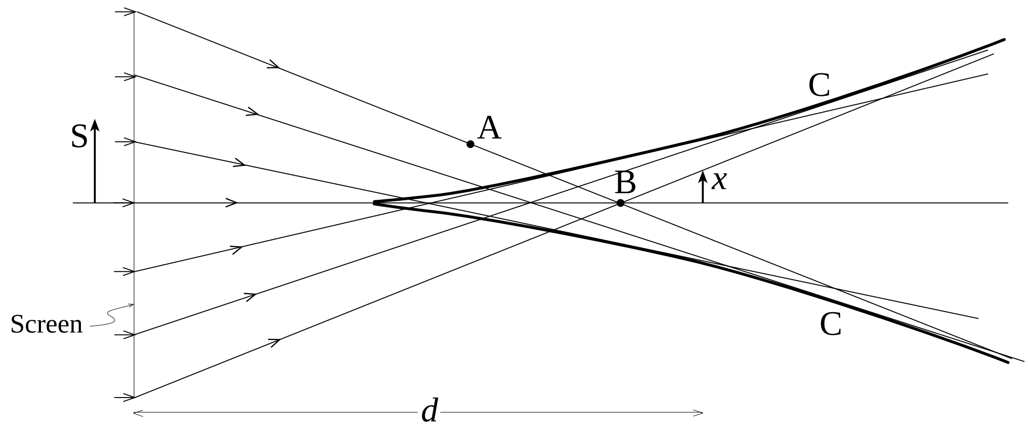

Fig. 6.13: The formation of caustics by a circularly symmetric lens. Light from a distant source is refracted at the plane of the lens. The envelope of the refracted rays forms a caustic surface C. An observer at a point A outside the caustic will see a single image of the distant source, whereas one at point B between the caustics will see three images. If the observer at B moves toward the caustic, then he will see two of the images approach each other, merge and then vanish. If the source has a finite angular size, the angular size of the two images will increase as they merge, and the energy flux from the two images will also increase. In this example, the caustic terminates in a cusp point. THIS DIAGRAM MUST BE REDRAWN.

a *caustic*.[7] Caustics are quite familiar and can be formed when sunlight is refracted or reflected by the choppy water on the surface of a swimming pool. The bright lines one sees on the pool's bottom are intersections of 2D caustics with the pool's 2D surface Another good example is the cusped curve (called a nephroid) formed by light from a distant source reflected off the cylindrical walls of a teacup onto the surface of the tea. What is surprising is that caustics formed under quite general conditions exhibit a surprising universality.

For simplicity let us consider the problem of the refraction of light by an axisymmetric, converging lens, for example a gravitational lens (c.f. Fig. 6.13). Consider a set of rays from a distant source with impact parameter s at the lens. Let these rays pass through a point of observation a distance d from the lens with radial coordinate $x = s - \theta d << d$. As we have just shown the true rays will be those for which the total phase $\phi(s, x)$ is stationary with respect to variations of s. Now, if x is small enough and d is large enough, then there will typically be three rays that pass through any point of observation. In the case of a gravitational lens, the astronomer would see three images of the source. However, when x is large and the astronomer is well away from the optic axis, there will only be one ray and one image. There is therefore an axisymmetric surface, called a *caustic* where the number of images changes from one to three.

Now let us consider the behavior of the phase ϕ as we cross this caustic. From Fig. 6.13, it is clear that the two disappearing images approach one another and then vanish. Algebraically, this means that, by changing the parameter (often called a *control* variable) x, the variation of $\phi(s, x)$ with s (often called a *state* variable) changes locally through a set of curves like those in Fig. 6.14, where a maximum and a minimum smoothly merge through

[7] See, for example, Berry (1982).

