where $U(P, \nu)$ is the Fourier spectrum of $u(P, t)$. Assuming the geometry of Fig. 3.6 show that if

$$\frac{\Delta\nu}{\bar{\nu}} \ll 1 \qquad \text{and} \qquad \frac{1}{\Delta\nu} \gg \frac{nr_{01}}{\upsilon}$$

then

$$u_{-}(P_0, t) = \frac{1}{j\bar{\lambda}} \iint_{-\infty}^{\infty} u_{-}(P_1, t) \frac{\exp(j\bar{k}r_{01})}{r_{01}} \cos(\vec{n}, \vec{r}_{01})\, ds$$

where $\bar{\lambda} = \upsilon/\bar{\nu}$ and $\bar{k} = 2\pi/\bar{\lambda}$. In the above equations, n is the refractive index of the medium and υ is the velocity of propagation.

3-7. For a wave that travels only in directions that have small angles with respect to the optical axis, the general form of the complex field may be approximated by

$$U(x, y, z) \approx A(x, y, z) \exp(jkz),$$

where $A(x, y, z)$ is a slowly varying function of z.

(a) Show that for such a wave the Helmholtz equation can be reduced to

$$\nabla_t^2 A + j2k \frac{\partial A}{\partial z} = 0,$$

where $\nabla_t^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse portion of the Laplacian. This equation is known as the *paraxial Helmholtz equation*.

(b) Show that a solution to this equation is given by

$$A(x, y, z) = \frac{A_1}{q(z)} \exp\left[jk\frac{x^2 + y^2}{2q(z)}\right]$$

for any complex $q(z)$ having $\frac{d}{dz}q(z) = 1$.

(c) Given

$$\frac{1}{q(z)} = \frac{1}{R(z)} + j\frac{\lambda}{\pi W^2(z)},$$

show that the solution $U(x, y, z)$ takes the form

$$U(x, y, z) = A_1 \frac{W_0}{W(z)} \exp\left[-\frac{\rho^2}{W^2(z)}\right] \exp\left[jkz + jk\frac{\rho^2}{2R(z)} + j\theta(z)\right]$$

where W_0 is a constant (independent of z) and $\theta(z)$ is a phase angle that changes with z. Note that this is a beam with a Gaussian profile and with a quadratic-phase approximation to a spherical wavefront.

4 Fresnel and Fraunhofer Diffraction

In the preceding chapter the results of scalar diffraction theory were presented in their most general forms. Attention is now turned to certain approximations to the general theory, approximations that will allow diffraction pattern calculations to be reduced to comparatively simple mathematical manipulations. These approximations, which are commonly made in many fields that deal with wave propagation, will be referred to as *Fresnel* and *Fraunhofer* approximations. In accordance with our view of the wave propagation phenomenon as a "system," we shall attempt to find approximations that are valid for a wide class of "input" field distributions.

4.1 Background

In this section we prepare the reader for the calculations to follow. The concept of the *intensity* of a wave field is introduced, and the Huygens-Fresnel principle, from which the approximations are derived, is presented in a form that is especially well suited for approximation.

4.1.1 The Intensity of a Wave Field

In the optical region of the spectrum, a photodetector responds directly to the optical power falling on its surface. Thus for a semiconductor detector, if optical power $\mathcal{P}$ is incident on the photosensitive region, absorption of a photon generates an electron in the conduction band and a hole in the valence band. Under the influence of internal and applied fields, the hole and electron move in opposite directions, leading to a photocurrent i that is the response to the incident absorbed photon. Under most circumstances the photocurrent is linearly proportional to the incident power,

$$i = \mathcal{R}\mathcal{P}. \tag{4-1}$$

The proportionality constant $\mathcal{R}$ is called the *responsivity* of the detector and is given by

$$\mathcal{R} = \frac{\eta_{qe} q}{h\nu}, \tag{4-2}$$

where η_{qe} is the *quantum efficiency* of the photodetector (the average number of electron-hole pairs released by the absorption of a photon, a quantity that is less than or equal to unity in the absence of internal gain), q is the electronic charge (1.602×10^{-19} coulombs), h is Planck's constant (6.626196×10^{-34} joule-second), and ν is the optical frequency.[1]

