Chapter 7

Diffraction

[This is version 0007.2, 8 November 2000]

7.1 Overview

The previous chapter was devoted to the classical mechanics of wave propagation. We showed how a classical wave equation can be solved in the short wavelength approximation to yield Hamilton's dynamical equations. We then specialized to stationary media, as we shall continue to do in this chapter. Under these conditions, the frequency of a wave packet is constant. We imported a result from classical mechanics, the principle of stationary action, to show that the true geometric-optics rays were those paths along which the action or the integral of the phase was stationary. Our physical interpretation of this result was that the waves did indeed travel along every path, from some source to a point of observation, where they were added together but they only gave a significant net contribution when they could add in phase, along the true rays. This is, essentially, Huygens' model of wave propagation, or, in modern language, a *path integral*.

Huygens' principle asserted that every point on a wave front acted as a source of secondary waves that combine so that their envelope constitutes the advancing wave front. This principle must be supplemented by two ancillary conditions, that the secondary waves are only formed in the direction of wave propagation and that a 90° phase shift be introduced into the secondary wave. The reason for the former condition is obvious, that for the latter, less so. We shall discuss both together with the formal justification of Huygens' construction below.

In this chapter, we begin our exploration of the "wave mechanics" of optics. This differs increasingly from geometric optics as the wave frequency decreases. The number of paths that can combine constructively increases and the rays that connect two points become blurred. In quantum mechanics, we recognize this phenomenon as the uncertainty principle and it is just as applicable to photons as to electrons. Solving the wave equation exactly is almost always too hard except in very simple circumstances. As we have emphasized, geometric optics, like classical mechanics, is one approximate method of solving the fundamental wave equation in the short wavelength limit. We must now develop useful approximate techniques when geometric optics becomes invalid.

We begin by making a somewhat artificial distinction between phenomena that arise when an effectively infinite number of paths are involved which we call *diffraction* and which we describe in this chapter, and those when a few paths, or, more correctly, a few tight bundles of rays are combined which we term *interference*, and whose discussion we defer to to the next chapter.

In Sec. 7.2, we shall present the Fresnel-Helmholtz-Kirchhoff theory that underlies most elementary discussions of diffraction, and we shall then distinguish between Fraunhofer diffraction (the limiting case when spreading of the wavefront mandated by the uncertainty principle is very important), and Fresnel diffraction (which arises when the diffracting screen is observed from much closer and it is the phase variation across the screen that is the dominant physical effect). In Sec. 7.3, we shall illustrate Fraunhofer diffraction by computing the expected angular resolution of Hubble Space Telescope and in Sec. 7.4, we shall analyze Fresnel diffraction and illustrate it using lunar occultation of radio waves and zone plates. Many contemporary optical devices are regarded as linear systems that transform an input wave signal and produce a linearly related output. Their operation, particularly as image processing devices can be considerably enhanced by processing the signal in the Fourier domain, a procedure known as spatial filtering. We shall illustrate these ideas in Sec. 7.5 using the phase contrast microscope and Gaussian beams. Finally, in Sec. 7.6 we shall analyze the effects of diffraction near a caustic of a wave's phase field, where geometric optics incorrectly predicts a divergent magnification of the wave. As we shall see, diffraction makes the magnification finite.

7.2 Helmholtz- Kirchhoff Integral

In this section, we shall derive a formalism for describing diffraction. We shall restrict our attention to the simplest (and fortunately the most widely useful) case: a scalar wave with field variable ψ of frequency $\omega = ck$ that satisfies the Helmholtz equation

$$\nabla^2\psi + k^2\psi = 0 \tag{7.1}$$

except at boundaries. Generally ψ will represent a real valued physical quantity (although it may, for mathematical convenience, be given a complex representation). This is in contrast to a quantum mechanical wave function satisfying the Schrödinger equation which is an intrinsically complex function. The wave is monochromatic and non-dispersive and the medium is isotropic and homogeneous so that k can be treated as constant. Each of these assumptions can be relaxed with some technical penalty.

The scalar formalism that we shall develop based on Eq. (7.1) is fully valid for weak sound waves in a fluid, e.g. air (Chap. 15). It is also fairly accurate, but not precisely so, for the most widely used application of diffraction theory: the propagation of electromagnetic waves in vacuo or in a medium with homogeneous dielectric constant. In this case ψ can be regarded as one of the Cartesian components of the electric field vector, e.g. E_x (or equally well a Cartesian component of the vector potential or the magnetic field vector). In vacuo or in a homogeneous dielectric medium, Maxwell's equations imply that this $\psi = E_x$ satisfies the scalar wave equation and thence, for fixed frequency, the Helmholtz equation (7.1).

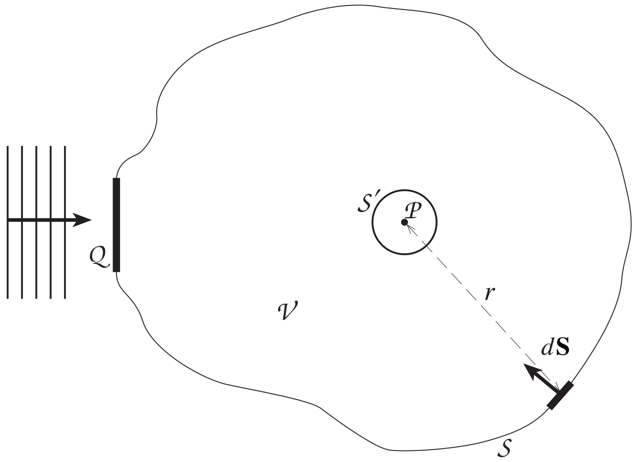

Fig. 7.1: Surface $\mathcal{S}$ for Helmholtz-Kirchhoff Integral. The surface $\mathcal{S}'$ surrounds the observation point $\mathcal{P}$ and $\mathcal{V}$ is the volume bounded by these two surfaces. The aperture $\mathcal{Q}$ and incoming wave to the left of it are irrelevant to the formulation of the Helmholtz-Kirchoff integral, but appear in subsequent applications.

However, when the wave hits a boundary of the medium (e.g. the edge of an aperture, or the surface of a mirror or lens), its interaction with the boundary can couple the various components of $\mathbf{E}$, thereby invalidating the simple scalar theory we shall develop. Fortunately, this polarizational coupling is usually very weak in the paraxial (small angle) limit, and also under a variety of other circumstances, thereby making our simple scalar formalism quite accurate.

The Helmholtz equation (7.1) is an elliptic, linear, partial differential equation, and it thus permits one to express the value $\psi_\mathcal{P}$ of ψ at any point $\mathcal{P}$ inside some closed surface $\mathcal{S}$ as an integral over $\mathcal{S}$ of some linear combination of ψ and its normal derivative; see Fig. 7.1). To derive such an expression, we first augment the actual wave ψ in the interior of $\mathcal{S}$ with a second solution of the Helmholtz equation, namely

$$\psi_0 = \frac{e^{ikr}}{r} \, . \tag{7.2}$$

This is a spherical wave originating from the point $\mathcal{P}$, and r is the distance from $\mathcal{P}$ to the field point. Next, we apply Gauss's theorem, Eq. (1.120), to the vector field $\psi\boldsymbol{\nabla}\psi_0 - \psi_0\boldsymbol{\nabla}\psi$ and invoke Eq. (7.1), thereby arriving at Green's theorem:

$$\begin{aligned} \int_{\mathcal{S}+S'} (\psi\boldsymbol{\nabla}\psi_0 - \psi_0\boldsymbol{\nabla}\psi)\cdot d\mathbf{S} &= -\int_\mathcal{V} (\psi\nabla^2\psi_0 - \psi_0\nabla^2\psi)dV \\ &= 0 \end{aligned} \tag{7.3}$$

Here we have introduced a small sphere $\mathcal{S}'$ of radius r' surrounding $\mathcal{P}$; $\mathcal{V}$ is the volume between the two surfaces $\mathcal{S}'$ and $\mathcal{S}$; and for future convenience we have made an unconventional choice of direction for the integration element $d\mathbf{S}$: it points into $\mathcal{V}$ instead of outward; see Fig. 7.1. As we let the radius r' decrease to zero, we find that,

$\psi \boldsymbol{\nabla}\psi_0 - \psi_0 \boldsymbol{\nabla}\psi \sim -\psi(0)/r'^2 + O(1/r')$ and so the integral over $\mathcal{S}'$ becomes $4\pi\psi_{\mathcal{P}}$. Rearranging, we obtain

$$\psi_{\mathcal{P}} = \frac{1}{4\pi}\int_{\mathcal{S}} \left(\psi \boldsymbol{\nabla}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\boldsymbol{\nabla}\psi\right) \cdot d\mathbf{S} \; . \tag{7.4}$$

Equation (7.4), known as the *Helmholtz-Kirchhoff formula*, is the promised expression for the field ψ at some point $\mathcal{P}$ in terms of a linear combination of its value and normal derivative on a surrounding surface. The specific combination of ψ and $d\mathbf{S}\cdot\boldsymbol{\nabla}\psi$ that appears in this formula is perfectly immune to contributions from any wave that might originate at $\mathcal{P}$ and pass outward through $\mathcal{S}$ (any "outgoing wave"). The integral thus is influenced only by waves that enter $\mathcal{V}$ through $\mathcal{S}$, propagate through $\mathcal{V}$, and then leave through $\mathcal{S}$. Moreover, if $\mathcal{P}$ is many wavelengths away from the boundary $\mathcal{S}$, then to high accuracy the integral is influenced by the waves ψ only when they are entering through $\mathcal{S}$ (when they are incoming), and not when they are leaving (outgoing). This fact is important for applications, as we shall see.

7.2.1 Diffraction by an Aperture

Next, let us suppose that some aperture $\mathcal{Q}$ of size much larger than a wavelength but much smaller than the distance to $\mathcal{P}$ is illuminated by a distant wave source (Fig. 7.1). (If the aperture were comparable to a wavelength in size, or if part of it were only a few wavelengths from $\mathcal{P}$, then polarizational coupling effects at the aperture would be large; our assumption avoids this complication.) Let the surface $\mathcal{S}$ pass through $\mathcal{Q}$, and denote by ψ' the wave incident on $\mathcal{Q}$. We assume that the diffracting aperture has a local and linear effect on ψ'. More specifically, we suppose that the wave transmitted through the aperture is given by

$$\psi_{\mathcal{Q}} = t\psi' \; , \tag{7.5}$$

where t is a complex transmission function that varies over the aperture. In practice, t is usually zero (completely opaque region) or unity (completely transparent region). However t can also represent a variable phase factor when, for example, the aperture comprises a medium of variable thickness and of different refractive index from that of the homogeneous medium outside the aperture.