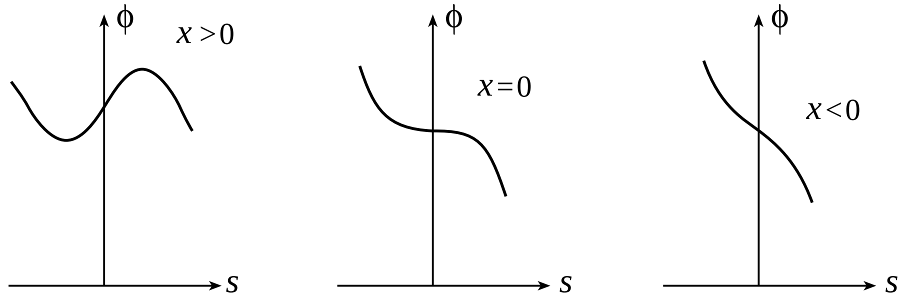

Fig. 6.14: Optical phase for three different observer locations measured by the transverse coordinate x. The true rays are refracted at the values of s corresponding to the maxima and minima of the phase.

a point of inflexion and then vanish. It is clear that close enough to the caustic, $\phi(s,x)$, for given s, has the form of a cubic. By changing the origins of s and x, this cubic can be written in the form of a Taylor series for which the leading terms are:

$$\phi(s;d,x) = \frac{1}{3}as^3 - bxs + \ldots \tag{6.106}$$

where the factor 1/3 is just a convention and we have dropped a constant. Note that, by changing coordinates, we have removed the quadratic terms. Now, 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 as we are only concerned with scaling laws.

Now, invoking Fermat's Principle and differentiating Eq. (6.106) with respect to s, we see that there are two true rays and two images for $x > 0$, (passing through $s = \pm(bx/a)^{1/2}$), and no images for $x < 0$. $x = 0$ marks the location of the caustic at this distance behind the lens. We can now compute the magnification of the images as the caustic is approached. This is given by

$$M \propto \frac{ds}{dx} = \frac{1}{2}\left(\frac{b}{ax}\right)^{1/2} \quad . \tag{6.107}$$

Notice that the magnification, and thus also the total flux in each image, scales inversely with the square root of the distance from the caustic. This does not depend on the optical details (i.e. on the coefficients in our power series expansions). It therefore is equally true for reflection at a spherical mirror, or refraction by a gravitational lens, or refraction by the rippled surface of the water in a swimming pool. 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*, and caustics are examples of *catastrophes*. In this 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 the solution should be structurally stable. That is to say, if we make small changes in the physical conditions, the scaling laws etc are robust.

The catastrophe that we have just considered is the most elementary example and is called the *fold*. The next simplest catastrophe, known as the *cusp*, is the curve where two fold surfaces meet. (The point cusp dsiplayed in Fig. 6.14, is actually structurally unstable as a consequence of the assumed strict axisymmetry. However if we regard s, x as just 1D Cartesian coordinates, then Fig. 6.14 provides a true representation of the geometry.) In total there are seven elementary catastrophes. Catastrophe theory has many interesting applications in optics, dynamics, and other branches of physics.

Let us conclude with an important remark. If we have a point source, the magnification will diverge to infinity as the caustic is approached, according to Eq. 6.107. 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. Diffraction is the subject of the next chapter.

****

EXERCISES

Exercise 6.13 *Example: Point-mass gravitational lens*

Consider a point mass M that is located a distance D from us and acts as a gravitational lens to produce multiple images of a very small, distant source of light.

(a) Use Eq. (6.95) to show that when the source lies on the continuation of the observer-lens line, it will produce a thin-ring image at the observer of angular radius

$$\theta_E = \left(\frac{4GM}{Dc^2}\right)^{1/2} . \tag{6.108}$$

(This ring is known as the Einstein ring.)

(b) Show that when the source is displaced from this line, there will be just two images, one lying within the Einstein ring, the other lying outside. Find their locations.

(c) Denote the ratio of the fluxes in these two images by R. Show that the angular radii of the two images can be expressed in the form $\theta_\pm = \pm\theta_E R^{\pm 1/4}$.

Exercise 6.14 *Challenge: Catastrophe Optics of an Elliptical 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.

****

Bibliography

Born, M. & Wolf, E. 1970 *Principles of Optics* Oxford: Oxford University Press

Berry, M. V. 1990 Physics Today 43:12 34

Berry, M. V. & Upstill, C. 1980 Progress in Optics 18

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

Goldstein, H. 1980 *Classical Mechanics* New York: Addison Wesley

Guenther, R. 1990 *Modern Optics* New York: Wiley

Hecht, E. 1987 *Optics* New York: Wiley

Klein, M. V. & Furtak, T. E. 1986 *Optics*New York: Wiley

Welford, W. T. 1990 *Optics* Oxford: Oxford Science Publications