Thus in optics the directly measurable quantity is optical power, and it is important to relate such power to the complex scalar fields $u(P, t)$ and $U(P)$ dealt with in earlier discussions of diffraction theory. To understand this relation requires a return to an electromagnetic description of the problem. We omit the details here, referring the reader to Ref. [271], Sections 5.3 and 5.4, and simply state the major points. Let the medium be isotropic, and the wave monochromatic. Assuming that the wave behaves *locally* as a transverse electromagnetic plane wave (i.e., $\vec{\mathcal{E}}$, $\vec{\mathcal{H}}$, and $\vec{k}$ form a mutually orthogonal triplet), then the electric and magnetic fields can be expressed locally as

$$\begin{aligned} \vec{\mathcal{E}} &= \operatorname{Re}\{\vec{E}_0 \exp[-j(2\pi\nu t - \vec{k}\cdot\vec{r})]\} \\ \vec{\mathcal{H}} &= \operatorname{Re}\{\vec{H}_0 \exp[-j(2\pi\nu t - \vec{k}\cdot\vec{r})]\}, \end{aligned} \tag{4-3}$$

where $\vec{E}_0$ and $\vec{H}_0$ are locally constant and have complex components. The power flows in the direction of the vector $\vec{k}$ and the power density can be expressed as

$$p = \frac{\vec{E}_0 \cdot \vec{E}_0^*}{2\eta} = \frac{E_{0X}^2 + E_{0Y}^2 + E_{0Z}^2}{2\eta}, \tag{4-4}$$

where η is the *characteristic impedance* of the medium and is given by

$$\eta = \sqrt{\frac{\mu}{\epsilon}}.$$

In vacuum, η is equal to 377 Ω. The total power incident on a surface of area $\mathcal{A}$ is the integral of the power density over A, taking into account that the direction of power flow is in the direction of $\vec{k}$,

$$\mathcal{P} = \iint_A p \frac{\vec{k}\cdot\hat{n}}{|\vec{k}|}\, dx\, dy.$$

[1] The reader may wonder why the generation of both an electron and a hole does not lead to a charge $2q$ rather than q in this equation. For an answer, see [271], p. 653.

Here $\hat{n}$ is a unit vector pointing into the surface of the detector, while $\vec{k}/|\vec{k}|$ is a unit vector in the direction of power flow. When $\vec{k}$ is nearly normal to the surface, the total power $\mathcal{P}$ is simply the integral of the power density p over the detector area.

The proportionality of power density to the squared magnitude of the $\vec{E}_0$ vector seen in Eq. (4-4) leads us to define the *intensity* of a scalar monochromatic wave at point P as the squared magnitude of the complex phasor representation $U(P)$ of the disturbance,

$$I(P) = |U(P)|^2. \tag{4-5}$$

Note that power density and intensity are not identical, but the latter quantity is directly proportional to the former. For this reason we regard the intensity as the physically measurable attribute of an optical wavefield.

When a wave is not perfectly monochromatic, but is narrow band, a straightforward generalization of the concept of intensity is given by

$$I(P) = \langle |u(P, t)|^2 \rangle, \tag{4-6}$$

where the angle brackets signify an infinite time average. In some cases, the concept of *instantaneous intensity* is useful, defined as

$$I(P, t) = |u(P, t)|^2. \tag{4-7}$$

When calculating a diffraction pattern, we will generally regard the intensity of the pattern as the quantity we are seeking.