What this formalism does not allow, though, is that $\psi_{\mathcal{Q}}$ at any point on the aperture be influenced by the wave's interaction with other parts of the aperture. For this reason, not only the aperture, but any structure that it contains must be many wavelengths across. To give a specific example of what might go wrong, suppose that electromagnetic radiation is normally incident upon a wire grid. Surface currents will be induced in the wires by the wave's electric field, and those currents will produce a secondary wave that cancels the primary wave immediately behind each wire, thereby "eclipsing" the wave. If the secondary wave from the currents flowing in the next wire is comparable with the first wire's secondary wave, then the transmitted net wave field will get modified in a complex, polarization-dependent manner. Such modifications are negligble if the wires are a number of wavelengths apart.

Let us now use the Helmholtz-Kirchoff formula (7.4) to compute the field at $\mathcal{P}$ due to the wave $\psi_{\mathcal{Q}} = t\psi'$ transmitted through the aperture. Let the surface $\mathcal{S}$ of Fig. 7.1 comprise the aperture $\mathcal{Q}$, a sphere of radius $R \gg r$ centered on $\mathcal{P}$, and the linear extension of the aperture to meet the sphere; and assume that the only incoming waves are those which pass through the aperture. Then, as noted above, when the incoming waves subsequently pass on outward through $\mathcal{S}$, they contribute negligibly to the integral (7.4), so the only contribution is from the aperture itself.[1]

On the aperture, because $kr \gg 1$, we can write $\nabla(e^{ikr}/r) \simeq -ik\mathbf{n}e^{ikr}/r$ where $\mathbf{n}$ is a unit vector pointing towards $\mathcal{P}$. Similarly, we write $\boldsymbol{\nabla}\psi \simeq ikt\mathbf{n}'\psi'$, where $\mathbf{n}'$ is a unit vector along the direction of propagation of the incident wave, and where our assumptions have permitted us to ignore the gradient of t. Inserting these gradients into the Helmholtz-Kirchoff formula, we obtain

$$\psi_P = -\frac{ik}{2\pi}\int_{\mathcal{Q}} d\mathbf{S}\cdot\left(\frac{\mathbf{n}+\mathbf{n}'}{2}\right)\frac{e^{ikr}}{r}t\psi'. \tag{7.6}$$

Eq. (7.6) can be used to compute the wave from a small aperture at any point in the far field. It has the form of an integral transform of the incident field variable, ψ', where the integral is over the area of the aperture. The kernel of the transform is the product of several factors. There is a factor $1/r$. This guarantees that the flux falls off as the inverse square of the distance to aperture as we might have expected. There is also a phase factor $-ie^{ikr}$ which advances the phase of the wave by an amount equal to the optical path length between the element of the aperture and $\mathcal{P}$, minus $\pi/2$. The amplitude and phase of the wave can also be changed by the transmission function t. Finally there is the geometric factor $d\hat{\mathbf{S}}\cdot(\mathbf{n}+\mathbf{n}')/2$ (with $d\hat{\mathbf{S}}$ the unit vector normal to the aperture). This is known as the *obliquity factor*, and it ensures that the waves from the aperture propagate only forward with respect to the original wave and not backward (not in the direction $\mathbf{n} = -\mathbf{n}'$). More specifically, this factor prevents the backward propagating secondary wavelets in Huygens construction from reinforcing each other to produce a back-scattered wave. When dealing with paraxial optics (later in this chapter), we can usually set the obliquity factor to unity.

It is instructive to specialize to a point source seen through a small diffracting aperture. If we suppose that the source has unit strength and is located at $\mathcal{P}'$, a distance r' before $\mathcal{Q}$, then $\psi' = -e^{ikr'}/4\pi r'$, and ψ_P can be written in the symmetric form

$$\psi_P = \int\left(\frac{e^{ikr}}{4\pi r}\right) it(\mathbf{k}'+\mathbf{k})\cdot d\mathbf{S}\left(\frac{e^{ikr'}}{4\pi r'}\right). \tag{7.7}$$

We can think of this expression as the Greens function response at $\mathcal{P}$ to a δ-function source at $\mathcal{P}'$. Alternatively, we can regard it as a *propagator* from $\mathcal{P}'$ to $\mathcal{P}$ by way of the aperture.

7.2.2 Spreading of the Wavefront

Eq. (7.6) [or (7.7)] gives a general prescription for computing the diffraction pattern from an illuminated aperture. It is commonly used in two complementary limits, called "Frauhn-

[1]Actually, the incoming waves will diffract around the edge of the aperture onto the back side of the screen that bounds the aperature, i.e. the side facing $\mathcal{P}$; and this diffracted wave will contribute to the Helmholtz-Kirchhoff integral in a polarization-dependent way. However, because the diffracted wave decays along the screen with an e-folding length of order a wavelength, its contribution will be negligible if the aperture is many wavelengths across and $\mathcal{P}$ is many wavelengths away from the edge of the aperture, as we have assumed.

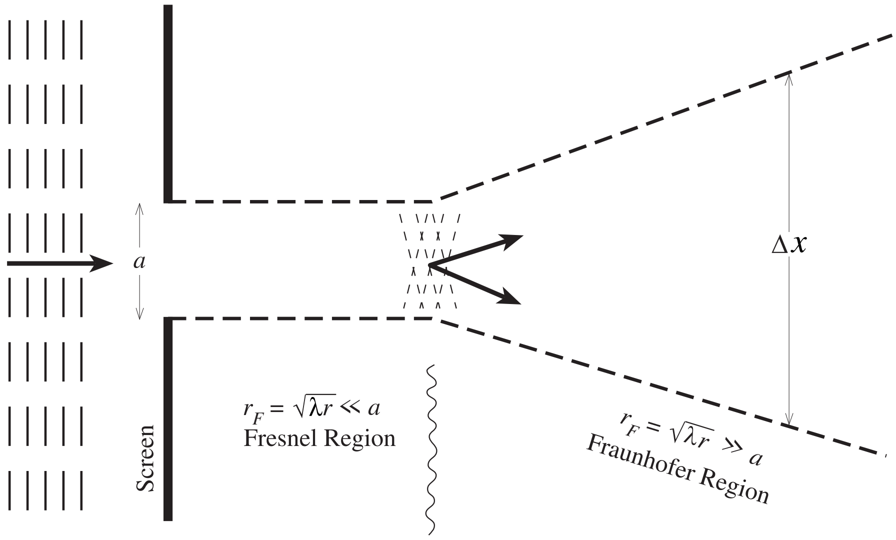

Fig. 7.2: Fraunhofer and Fresnel Diffraction.

hofer" and "Fresnel".

Suppose that the aperture has linear size a and is roughly centered on the geometric ray from the source point $\mathcal{P}'$ to the field point $\mathcal{P}$. Consider the variations of the phase ϕ of the contributions to $\psi_{\mathcal{P}}$ that come from various places in the aperture. Using elementary trigonometry, we can estimate that locations on the aperture's opposite sides have phases that differ by $\Delta\phi \sim ka^2/2r$. There are two limiting regimes depending on whether the aperture is large or small compared with the so-called *Fresnel length*

$$r_F \equiv \left(\frac{2\pi r}{k}\right)^{1/2} = (\lambda r)^{1/2}. \tag{7.8}$$

(Note that the Fresnel length depends on the distance r of the field point from the aperture.) When $a \ll r_F$, the phase variation $\Delta\phi$ can be ignored; the contributions from different parts of the aperture are essentially in phase with each other. This is the *Fraunhofer* regime. When $a \gg r_F$, the phase variation is of upmost importance in determining the observed intensity pattern $|\psi_{\mathcal{P}}|^2$. This is the *Fresnel* regime; see Fig. 7.2.

We can use an argument familiar, perhaps, from quantum mechanics to deduce the qualitative form of the intensity patterns in these two regimes. For simplicity, let the incoming wave be planar (r' huge) and let it propagate perpendicular to the aperture as shown in Fig. 7.2. Then geometric optics (photons treated like classical particles) would predict that an opaque screen will cast a sharp shadow; the wave leaves the aperture plane as a beam with a sharp edge. However, wave optics insists that the transverse localization of the wave into a region of size $\Delta x \sim a$ must produce a spread in its transverse wave vector, $\Delta k_x \sim 1/a$ (a momentum uncertainty $\Delta p_x = \hbar\Delta k_x \sim \hbar/a$ in the language of the Heisenberg uncertainty principle). This uncertain transverse wave vector produces, after propagating a distance r, a corresponding uncertainty $(\Delta k_x/k)r \sim r_F^2/a$ in the beam's transverse size; and this uncertainty superposes incoherently on the aperture-induced size a of the beam to produce a net

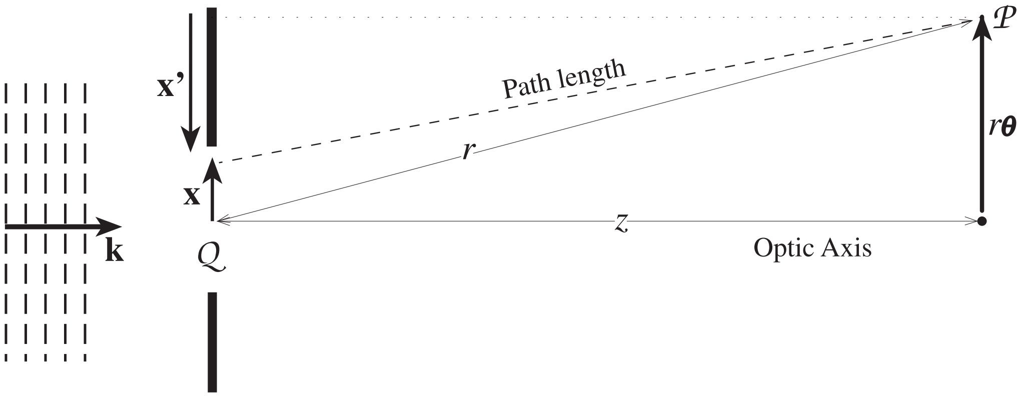

Fig. 7.3: Geometry for computing the path length between a point $\mathcal{Q}$ in the aperture and the point of observation $\mathcal{P}$. The transverse vector $\mathbf{x}$ is used to identify $\mathcal{Q}$ in our Fraunhofer analysis (Sec. 7.3), and $\mathbf{x}'$ is used in our Fresnel analysis (Sec. 7.4).

transverse beam size

$$\begin{aligned}\Delta x &\sim \sqrt{a^2 + (r_F^2/a)^2} \\ &\sim a \quad \text{if} \quad r \ll a^2/\lambda \text{ (Fresnel regime)} \\ &\sim \lambda r/a \quad \text{if} \quad r \gg a^2/\lambda \text{ (Fraunhofer regime).}\end{aligned} \tag{7.9}$$

In the nearby, Fresnel regime, the aperture creates a beam whose edges will have the same shape and size as the aperture itself, and will be reasonably sharp (but with some oscillatory blurring, associated with the wave-packet spreading, that we shall analyze below). By contrast, in the more distant Fraunhofer regime, wave-front spreading will cause the transverse size of the entire beam to grow linearly with distance; and, as we shall see, the intensity pattern typically will not resemble the aperture at all.