4.1.2 The Huygens-Fresnel Principle in Rectangular Coordinates

Before introducing a series of approximations to the Huygens-Fresnel principle, it will be helpful to first state the principle in more explicit form for the case of rectangular coordinates. As shown in Fig. 4.1, the diffracting aperture is assumed to lie in the (ξ, η) plane, and is illuminated in the positive z direction. We will calculate the wavefield across the (x, y) plane, which is parallel to the (ξ, η) plane and at normal distance z from it. The

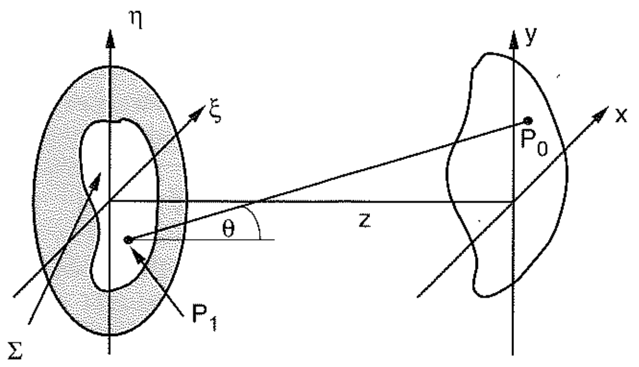

Figure 4.1 Diffraction geometry.

z axis pierces both planes at their origins. According to Eq. (3-41), the Huygens-Fresnel principle can be stated as

$$U(P_0) = \frac{1}{j\lambda}\iint_{\Sigma} U(P_1)\frac{\exp(jkr_{01})}{r_{01}}\cos\theta\, ds, \tag{4-8}$$

where θ is the angle between the outward normal $\hat{n}$ and the vector $\vec{r}_{01}$ pointing from P_0 to P_1. The term $\cos\theta$ is given exactly by

$$\cos\theta = \frac{z}{r_{01}},$$

and therefore the Huygens-Fresnel principle can be rewritten

$$U(x, y) = \frac{z}{j\lambda}\iint_{\Sigma} U(\xi, \eta)\frac{\exp(jkr_{01})}{r_{01}^2}\, d\xi\, d\eta, \tag{4-9}$$

where the distance r_{01} is given exactly by

$$r_{01} = \sqrt{z^2 + (x-\xi)^2 + (y-\eta)^2}. \tag{4-10}$$

There have been only two approximations in reaching this expression. One is the approximation inherent in the scalar theory. The second is the assumption that the observation distance is many wavelengths from the aperture, $r_{01} \gg \lambda$. We now embark on a series of additional approximations.

4.2 The Fresnel Approximation

To reduce the Huygens-Fresnel principle to a more simple and usable expression, we introduce approximations for the distance r_{01} between P_1 and P_0. The approximations are based on the binomial expansion of the square root in Eq. (4-10). Let b be a number that is less than unity, and consider the expression $\sqrt{1+b}$. The binomial expansion of the square root is given by

$$\sqrt{1+b} = 1 + \frac{1}{2}b - \frac{1}{8}b^2 + \cdots, \tag{4-11}$$

where the number of terms needed for a given accuracy depends on the magnitude of b.

To apply the binomial expansion to the problem at hand, factor a z outside the expression for r_{01}, yielding

$$r_{01} = z\sqrt{1 + \left(\frac{x-\xi}{z}\right)^2 + \left(\frac{y-\eta}{z}\right)^2}. \tag{4-12}$$

Let the quantity b in Eq. (4-11) consist of the second and third terms under the square root in (4-12). Then, retaining only the first two terms of the expansion (4-11), we have

$$r_{01} \approx z\left[1 + \frac{1}{2}\left(\frac{x-\xi}{z}\right)^2 + \frac{1}{2}\left(\frac{y-\eta}{z}\right)^2\right]. \tag{4-13}$$

The question now arises as to whether we need to retain all the terms in the approximation (4-13), or whether only the first term might suffice. The answer to this question depends on which of the several occurrences of r_{01} is being approximated. For the r_{01}^2 appearing in the denominator of Eq. (4-9), the error introduced by dropping all terms but z is generally acceptably small. However, for the r_{01} appearing in the exponent, errors are much more critical. First, they are multiplied by a very large number k, a typical value for which might be greater than 10^7 in the visible region of the spectrum (e.g., $\lambda = 5\times10^{-7}$meters). Second, phase changes of as little as a fraction of a radian can change the value of the exponential significantly. For this reason we retain both terms of the binomial approximation in the exponent. The resulting expression for the field at (x, y) therefore becomes

$$U(x, y) = \frac{e^{jkz}}{j\lambda z}\iint_{-\infty}^{\infty} U(\xi, \eta)\exp\left\{j\frac{k}{2z}\left[(x-\xi)^2 + (y-\eta)^2\right]\right\} d\xi\, d\eta, \tag{4-14}$$

where we have incorporated the finite limits of the aperture in the definition of $U(\xi, \eta)$, in accord with the usual assumed boundary conditions.

Equation (4-14) is readily seen to be a convolution, expressible in the form

$$U(x, y) = \iint_{-\infty}^{\infty} U(\xi, \eta)\, h(x-\xi, y-\eta)\, d\xi\, d\eta \tag{4-15}$$

where the convolution kernel is

$$h(x, y) = \frac{e^{jkz}}{j\lambda z}\exp\left[\frac{jk}{2z}\left(x^2 + y^2\right)\right]. \tag{4-16}$$

We will return to this viewpoint a bit later.

Another form of the result (4-14) is found if the term

$$\exp\left[\frac{jk}{2z}\left(x^2 + y^2\right)\right]$$

is factored outside the integral signs, yielding

$$U(x, y) = \frac{e^{jkz}}{j\lambda z}e^{j\frac{k}{2z}(x^2+y^2)}\iint_{-\infty}^{\infty}\left\{U(\xi, \eta)e^{j\frac{k}{2z}(\xi^2+\eta^2)}\right\}e^{-j\frac{2\pi}{\lambda z}(x\xi+y\eta)}\, d\xi d\eta, \tag{4-17}$$

which we recognize (aside from multiplicative factors) to be the *Fourier transform* of the product of the complex field just to the right of the aperture and a quadratic phase exponential.

We refer to both forms of the result, (4-14) and (4-17), as *the Fresnel diffraction integral*. When this approximation is valid, the observer is said to be in the region of Fresnel diffraction, or equivalently in the *near field* of the aperture.[2]

[2] Recently an interesting relation between the Fresnel diffraction formula and an entity known as the "fractional Fourier transform" has been found. The interested reader can consult Ref. [242] and the references contained therein.

4.2.1 Positive vs. Negative Phases

We have seen that it is common practice when using the Fresnel approximation to replace expressions for spherical waves by quadratic-phase exponentials. The question often arises as to whether the sign of the phase should be positive or negative in a given expression. This question is not only pertinent to quadratic-phase exponentials, but also arises when considering exact expressions for spherical waves and when considering plane waves propagating at an angle with respect to the optical axis. We now present the reader with a methodology that will help determine the proper sign of the exponent in all of these cases.

The critical fact to keep in mind is that we have chosen our phasors to rotate in the *clockwise* direction, i.e., their time dependence is of the form $\exp(-j2\pi\nu t)$. For this reason, if we move in space in such a way as to intercept portions of a wavefield that were emitted *later* in time, the phasor will have advanced in the clockwise direction, and therefore the phase must become more *negative*. On the other hand, if we move in space to intercept portions of a wavefield that were emitted *earlier* in time, the phasor will not have had time to rotate as far in the clockwise direction, and therefore the phase must become more *positive*.

If we imagine observing a spherical wave that is diverging from a point on the z axis, the observation being in an (x, y) plane that is normal to that axis, then movement away from the origin always results in observation of portions of the wavefront that were emitted earlier in time than that at the origin, since the wave has had to propagate further to reach those points. For that reason the phase must increase in a positive sense as we move away from the origin. Therefore the expressions

$$\exp(jkr_{01}) \quad \text{and} \quad \exp\left[j\frac{k}{2z}\left(x^2+y^2\right)\right]$$

(for positive z) represent a diverging spherical wave and a quadratic-phase approximation to such a wave, respectively. By the same token, $\exp(-jkr_{01})$ and $\exp[-j(k/2z)(x^2+y^2)]$ represent a converging spherical wave, again assuming that z is positive. Clearly, if z is a negative number, then the interpretation must be reversed, since a negative sign is hidden in z.