7.3 Fraunhofer Diffraction

Consider the Fraunhofer regime of strong wavefront spreading, $a \ll r_F$, and for simplicity specialize to the case of an incident plane wave with wave vector $\mathbf{k}$ orthogonal to the aperture plane; see Fig. 7.3. Regard the line along $\mathbf{k}$ through the center of the aperture $\mathcal{Q}$ as the "optic axis;" identify points in the aperture by their two-dimensional vectorial separation $\mathbf{x}$ from that axis; identify $\mathcal{P}$ by its distance r from the aperture center and its 2-dimensional transverse separation $r\boldsymbol{\theta}$ from the optic axis; and restrict attention to small-angle diffraction $|\boldsymbol{\theta}| \ll 1$. Then the geometric path length between $\mathcal{P}$ and a point on $\mathcal{Q}$ [denoted r in Eq. (7.6)—note our change of the meaning of r] can be expanded as

$$\text{Path length } = (r^2 - 2r\mathbf{x}\cdot\boldsymbol{\theta} + x^2)^{1/2} \sim r - \mathbf{x}\cdot\boldsymbol{\theta} + \frac{x^2}{2r} + \ldots \tag{7.10}$$

cf. Fig. 7.3. The first term in this expression, r, just contributes an $\mathbf{x}$-independent phase e^{ikr} to the $\psi_{\mathcal{P}}$ of Eq. (7.6). The third term, $x^2/2r$, contributes a phase variation that is

$\ll 1$ here in the Fraunhofer region (but that will be important in the Fresnel region, Sec. 7.4 below). Therefore, in the Fraunhofer region we can retain just the second term, $-\mathbf{x}\cdot\boldsymbol{\theta}$ and write Eq. (7.6) in the form

$$\psi_P(\boldsymbol{\theta}) \propto \int e^{-ik\mathbf{x}\cdot\boldsymbol{\theta}} t(\mathbf{x}) d^2x = \tilde{t}(\boldsymbol{\theta}) \ , \tag{7.11}$$

where d^2x is the surface area element in the aperture plane and we have dropped a constant phase factor and constant multiplicative factors. Thus, $\psi_{\mathcal{P}}(\boldsymbol{\theta})$ in the Fraunhofer regime is given by the two-dimensional Fourier transform, denoted $\tilde{t}(\boldsymbol{\theta})$, of the transmission function $t(\mathbf{x})$, with $\mathbf{x}$ made dimensionless in the transform by multiplying by $k = 2\pi/\lambda$.

It is usually uninteresting to normalise Fraunhofer diffraction patterns. Moreover, on those occasions when the absolute value of the observed flux is needed, rather than just the angular shape of the diffraction pattern, it typically can be derived most easily from conservation of the total wave energy. This is why we ignore the proportionality factor in Eq. (7.11).

All of the techniques for handling Fourier transforms that should be familiar from quantum mechanics and elsewhere can be applied to derive Fraunhofer diffraction patterns. In particular, the convolution theorem turns out to be very useful. Let us give an example.

7.3.1 Diffraction Grating

A diffraction grating can be modeled as a finite series of alternating transparent and opaque, long, parallel stripes. Let there be N transparent and opaque stripes each of width $a \gg \lambda$ (Fig. 7.4 a), and idealize them as infinitely long so their diffraction pattern is one-dimensional. We shall outline how to use the convolution theorem to derive their Fraunhofer diffraction pattern. The details are left as an exercise for the reader (Ex. 7.2).

First consider a single transparent stripe (slit) of width a centered on $x = 0$, and measure the scalar angle θ from the direction of the incident radiation. This single stripe has the transmission function

$$\begin{aligned} t_1(x) &= 1 \quad |x| < a/2 \\ &= 0 \quad |x| > a/2 \ , \end{aligned} \tag{7.12}$$

so its diffraction pattern is

$$\begin{aligned} \psi_P(\boldsymbol{\theta}) &\propto \tilde{t}_1 \\ &\propto \int_{-a/2}^{a/2} e^{ikx\theta} dx \\ &\propto \operatorname{sinc}\left(\frac{ka\theta}{2}\right) \ , \end{aligned} \tag{7.13}$$

where $\operatorname{sinc}(x) \equiv \sin(x)/x$.

Now the idealized N-slit grating can be considered as an infinite series of δ-functions with separation $2a$ convolved with the transmission function for a single slit,

$$\int_{-\infty}^{\infty} \left[\sum_{n=-\infty}^{+\infty} \delta(y - 2an) \right] t_1(x-y) dy \tag{7.14}$$

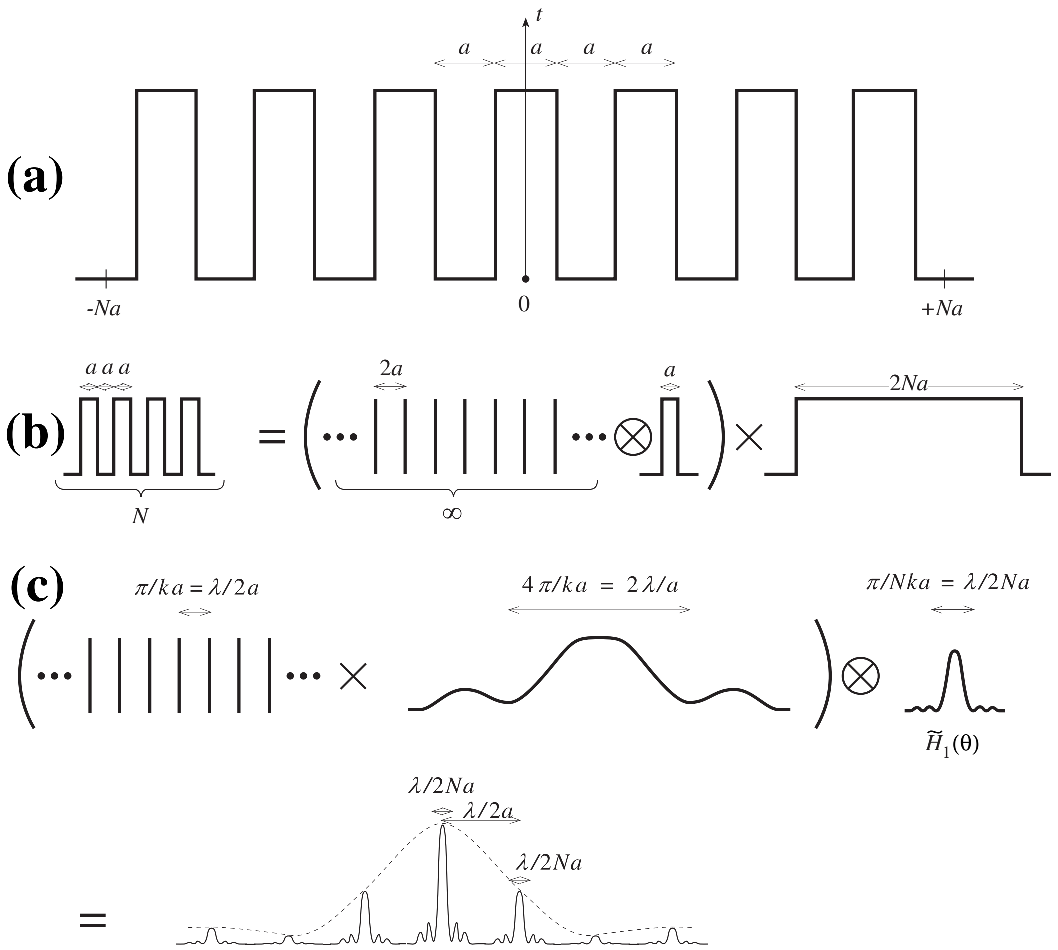

Fig. 7.4: a) Diffraction grating formed by N alternating transparent and opaque stripes each of width a. b) Decomposition of a finite grating into an infinite series of equally spaced δ-functions that are convolved (the symbol $\otimes$) with the shape of an individual transparent stripe and then multiplied (the symbol $\times$) by a large aperture function covering N such stripes; cf. Eq. (7.16) c) The resulting Fraunhofer diffraction pattern shown schematically as the Fourier transform of a series of delta functions multiplied by the Fourier transform of the large aperture and then convolved with the transform of a single stripe.

and then multiplied by the aperture function

$$\begin{aligned} H(x) &= 1 \quad |x| < Na \\ &= 0 \quad |x| > Na; \end{aligned} \tag{7.15}$$

more explicitly,

$$t(x) = \left(\int_{-\infty}^{\infty} \left[\sum_{n=-\infty}^{+\infty} \delta(y - 2an) \right] t_1(x - y) dy \right) H(x) \ , \tag{7.16}$$

which is shown graphically in Fig. 7.4 b.

The convolution theorem says that the Fourier transform of a convolution of two functions is the product of the functions' Fourier transforms, and conversely. Let us apply this theorem to expression (7.16) for our transmission grating. The diffraction pattern of the infinite series of δ-functions with spacing $2a$ is itself an infinite series of δ-functions with reciprocal spacing $2\pi/(2ka) = \lambda/2a$ (see the hint in Exercise 7.2). This must be multiplied by the Fourier transform $\tilde{t}_1(\theta)$ of the single slit, and then convolved with the Fourier transform of $H(x)$, $\tilde{H}(\theta) \propto \mathrm{sinc}(Nka\theta)$. The result is shown schematically in Fig. 7.4 c. (Each of the transforms is real, so the one-dimensional functions shown in the figure fully embody them.)

The diffracted energy flux is $|\psi_{\mathcal{P}}|^2$, where ψ_P is shown at the bottom of Fig. 7.4. What the grating has done is channel the incident radiation into a few equally spaced beams with directions $\theta = \pi p/ka$, where p is an integer known as the *order* of the beam. Each of these beams has a shape given by $|\tilde{H}(\theta)|^2$: a sharp central peak with half width (distance from center of peak to first null of the intensity) $\lambda/2Na$, followed by a set of *side lobes* whose intensities are $\propto N^{-1}$.

The fact that the deflection angles $\theta = \pi p/ka$ of these beams are proportional to $k^{-1} = \lambda/2\pi$ underlies the use of diffraction gratings for spectroscopy. It is of interest to ask what the wavelength resolution of such an idealized grating might be. If one focuses attention on the p'th order beams at two wavelengths λ and $\delta\lambda$ (which are located at $\theta = p\lambda/2a$ and $p(\lambda + \delta\lambda)/2a$, then one can distinguish the beams from each other when their separation $\delta\theta = p\delta\lambda/2a$ is at least as large as the angular distance $\lambda/2Na$ between the maximum of each beam's diffraction pattern and its first minimum, i.e., when

$$\frac{\lambda}{\delta\lambda} \lesssim \mathcal{R} \equiv Np \ . \tag{7.17}$$

$\mathcal{R}$ is called the grating's *chromatic resolving power.*

Real gratings are not this simple. First they usually work not by modulating the amplitude of the incident radiation in this simple manner, but instead by modulating the phase. Second, the manner in which the phase is modulated is such as to channel most of the incident power into a particular order, a technique known as *blazing.* Third, gratings are often used in reflection rather than transmission. Despite these complications, the principles of a real grating's operation are essentially the same as our idealized grating. Manufactured gratings typically have $N \gtrsim 10,000$, giving a wavelength resolution for visual light that can be as small as $\sim 10\,\mathrm{pm}$, i.e. $10^{-11}\mathrm{m}$.

Fig. 7.5: Two complementary apertures used to illustrate Babinet's Principle.

7.3.2 Babinet's Principle

We have shown how to compute the Fraunhofer diffraction pattern formed by, for example, a narrow slit. We might also be interested in the pattern from a complementary aperture, a needle of width and length the same as those for the slit. We can derive the needle's pattern by observing that the sum of the waves from the two apertures should equal the wave from a completely unaltered incident wave front. That is to say if we exclude the direction of the incident wave, the field amplitudes diffracted by the two apertures are the negative of each other, and hence the intensities $|\psi|^2$ are the same. Therefore, the Fraunhofer diffraction patterns from the needle and the slit—and indeed from any pair of complementary apertures, e.g., Fig. 7.5—are identical, except in the direction of the incident wave (the "precisely forward" direction).

7.3.3 Hubble Space Telescope

The Hubble Space Telescope was launched in April 1990 to observe planets, stars and galaxies above the earth's atmosphere. One reason for going into space is to avoid the irregular refractive index variations in the earth's atmosphere, known generically as *seeing*, which degrade the quality of the images. (Another reason is to observe the ultraviolet part of the spectrum, which is absorbed in the earth's atmosphere.) Seeing typically limits the angular resolution of Earth-bound telescopes at visual wavelengths to $\sim 1''$. We wish to compute how much the angular resolution improves by going into space. As we shall see, the computation is essentially an exercise in Fraunhofer diffraction theory.

The essence of the computation is to idealise the telescope as a circular aperture with diameter equal to the diameter of the primary mirror. Light from this mirror is actually reflected onto a secondary mirror and then follows a complex optical path before being focused onto a variety of detectors. However, this path is irrelevant to the angular resolution. The purpose of the optics is merely to bring the light to a focus close to the mirror, in order to produce an instrument that is compact enough to be launched and to match the sizes of stars' images to the pixel size on the detector. In doing so, however, the optics leaves the angular resolution unchanged; the resolution is the same as if we were simply to observe the light, which passes through the primary mirror's circular aperture, far beyond the mirror, in the Fraunhofer region.

If the telescope aperture were very small, for example a pin hole, then the light from

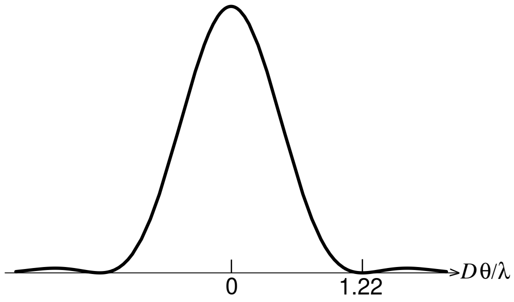

Fig. 7.6: Airy diffraction pattern produced by a circular aperture.

a point source (a very distant star) would create a broad diffraction pattern, and the telescope's angular resolution would be correspondingly poor. As we increase the diameter of the aperture, we still see a diffraction pattern, but its width diminishes.

Using these considerations, we can compute how well the telescope can distinguish neighboring stars. We do not expect it to fully resolve them if they are closer together on the sky than the angular width of the diffraction patttern. Of course, optical imperfections in a real telescope may degrade the image quality even further, but this is the best that we can do, limited only by the uncertainty principle.

The calculation of the Fraunhofer amplitude far from the aperture is straightforward:

$$\begin{aligned} \psi(\theta) &\propto \int e^{-ik\mathbf{x}\cdot\boldsymbol{\theta}} d^2x \\ &\propto \text{jinc}\left(\frac{kD\theta}{2}\right) \end{aligned} \tag{7.18}$$

where D is the diameter of the aperture (i.e., of the telescope's primary mirror) and $\text{jinc}(x) \equiv J_1(x)/x$ with J_1 the Bessel function of order one. The flux from the star observed at angle θ is therefore $\propto \text{jinc}^2(kD\theta/2)$. This intensity pattern, known as the *Airy pattern*, is shown in Fig. 7.6. There is a central "Airy disk" surrounded by a circle where the flux vanishes, and then further surrounded by a series of concentric rings whose flux diminishes with radius. Only 16 percent of the total light falls outside the central Airy disk. The angular radius θ_A of the Airy disk, i.e. the radius of the dark circle surrounding it, is determined by the first zero of $J_1(kD\theta/2)$: $\theta_A = 1.22\lambda/D$.

A conventional, though essentially arbitrary, criterion for angular resolution is to say that two point sources can be distinguished if they are separated in angle by more than θ_A. For the Hubble Space Telescope, $D = 2.4$m and $\theta_A \sim 0.04''$ at visual wavelengths, which is over ten times better than is achievable on the ground with conventional (non-adaptive) optics.

Initially, there was a serious problem with Hubble's telescope optics. The hyperboloidal primary mirror was ground to the wrong shape, so rays parallel to the optic axis did not pass through a common focus after reflection off a convex hyperboloidal secondary mirror. This defect, known as *spherical aberration*, created blurred images. Correcting optics were developed, and were installed by astronauts several years after launch.

****

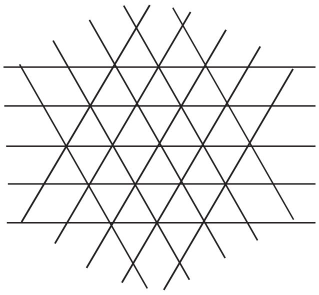

Fig. 7.7: Diffraction grating formed from three groups of parallel lines.

EXERCISES

Exercise 7.1 *Problem: Pointillist Painting*

The neo-impressionist painter George Seurat was a member of the pointillist school. His paintings consisted of an enormous number of closely spaced dots (of size $\sim$ 0.4mm) of pure pigment. The illusion of color mixing was produced only in the eye of the observer. How far from the painting should one stand in order to obtain the desired blending of color?

Exercise 7.2 *Problem: Thickness of a Human Hair*

Conceive and carry out an experiment using light diffraction to measure the thickness of a hair from your head, accurate to within a factor ~ 2. [Hint: make sure the source of light that you use is small enough that its finite size has negligible influence on your result.]

Exercise 7.3 *Derivation: Diffraction Grating*

Use the convolution theorem to carry out the calculation of the Fraunhofer diffraction pattern from the grating shown in Fig. 7.4. [*Hint:* To show that the Fourier transform of the infinite sequence of equally spaced delta functions is a similar sequence of delta functions, perform the Fourier transform to get $\sum_{n=\infty}^{+\infty} e^{i2kan\theta}$ (aside from a multiplicative factor); then use the formulas for a Fourier *series* expansion, and its inverse, for any function that is periodic with period π/ka to show that $\sum_{n=-\infty}^{+\infty} e^{i2kan\theta}$ is a sequence of delta functions.]

Exercise 7.4 *Problem: Triangular Diffraction Grating*

Sketch the Fraunhofer diffraction pattern you would expect to see from a diffraction grating made from three groups of parallel lines aligned at angles of 120° to each other (Fig. 7.7).

Exercise 7.5 *Problem: Light Scattering by Particles*

Consider the scattering of light by an opaque particle of size $a \gg 1/k$. One component of the scattered radiation is due to diffraction around the particle. This component is confined to a cone with opening angle $\Delta\theta \sim 1/ka \ll 1$ about the incident wave direction. It contains power $P_S = FA$, where F is the incident energy flux and A is the cross sectional area of the particle perpendicular to the incident wave.

(a) Give a semi-quantitative derivation of $\Delta\theta$ and P_S using Babinet's principle.

(b) Explain why the total "extinction" (absorption plus scattering) cross section is equal to $2A$ independent of the shape of the opaque particle.

****

7.4 Fresnel Diffraction

We now turn to the Fresnel regime, where the aperture is far larger than the Fresnel length r_F and there is a large phase variation over the aperture. We specialize to incoming wave vectors that are approximately orthogonal to the aperture plane and to small diffraction angles so that we can ignore the obliquity factor. By contrast with the Frauhofer case, however, we identify $\mathcal{P}$ by its distance z from the aperture plane instead of its distance r from the aperture center, and we use as our integration variable in the aperture $\mathbf{x}' \equiv \mathbf{x} - r\boldsymbol{\theta}$ (cf. Fig. 7.3.), thereby writing the dependence of the phase at $\mathcal{P}$ on $\mathbf{x}$ in the form

$$\Delta\phi \equiv k \times [(\text{path length from } \mathbf{x} \text{ to } \mathcal{P}) - z] = \frac{k\mathbf{x}'^2}{2z} + O\left(\frac{kx'^4}{z^3}\right). \tag{7.19}$$

In the Fraunhofer region (Sec. 7.3 above), only the linear term $-k\mathbf{x}\cdot\boldsymbol{\theta}$ in $k\mathbf{x}'^2/2z \simeq k(\mathbf{x} - r\boldsymbol{\theta})^2/r$ was significant. In the Fresnel region the term quadratic in $\mathbf{x}$ is also significant (and we have changed variables to $\mathbf{x}'$ so as to simplify it), but the $O(x'^4)$ term is negligible.

Let us consider the Fresnel diffraction pattern formed by a simple aperture of arbitrary shape, illuminated by a normally incident plane wave. It is convenient to introduce Cartesian coordinates (x', y') and to define

$$s = \left(\frac{k}{\pi z}\right)^{1/2} x', \quad t = \left(\frac{k}{\pi z}\right)^{1/2} y'. \tag{7.20}$$

We can thereby rewrite Eq. (7.6) (setting the obliquity factor to one) in the form

$$\psi_{\mathcal{P}} = -\frac{ike^{ikz}}{2\pi z}\int_{\mathcal{Q}} e^{i\Delta\phi}\psi_{\mathcal{Q}}dx'dy' = -\frac{i}{2}\int\int e^{i\pi s^2/2}e^{i\pi t^2/2}\psi_{\mathcal{Q}}e^{ikz}dsdt\,. \tag{7.21}$$

We shall use this rather general expression in the next section, when discussing Fourier optics. In this section we shall focus on the details of the Fresnel diffraction pattern for an

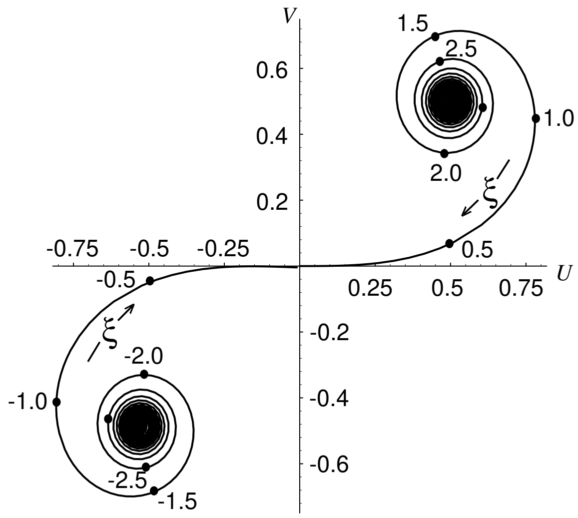

Fig. 7.8: Cornu Spiral.

incoming plane wave that falls perpendicularly on the aperture, so $\psi_{\mathcal{Q}}$ is constant over the aperture. For simplicity, we initially confine attention to a rectangular aperture with edges along the x' and y' directions. Then the two integrals have limits that are independent of each other and the integrals can be expressed in the form $S(s_{max})-S(s_{min})$ and $S(t_{max})-S(t_{min})$, so

$$\psi_{\mathcal{P}} = \frac{-i}{2}[S(s_{max}) - S(s_{min})][S(t_{max}) - S(t_{min})]\psi_{\mathcal{Q}}e^{ikz} \equiv \frac{-i}{2}\Delta S_s \Delta S_t \psi_{\mathcal{Q}} e^{ikz} \ , \tag{7.22}$$

where the arguments are the limits of integration and where

$$S(\xi) \equiv \int_0^{\xi} e^{i\pi s^2/2} ds \equiv U(\xi) + iV(\xi) \tag{7.23}$$

with

$$U(\xi) \ = \ \int_0^{\xi} ds \cos(\pi s^2/2) \ , \tag{7.24}$$

$$V(\xi) \ = \ \int_0^{\xi} ds \sin(\pi s^2/2) \ . \tag{7.25}$$

The real functions $U(\xi), V(\xi)$ are known as Fresnel integrals.

It is convenient to exhibit the Fresnel integrals graphically using a *Cornu Spiral* (Fig. 7.8). This is a graph of the parametric equation $[U(\xi), V(\xi)]$, or equivalently a graph of $S(\xi) = U(\xi) + iV(\xi)$ in the complex plane. The two terms in Eq. (7.23) can be represented in amplitude and phase by arrows in the (U,V) plane reaching from $\xi = s_{min}$ on the Cornu spiral to $\xi = s_{max}$, and from $\xi = t_{min}$ to $\xi = t_{\max}$.

The simplest illustration is the totally unobscured, plane wavefront. In this case, the limits of both integrations extend from $-\infty$ to $+\infty$, which as we see in Fig. 7.8 is an arrow of length $2^{1/2}$ and phase $\pi/4$. Therefore, ψ_P is equal to $(2^{1/2}e^{i\pi/4})^2(-i/2)\psi_Q e^{ikz} = \psi_Q e^{ikz}$, as we could have deduced simply by solving the Helmholtz equation (7.1) for a plane wave.

This unobscured-wavefront calculation vindicates three procedures that we have already used. First, it illustrates our interpretation of Fermat's principle in geometric optics. In the limit of short wavelength, the paths that contribute to the wave field are just those along which the phase is stationary to small variations in path. Our present calculation shows that, because of the tightening of the Cornu spiral as one moves toward a large argument, the contributing paths are those that are separated from the geometric-optics one by less than a few Fresnel lengths at $\mathcal{Q}$. (For a laboratory experiment with light and $z \sim 2$m, a Fresnel length is typically ~ 1mm.)

A second, and related, point is that in computing the diffraction pattern from a more complicated aperture, we need only perform the integral (7.6) in the immediate vicinity of the geometric-optics ray. We can ignore the contribution from the extension of the aperture $\mathcal{Q}$ to meet the "sphere at infinity" even when the wave is unobstructed there. The rapid phase variation makes this contribution sum to zero.

Third, in integrating over the whole area of the wave front at $\mathcal{Q}$, we have summed contributions with increasingly large phase differences that add in such a way that the total has a net extra phase of $\pi/2$, relative to the geometric-optics ray. This phase factor cancels exactly the prefactor $-i$ in the Fresnel-Kirchhoff integral, Eq. (7.6). (This phase factor is unimportant in the limit of geometric optics.)

7.4.1 Lunar Occultation of a Radio Source

The next simplest case of Fresnel diffraciton is the pattern formed by a straight edge. To give a specific example consider a cosmologically distant source of radio waves that is occulted by the moon. If we treat the lunar limb as a straight edge, then as it passes in front of the radio source, a changing diffraction pattern will be sampled by a telescope on earth. We orient our coordinates so the moon's edge is along the y' direction (t direction). Then in Eq. (7.22) $\Delta S_t \equiv S(t_{\rm max}) - S(t_{\rm min}) = \sqrt{2i}$ is constant, and $\Delta S_s \equiv S(s_{\rm max}) - S(s_{\rm min})$ is described by the Cornu spiral. Long before the occultation, ΔS_s will be given by the arrow from $(-1/2, -1/2)$ to $(1/2, 1/2)$, i.e. $\Delta S_s = \sqrt{2i}$. The observed wave amplitude, Eq. (7.22), is therefore $\psi_Q e^{ikz}$.

Now when the moon starts to occult the radio source, the upper bound on the Fresnel integral must diminish from $s_{max} = +\infty$, and the complex vector on the Cornu spiral begins to oscillate in length (e.g., from A to B in Fig. 7.9) and in phase. The observed flux will also oscillate, more and more strongly as geometric occultation is approached. At the point of geometric occultation, the complex vector extends from $(-1/2, -1/2)$ to $(0, 0)$ and so the observed wave intensity is one quarter the unocculted value. As the occultation proceeds, the length of the complex vector and the observed flux will decrease monotonically to zero, while the phase continues to oscillate.

Historically, diffraction of a radio-source's waves by the moon led to the discovery of quasars—the hyperactive nuclei of distant galaxies. In the early 1960s, a team of British radio observers led by Cyril Hazard knew that the moon would occult a powerful radio source named 3C273, so they set up their telescope to observe the development of the diffraction

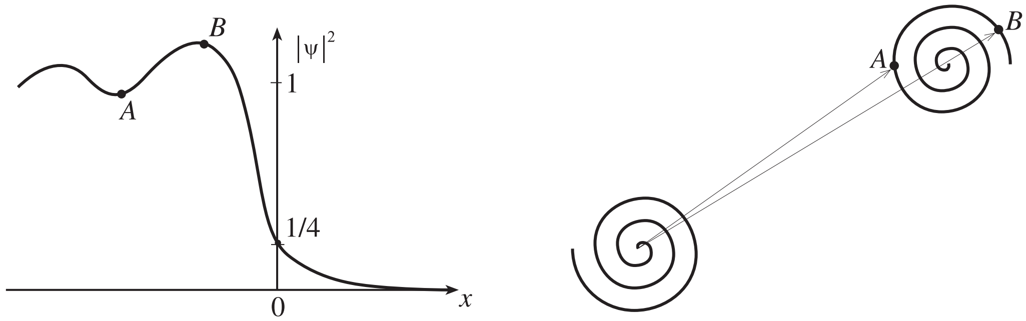

Fig. 7.9: Diffraction pattern formed by a straight edge and graphical interpretation using Cornu Spiral.

pattern as the occultation proceded. From the pattern's observed times of ingress (passage into the moon's shadow) and egress (emergence from the moon's shadow), Hazard determined the coordinates of 3C273 on the sky. These coordinates enabled Maarten Schmidt at the 200-inch telescope on Palomar Mountain to identify 3C273 optically and discover (from its optical redshift) that it was surprisingly distant and consequently had an unprecedented luminosity.

In Hazard's occultation measurements, the observing wavelength was $\lambda \sim 0.2$m. Since the moon is roughly $z \sim 400,000$ km distant, the Fresnel length was about $r_F = \sqrt{\lambda z} \sim 10$km. The orbital speed of the moon is $u \sim 200$m s^{-1}, so the diffraction pattern took a time $\sim 5r_F/u \sim 4$min to pass through the telescope.

The straight-edge diffraction pattern of Fig. 7.9 occurs universally along the edge of the shadow of any object, so long as the source of light is sufficiently small and the shadow's edge bends on lengthscales long compared to the Fresnel length $r_F = \sqrt{\lambda z}$.

7.4.2 Circular Apertures

We have shown how the diffraction pattern for a plane wave can be thought of as formed by waves that derive from a patch a few Fresnel lengths in size. This notion can be made quantitatively useful if we re-analyze the unobstructed wave front in circular polar coordinates. Consider a plane wave incident on an aperture $\mathcal{Q}$ and define $\rho \equiv |\mathbf{x}'|/r_F = \sqrt{\frac{1}{2}(s^2+t^2)}$. Then the phase factor in Eq. (7.21) is $\Delta\phi = \pi\rho^2$ and the observed wave will thus be given by

$$\begin{aligned}\psi_{\mathcal{P}} &= -i\int_0^\rho 2\pi\rho d\rho e^{i\pi\rho^2}\psi_{\mathcal{Q}}e^{ikz} \\ &= (1-e^{i\pi\rho^2})\psi_{\mathcal{Q}}e^{ikz} \ . \end{aligned} \tag{7.26}$$

Now, this integral does not appear to converge as $\rho \to \infty$. We can see what is happening if we sketch an amplitude-and-phase diagram (Fig. 7.10). Adding up the contributions to $\psi_{\mathcal{P}}$ from each annular ring, we see that as we integrate outward from $\rho = 0$, the complex vector has the initial phase retardation of $\pi/2$ but then moves on a semi-circle so that by the time we have integrated out to a radius of r_F, i.e. $\rho = 1$, the contribution to the observed

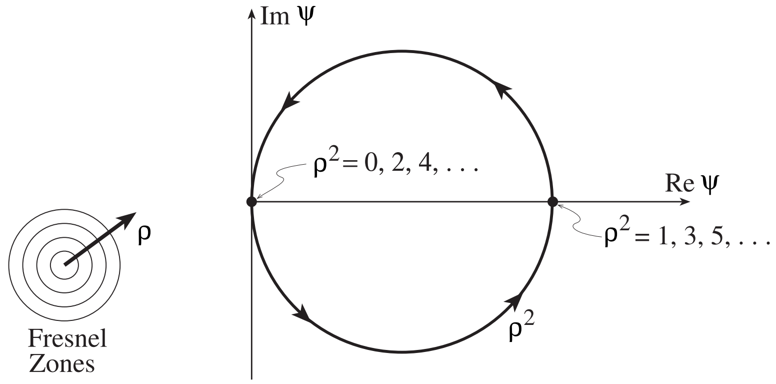

Fig. 7.10: Amplitude-and-phase diagram for an unobstructed plane wave front, decomposed into Fresnel zones.

wave is $\psi_{\mathcal{P}} = 2\psi_{\mathcal{Q}}$ in phase with the incident wave. Then, when the integration has been extended onward to $\sqrt{2}\, r_F$, $\rho = \sqrt{2}$, the circle has been completed and $\psi_{\mathcal{P}} = 0$! The integral continues on around the same circle as the upper-bound radius is further increased.

Of course, the field must actually have a well-defined value, despite this apparent failure of the integral to converge. To understand how the field becomes well-defined, imagine splitting the aperture $\mathcal{Q}$ up into concentric annular rings, known as *Fresnel half-period zones*, of radius $\sqrt{n}\, r_F$, where $n = 1, 2, 3 \ldots$. The integral fails to converge because the contribution from each odd-numbered ring cancels that from an adjacent even-numbered ring. However, the thickness of these rings decreases as $1/\sqrt{n}$, and eventually we must allow for the fact that any physical source of waves will have a finite angular size. The finite size causes different pieces of the source to have their Fresnel rings centered at slightly different points in the aperture plane, and this causes our computation of $\psi_{\mathcal{P}}$ to begin averaging over rings, and the averaging forces the tip of the complex vector to asymptote to the center of the circle in Fig. 7.10. Correspondingly, due to the averaging, the observed intensity asymptotes to $|\psi_{\mathcal{Q}}|^2$.

Although this may not have seemed a particularly wise way to decompose a plane wave front, it does allow a particularly striking experimental verification of our theory of diffraction. Suppose that we fabricate an aperture (called a *zone plate*) in which, for a chosen observation point $\mathcal{P}$ on the optic axis, alternate half-period zones are obscured. Then the wave observed at $\mathcal{P}$ will be the linear sum of several diameters of the circle in Fig. 7.10, and therefore will be far larger than $\psi_{\mathcal{Q}}$. This strong amplification is confined to our chosen spot on the optic axis; most everywhere else the field's intensity is reduced, thereby conserving energy. Thus, the zone plate behaves like a lens (a "Fresnel lens"). The lens's focal length is $f = kA/2\pi^2$, where A (typically chosen to be a few mm^2 for light) is the area of the first half-period zone.

Zone plates are only good lenses when the radiation is monochromatic, since the focal length is wavelength-dependent, $f \propto \lambda^{-1}$. They have the further interesting property that they possess secondary foci, where the fields from $3, 5, 7, \ldots$ contiguous zones add up coherently (Ex. 7.5).

****

EXERCISES

Exercise 7.6 *Problem: Zone Plate*

(a) Use an amplitude-and-phase diagram to explain why a zone plate has secondary foci at distances of $f/3, f/5, f/7 \ldots$.

(b) An opaque, perfectly circular disk of diameter D is placed perpendicular to an incoming plane wave. Show that, at distances r such that $r_F \ll D$, the disk casts a rather sharp shadow, but at the precise center of the shadow there should be a bright spot. How bright?

Exercise 7.7 *Problem: Seeing in the atmosphere.*

Stars viewed through the atmosphere appear to have angular diameters of order an arc second and to exhibit large amplitude fluctuations of flux with characteristic frequencies that can be as high as 100Hz. Both of these phenomena are a consequence of irregular variations in the refractive index of the atmosphere. An elementary model of this effect consists of a thin phase-changing screen, about a km above the ground, on which the rms phase variation is $\Delta\phi \gtrsim 1$ and the characteristic spatial scale, on which the phase changes by $\sim \Delta\phi$, is a.

(a) Explain why the rays will be irregularly deflected through a scattering angle $\Delta\theta \sim (\lambda/a)\Delta\phi$. Strong intensity variation requires that several rays deriving from points on the screen separated by more than a, combine at each point on the ground. These rays combine to create a diffraction pattern on the ground with scale b.

(b) Show that the Fresnel length in the screen is $\sim \sqrt{ab}$. Now the time variation arises because winds in the upper atmosphere with speeds $u \sim 30\text{m s}^{-1}$ blow the irregularities and the diffraction pattern past the observer. Use this information to estimate the Fresnel length, r_F, the atmospheric fluctuation scale size a, and the rms phase variation $\Delta\phi$. Do you think the assumptions of this model are well satisfied?

Exercise 7.8 *Problem: Spy Satellites*

Telescopes can also look down through the same atmospheric irregularities as those discussed in the previous example. In what important respects will the optics differ from that for telescopes looking upward?

****

7.5 Fourier Optics

We have developed a linear theory of wave optics which has allowed us to calculate diffraction patterns in the Fraunhofer and Fresnel limiting regimes. That these calculations agree with laboratory measurements provides some vindication of the theory and the assumptions implicit in it. We now turn to practical applications of these ideas, specifically to the acquisition and processing of images by instruments operating throughout the electromagnetic spectrum. As we shall see, these instruments rely on an extension of paraxial geometric optics (Sec. 6.4) to situations where diffraction effects are important. Because of the central role played by Fourier transforms in diffraction [e.g. Eq. (7.11)], the theory underlying these instruments is called paraxial Fourier optics, or just Fourier optics.

Although the conceptual framework and mathematical machinery for image processing by Fourier optics were developed over a century ago, Fourier optics has only been widely exploited during the past thirty years. This maturation has been driven in part by a growing recognition of similarities between optics and communication theory — for example, the realization that a microscope is simply an image processing device. The development of electronic computation has also triggered enormous strides; computers are now seen as extensions of optical devices, and visa versa. It is a matter of convenience, economics and practicality to decide which parts of the image processing are carried out with mirrors, lenses, etc., and which parts are performed numerically.

One conceptually simple example of optical image processing would be an improvement in one's ability to identify a faint star in the Fraunhofer diffraction rings ("fringes") of a much brighter star. As we shall see below [Eq. (7.32) and subsequent discussion], the bright star's image in a telescope's focal plane has the same Airy diffraction pattern as we met in Eq. (7.18) and Fig. 7.6. If the shape of that image could be changed from the ring-endowed Airy pattern to a Gaussian, then it would be far easier to identify the nearby faint star. One way to achieve this would be to attenuate the incident radiation at the telescope aperture in such a way that, immediately after passing through the aperture, it has a Gaussian profile instead of a sharp-edged profile. Its Fourier transform (the diffraction pattern in the focal plane) would then also be a Gaussian. Such a Gaussian-shaped attenuation is difficult to achieve in practice, but it turns out—as we shall see–that there are easier options.

Before exploring these options, we must lay some foundations, beginning with the concept of coherent illumination in Sec. 7.5.1, and then point spread functions in Sec. 7.5.2.

7.5.1 Coherent Illumination

If the radiation that arrives at the input of an optical system derives from a single source, e.g. a point source that has been collimated into a parallel beam by a converging lens, then the radiation is best described by its complex amplitude ψ (as we are doing in this chapter). An example might be a biological specimen on a microscope slide, illuminated by an external point source, for which the phases of the waves leaving different parts of the slide are strongly correlated with each other. This is called *coherent illumination*. If, by contrast, the source is self luminous and of non-negligible size, with the atoms or molecules in its different parts radiating independently—for example a cluster of stars—then the phases of the radiation from different parts are uncorrelated, and it may be the intensity of the radiation, not the complex amplitude, that obeys well-defined (non-probabilistic) evolution laws. This is called *incoherent illumination*. In this chapter we shall develop Fourier optics for a coherently illuminating source (the kind of illumination tacitly assumed in previous sections of the chapter) . A parallel theory with a similar vocabulary can be developed for incoherent sources, and some of the foundations for it will be laid in Chap. 8. In Chap. 8 we shall also develop a more precise formulation of the concept of coherence.

7.5.2 Point Spread Functions

In our treatment of paraxial geometric optics (Sec. 6.4), we showed how it is possible to regard a group of optical elements as a sequence of linear devices and relate the output rays to the input by linear operators, i.e. matrices. This chapter's theory of diffraction is also linear and so a similar approach can be followed. As in Sec. 6.4, we will restrict attention to small angles relative to some optic axis ("paraxial Fourier optics"). We shall describe the wave field at some distance z_j along the optic axis by the function $\psi_j(\mathbf{x})$, where $\mathbf{x}$ is a two dimensional vector perpendicular to the optic axis as in Fig. 7.3. If we consider a single linear optical device, then we can relate the output field ψ_2 at z_2 to the input ψ_1 at z_1 using a Greens' function denoted $P_{21}(\mathbf{x}_2, \mathbf{x}_1)$:

$$\psi_2(\mathbf{x}_2) = \int P_{21}(\mathbf{x}_2, \mathbf{x}_1) d^2x_1 \psi_1. \tag{7.27}$$

If ψ_1 were a δ-function, then the output would be simply given by the function P_{21}, up to normalization. For this reason, P_{21} is usually known as the *Point Spread Function*. Alternatively, we can think of it as a propagator. If we now combine two optical devices sequentially, so the output of the first device ψ_2 is the input of the second, then the point spread functions combine in the natural manner of any linear propagator to give a total point spread function

$$P_{31}(\mathbf{x}_3, \mathbf{x}_1) = \int P_{32}(\mathbf{x}_3, \mathbf{x}_2) d^2x_2 P_{21}(\mathbf{x}_2, \mathbf{x}_1) \ . \tag{7.28}$$

Just as the simplest matrix for paraxial, geometric-optics propagation is that for free propagation through some distance d, so also the simplest point spread function is that for free propagation. From Eq. (7.21) we see that it is given by

$$P_{21} = \frac{-ik}{2\pi d} e^{ikd} \exp\left(\frac{ik(\mathbf{x}_1 - \mathbf{x}_2)^2}{2d}\right) \ , \tag{7.29}$$

where $d = z_2 - z_1$ is the distance of propagation along the optic axis. Note that this P_{21} depends upon only on $\mathbf{x}_1 - \mathbf{x}_2$ and not on $\mathbf{x}_1$ or $\mathbf{x}_2$ individually, as it should because there is translational invariance in the $\mathbf{x}_1, \mathbf{x}_2$ planes.

A thin lens adds or subtracts an extra phase $\Delta\phi$ to the wave, and $\Delta\phi$ depends quadratically on distance from the optic axis ($|\mathbf{x}|$), so that the angle of deflection, which is proportional to the gradient of the phase, will depend linearly on $\mathbf{x}$. Correspondingly, the

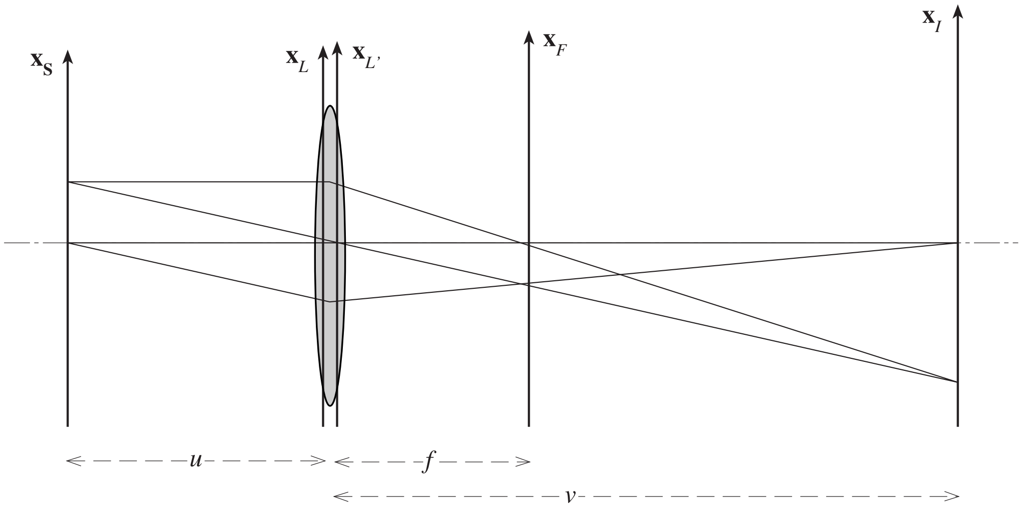

Fig. 7.11: Wave theory of a single converging lens.

point-spread function for a thin lens is

$$P_{21} = \exp\left(\frac{-ik|\mathbf{x}_1|^2}{2f}\right)\delta(\mathbf{x}_2 - \mathbf{x}_1) \tag{7.30}$$

where f is the focal length, positive for a converging lens and negative for a diverging lens.

7.5.3 Abbé's Description of Image Formation by a Thin Lens

We can use these two point spread functions to give a wave description of the production of images by a single converging lens, in parallel to the geometric-optics description of Figs. 6.3 and 6.5. We shall do this in two stages. First, we shall propagate the wave from the source plane S a distance u in front of the lens, through the lens L, to its focal plane F a distance f behind the lens (Fig. 7.11). Then we shall propagate the wave a further distance $v - f$ from the focal plane to the image plane. We know from geometric optics that $v = fu/(f-u)$ [Eq. (6.41)].

Using Eqs. (7.28), (7.29), (7.30), we obtain for the propagator from the source plain to the focal plane

$$\begin{aligned}
P_{FS} &= \int P_{FL'} d^2x'_L P_{L'L} d^2x_L P_{LS} \\
&= \int \frac{ik}{2\pi f} e^{ikf} \exp\left(\frac{ik(\mathbf{x}_F - \mathbf{x}'_L)^2}{2f}\right) d^2x_{L'} \delta(\mathbf{x}_{L'} - \mathbf{x}_L) \exp\left(\frac{-ik|\mathbf{x}_L|^2}{2f}\right) \\
&\quad \times \frac{-ik}{2\pi u} e^{iku} \exp\left(\frac{ik(\mathbf{x}_S - \mathbf{x}_L)^2}{2u}\right) d^2x_L \\
&= \frac{-ik}{2\pi f} e^{ik(f+u)} \exp\left(-\frac{ikx_F^2}{2(v-f)}\right) \exp\left(-\frac{ik\mathbf{x}_F \cdot \mathbf{x}_S}{f}\right) \; .
\end{aligned} \tag{7.31}$$

Here we have extended all integrations to $\pm\infty$ and have used the values of the Fresnel integrals at infinity, $S(\pm\infty) = \pm(1+i)/2$ to get the expression on the last line. The wave in the focal plane is given by

$$\begin{aligned}\psi_F(\mathbf{x}_F) &= \int P_{FS} d^2x_S \psi_S(\mathbf{x}_S) \\ &= -\frac{ik}{2\pi f} e^{ik(f+u)} \exp\left(-\frac{ikx_F^2}{2(v-f)}\right) \tilde{\psi}_S(\mathbf{x}_F/f) \qquad (7.32)\end{aligned}$$

where

$$\tilde{\psi}_S(\boldsymbol{\theta}) = \int d^2x_S \psi_S(\mathbf{x}_S) e^{-i\boldsymbol{\theta}\cdot\mathbf{x}_S} \; . \qquad (7.33)$$

Thus, we have shown that the field in the back focal plane is, apart from an unimportant phase factor, proportional to the Fourier transform of the field in the source plane; in other words, *the focal-plane field is the Fraunhofer diffraction pattern of the input wave.* That this has to be the case can be understood from Fig. 7.11. The focal plane F is where the converging lens brings parallel rays from the source plane to a focus. By doing so, *the lens in effect brings in from "infinity" the Fraunhofer diffraction pattern of the source, and places it into the focal plane.*

It now remains to propagate the final distance from the focal plane to the image plane. We do so with the free-propagation point-spread function of Eq. (7.29):

$$\begin{aligned}\psi_I &= \int P_{IF} d^2x_F \psi_F \qquad (7.34) \\ &= -\left(\frac{u}{v}\right) e^{ik(u+v)} \exp\left(\frac{ikx_I^2}{2(v-f)}\right) \psi_S(\mathbf{x}_S = -\mathbf{x}_I u/v) \; . \qquad (7.35)\end{aligned}$$

This says that (again ignoring a phase factor) *the wave in the image plane is just a magnified version of the wave in the source plane*, as we might have expected from geometric optics. In words, *the lens acts by taking the Fourier transform of the source and then takes the Fourier transform again to recover the source structure.* This description of image formation was developed by Ernst Abbé in 1873.

The focal plane is a convenient place to process the image by altering its Fourier transform—a process known as *spatial filtering.* One simple example is a *low-pass filter* in which a small circular aperture or "stop" is introduced into the focal plane, thereby allowing only the low-order spatial Fourier components to be transmitted to the image plane. This will obviously lead to considerable smoothing of the wave. An application is to the output beam from a laser (Chap. 9), which ought to be smooth but has high spatial frequency structure on account of noise and imperfections in the optics. A low-pass filter can be used to clean the beam. In the language of Fourier transforms, if we multiply the transform of the source, in the back focal plane, by a small-diameter circular aperture function, we will thereby convolve the image with a broad Airy-disk smoothing function. Conversely, we can exclude the low spatial frequencies with a high-pass filter, e.g. by placing an opaque circular disk in the focal plane, centered on the optic axis. This will have the effect of accentuating boundaries and discontinuities in the source and can be used to highlight features where the gradient of the

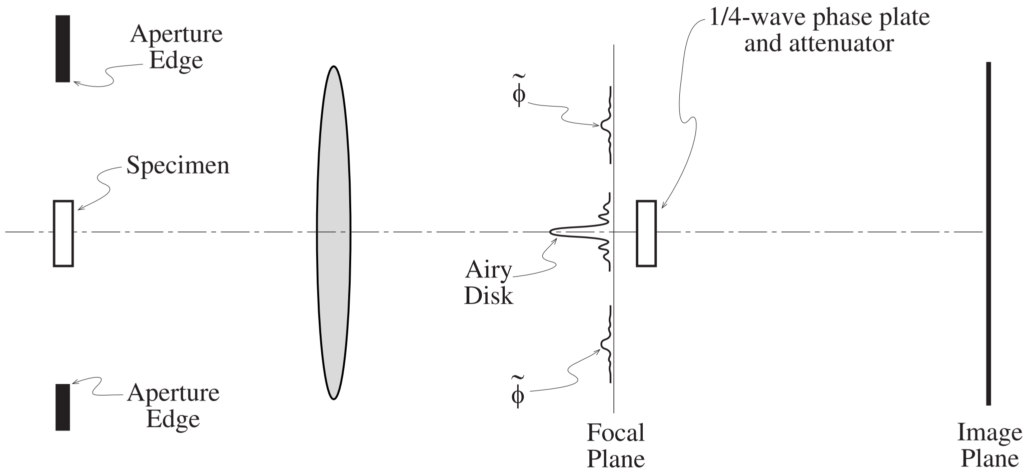

Fig. 7.12: Schematic Phase Contrast Microscope.

brightness is considerable. Another type of filter is used when the image is pixellated and thus has unwanted structure with wavelength equal to pixel size: a narrow range of frequencies centered around this spatial frequency is removed by putting an appropriate filter in the back focal plane.

7.5.4 Phase Contrast Microscopy

"Phase contrast microscopy" is a very useful technique for studying small objects, such as transparent biological specimens, that modify the phase of coherent illuminating light, but not its amplitude. Let us suppose that the phase change in the specimen, $\phi(\mathbf{x})$, is small, as often is the case for biological specimens. We can then write the field just after it passes through the specimen as

$$\psi_S(\mathbf{x}) = H(\mathbf{x})e^{i\phi(\mathbf{x})} \simeq H(\mathbf{x}) + i\phi(\mathbf{x})H(\mathbf{x}) \; ; \tag{7.36}$$

see Fig. 7.12. Here H is the microscope's aperture function, unity for $|\mathbf{x}| < D/2$ and zero for $|\mathbf{x}| > D/2$, with D the aperture diameter. The intensity is not modulated, and therefore the effect of the specimen on the wave is very hard to observe unless one is clever.

Equation (7.36) and the linearity of the Fourier transform imply that the wave in the focal plane is the sum of (i) the Fourier transform of the aperture function, i.e. an Airy function (bright spot with very small diameter), and (ii) the transform of the phase function convolved with that of the aperture (in which the fine-scale variations of the phase function dominate and push $\tilde{\phi}$ to large radii in the focal plane, Fig. 7.12, and the aperture has little influence):

$$\psi_F \sim \text{jinc}\left(\frac{kD|\mathbf{x}_F|}{2f}\right) + i\tilde{\phi}\left(\frac{k\mathbf{x}_F}{f}\right) \; . \tag{7.37}$$

If a high pass filter is used to remove the Airy disk completely then the remaining wave in the image plane will be essentially ϕ magnified by v/u. The flux will still be quadratic in the phase and so the contrast in the image will be small. A better technique is to phase shift the Airy disk in the focal plane by $\pm\pi/2$ so that the two terms in Eq. (7.37) are in phase. The intensity variations $[\propto (1\pm\phi)^2]$ will now be linear in the phase ϕ. An even better procedure is to attenuate the Airy disk until its amplitude is comparable with the rms value of ϕ and also phase shift it by $\pm\pi/2$. This will maximise the contrast in the final image. Analogous techniques are used in communications to inter-convert amplitude-modulated and phase-modulated signals.

7.5.5 Gaussian Beams

The mathematical techniques of Fourier optics enable us to analyze the structure and propagation of light beams that have Gaussian profiles. (Such Gaussian beams are the natural output of ideal lasers, they are the real output of spatially filtered lasers, and they are widely used for optical communications, interferometry and other practical applications. Moreover, they are the closest one can come in the real world of wave optics to the idealization of a geometric-optics pencil beam.)

Consider a beam that is precisely plane fronted, with a Gaussian profile, at location $z = 0$ on the optic axis,

$$\psi_0 = \exp\left(\frac{-\varpi^2}{\sigma_0^2}\right) \; ; \tag{7.38}$$

here $\varpi = |\mathbf{x}|$ is radial distance from the optic axis. The form of this same wave at a distance z further down the optic axis can be computed by folding this ψ_0 into the point spread function (7.29) (with the distance d replaced by z). The result is

$$\psi_z = \frac{1}{(1+z^2/z_0^2)^{1/2}} \exp\left(\frac{-\varpi^2}{\sigma_0^2(1+z^2/z_0^2)}\right) \exp\left[i\left(\frac{k\varpi^2}{2z(1+z_0^2/z^2)} - \tan^{-1}\frac{z}{z_0} + kz\right)\right], \tag{7.39}$$

where

$$z_0 = \frac{k\sigma_0^2}{2} = \frac{\pi\sigma_0^2}{\lambda} \; . \tag{7.40}$$

Formula (7.39) for the freely propagating beam is valid for negative z as well as positive. Notice that the beam's radius (distance from optic axis to the point of $1/e$ attenuation of its amplitude) is

$$\sigma_z = \sigma_0(1+z^2/z_0^2)^{1/2} \; . \tag{7.41}$$

This beam radius is a minimum at $z = 0$ (the beam's waist), and increases away from there in either direction.

The Gaussian beam's form (7.40) near some arbitrary location z is fully characterized by three parameters: the wavelength $\lambda = 2\pi/k$, the distance z to the waist, and the beam radius at the waist σ_0 [from which the local beam radius σ_z can be computed by Eq. (7.41)]. At location z, the beam's wave fronts (surfaces of constant phase) have radius of curvature $R_z = z(1 + z_0^2/z^2)$. The radius of curvature is infinite at the waist. Near the waist, in the Fresnel region$(\lambda z)^{1/2} \ll \sigma_0$, it decreases with distance as $R_z \simeq z_0^2/z$. It reaches a minimum value at the boundary between the Fresnel and the Fraunhofer regions, and it then begins to increase as $R_z \propto z$. Correspondingly, in the Fresnel region the beam radius is nearly constant, $\sigma_z \simeq \sigma_0$, while in the Fraunhofer region it increases linearly with distance,

$\sigma_z \simeq \sigma_0 z/z_0$. These are just the behaviors that one should expect from the uncertainty principle analysis at the end of Sec. 7.2.

It is easy to compute the effects of a thin lens on a Gaussian beam by folding the ψ_z at the lens's location into the lens point spread function (7.30). The result is a phase change that preserves the general Gaussian form of the wave, but alters the distance to the waist and the radius at the waist. Thus, by judicious placement of lenses (or, equally well curved mirrors), and judicious choices of their focal lengths, one can tailor the parameters of a Gaussian beam to fit whatever optical device one is working with. For example, if one wants to send a Gaussian beam into a self-focusing optical fiber, one should place its waist at the entrance to the fiber, and adjust its waist size there to coincide with that of the fiber's Gaussian mode of propagation (the mode analyzed in Ex. 7.8). The beam will then enter the fiber smoothly, and will propagate steadily along the fiber, with the effects of the transversely varying index of refraction continually compensating for the effects of diffraction so as to keep the phase fronts flat and the waist size constant.

****

EXERCISES

Exercise 7.9 *Problem: Guided Gaussian Beams*

Consider a self-focusing optical fiber discussed in the previous chapter in which the refractive index is

$$n(\mathbf{x}) = n_0(1 - \alpha^2\varpi^2)^{1/2} , \tag{7.42}$$

where $\varpi = |\mathbf{x}|$.

(a) Write down the Helmholtz equation in cylindrical polar coordinates and seek an axisymmetric mode for which $\psi = R(\varpi)Z(z)$, where R, Z are functions to be determined and z measures distance along the fiber. In particular show that there exists a mode with a Gaussian radial profile that propagates along the fiber without spreading.

(b) Compute the group and phase velocities along the fiber for this mode.

Exercise 7.10 *Problem: Convolution via Fourier Optics*

(a) Suppose that you have two thin sheets with transmission functions $t = g(x, y)$ and $t = h(x, y)$, and you wish to compute via Fourier optics the convolution

$$g \otimes h(x_o, y_o) \equiv \int\int g(x, y)h(x + x_o, y + y_o)dxdy . \tag{7.43}$$

Devise a method for doing so using Fourier optics. [Hint: use several lenses and a projection screen with a pinhole through which passes light whose intensity is proportional to the convolution; place the two sheets at strategically chosen locations along the optic axis, and displace one of the two sheets transversely with respect to the other.]

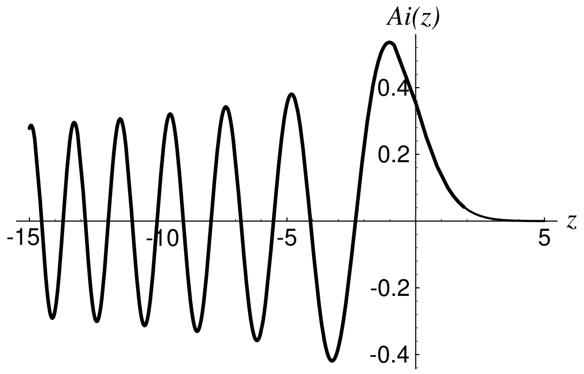

Fig. 7.13: The Airy Function $\mathrm{Ai}(z)$ describing diffraction at a caustic. The argument is $z = -bx/a^{1/3}$ where x is distance from the caustic and a, b are constants.

(b) Suppose you wish to convolve a large number of different one-dimensional functions simultaneously, i.e. you want to compute

$$g_j \otimes h_j(x_o) \equiv \int g_j(x) h_j(x + x_o) dx \tag{7.44}$$

for $j = 1, 2, \ldots$. Devise a way to do this via Fourier optics using appropriately constructed transmissive sheets and cylindrical lenses.

****

7.6 Diffraction at a Caustic

In Sec. 6.6, we described how caustics can be formed in general in the geometric-optics limit—e.g., on the bottom of a swimming pool when the water's surface is randomly rippled, or behind a gravitational lens. We chose as an example a simple phase changing screen illuminated by a point source and observed from some fixed distance r, and we showed how a pair of images would merge as the transverse distance x of the observer from the caustic decreases to zero. We expanded the phase in a Taylor series $\phi(s, x) = as^3/3 - bxs$, where the coefficients a, b are constant and s is a transverse coordinate in the screen (cf. Fig. 6.10). We were then able to show that the magnification of the images diverged $\propto x^{-1/2}$ [Eq. (6.64)], as the caustic was approached. This raised the question of what happens when we take into account the finite wavelength of the wave.

We are now in a position to answer this question. We simply use the Helmholtz-Kirchhoff integral (7.6) to write the expression for the amplitude measured at position x in the form

$$\psi(x) \propto \frac{1}{\lambda r} \int ds e^{i\phi(s,x)} , \tag{7.45}$$

ignoring multiplicative constants and constant phase factors. The phase ϕ varies rapidly with s at large $|s|$, so we can treat the limits of integration as $\pm\infty$. The integral turns out to be the Airy function

$$\int_{-\infty}^{\infty} ds \cos(as^3/3 - bxs) = \frac{2\pi}{a^{1/3}} \mathrm{Ai}(-bx/a^{1/3}) \ . \tag{7.46}$$

$\mathrm{Ai}(z)$ is displayed in Fig. 7.13.

The asymptotic behavior of $\mathrm{Ai}(z)$ is

$$\begin{aligned} \mathrm{Ai}(z) &\sim \pi^{-1/2} z^{-1/4} \sin(2z^{3/2}/3 + \pi/4), \quad z \to -\infty \\ &\sim \frac{e^{-2z^{3/2}/3}}{2\pi^{1/2} z^{1/4}}, \quad z \to \infty \ . \end{aligned} \tag{7.47}$$

We see that the amplitude ψ remains finite as the caustic is approached instead of diverging as in the geometric-optics limit. Furthermore, for $x > 0$ (left part of Fig. 7.13), where an observer sees two geometric-optics images, the envelope of ψ diminishes $\propto x^{-1/4}$, so that the intensity $|\psi|^2$ decreases $\propto x^{-1/2}$ just as in the geometric-optics limit. The peak magnification is $\propto a^{-2/3}$. What is actually seen is a series of bands alternating dark and light with spacing calculable using $\Delta(2z^{3/2}/3) = \pi$ or $\Delta x \propto x^{-1/2}$. At sufficient distance from the caustic, it will not be possible to resolve these bands and a uniform illumination of average intensity will be observed. In other words, we have recovered the geometric-optics limit. The scalings derived above, just like the geometric-optics scalings are a universal property of this type of caustic (the simplest caustic of all, the "fold").

There is a helpful analogy, familiar from quantum mechanics. Consider a particle in a harmonic potential well in a very excited state. Its wave function is given in the usual way using Hermite polynomials of large order. Close to the classical turning point, these functions change from being oscillatory to an exponential decay, just like the Airy function (and if we were to expand about the turning point, we would recover Airy functions). What is happening, of course is that the probability density of finding the particle close to its turning point diverges because it is moving very slowly there, and the oscillations are due to interference between waves associated with the particle moving in opposite directions. If we just consider the motions of photons parallel to the aperture then we have essentially the same problem here; the oscillations are associated with interference of the waves associated with the motions of the photons in two beams, one from each of the geometric-optics images. This is our first illustration of the formation of large contrast interference fringes when only a few beams are combined. We shall meet other examples of such interference in the following chapter.

****

EXERCISES

Exercise 7.11 *Wavelength scaling at a caustic*

Assume that the phase variation introduced at the screen in Fig. 6.9 is non-dispersive so that the $\phi(s, x)$ in Eq. (7.45) is $\phi \propto \lambda^{-1}$. Show that the peak magnification of the interference fringes at the caustic scales with wavelength $\propto \lambda^{-4/3}$. Also show that the spacing of the fringes at a given observing position is proportional to the wavelength.

****

Bibliography

Berry, M. V. & Upstill, C. 1980 *Prog. Optics* 18 257

Born, M. & Wolf, E. 1975 *Principles of Optics* Oxford: Pergamon

Goodman, J. W. *Introduction to Fourier Optics* New York: McGraw-Hill

Hecht, E. 1989 *Optics* New York: Addison Wesley

Longhurst, R. S. 1973 *Geometrical and Physical Optics* London: Longmans

Welford, W. T. 1988 *Optics* Oxford: Oxford University Press