Similar reasoning applies to the expressions for plane waves traveling at an angle with respect to the optical axis. Thus for positive α, the expression $\exp(j2\pi\alpha y)$ represents a plane wave with a wave vector in the (y, z) plane. But does the wave vector point with a positive angle with respect to the z axis or with a negative angle, keeping in mind that a positive angle is one that has rotated counterclockwise with respect to the z axis? If we move in the positive y direction, the argument of the exponential increases in a positive sense, and therefore we are moving to a portion of the wave that was emitted earlier in time. This can only be true if the wave vector points with a positive angle with respect to the z axis, as illustrated in Fig. 4.2.

4.2.2 Accuracy of the Fresnel Approximation

Considering the approximation in the exponent, which is the most critical approximation, it can be seen that the *spherical* secondary wavelets of the Huygens-Fresnel principle have

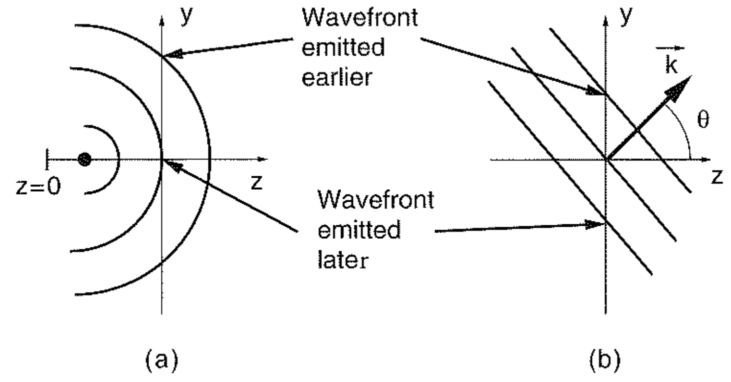

Figure 4.2 Determining the sign of the phases of exponential representations of (a) spherical waves and (b) plane waves.

been replaced by wavelets with parabolic wavefronts. The accuracy of this approximation is determined by the errors induced when terms higher than first order (linear in b) are dropped in the binomial expansion (4-11). A sufficient condition for accuracy would be that the maximum phase change induced by dropping the $b^2/8$ term be much less than 1 radian. This condition will be met if the distance z satisfies

$$z^3 \gg \frac{\pi}{4\lambda}[(x-\xi)^2+(y-\eta)^2]^2_{\max}. \tag{4-18}$$

For a circular aperture of size 1 cm, a circular observation region of size 1 cm, and a wavelength of 0.5 μm, this condition would indicate that the distance z must be $\gg$ 25 cm for accuracy. However, as the next comment will explain, this sufficient condition is overly stringent, and accuracy can be expected for much shorter distances.

For the Fresnel approximation to yield accurate results, it is not necessary that the higher-order terms of the expansion be small, only that they not change the value of the Fresnel diffraction integral significantly. Considering the convolution form of the result, Eq. (4-14), if the major contribution to the integral comes from points (ξ, η) for which $\xi \approx x$ and $\eta \approx y$, then the particular values of the higher-order terms of the expansion are unimportant.

To investigate this point more completely, expand the quadratic-phase exponential of Eq. (4-16) into its real and imaginary parts,

$$\frac{1}{j\lambda z}\exp\left[j\frac{\pi}{\lambda z}\left(x^2+y^2\right)\right] = \frac{1}{j\lambda z}\left\{\cos\left[\frac{\pi}{\lambda z}\left(x^2+y^2\right)\right] + j\sin\left[\frac{\pi}{\lambda z}\left(x^2+y^2\right)\right]\right\}, \tag{4-19}$$

where we have dropped the unit magnitude phasor e^{jkz} simply by redefining the phase reference, and we have replaced k by $2\pi/\lambda$. The volume under this function can readily be shown to be unity (Prob. 4-1). Figure 4.3 shows plots of one-dimensional quadratic-phase cosine and sine functions $\cos(\pi x^2)$ and $\sin(\pi x^2)$. Each of these functions has area $1/\sqrt{2}$. Using this fact it can be shown that all of the unit area under the two-dimensional quadratic-phase exponential is contributed by the two-dimensional sinusoidal term.

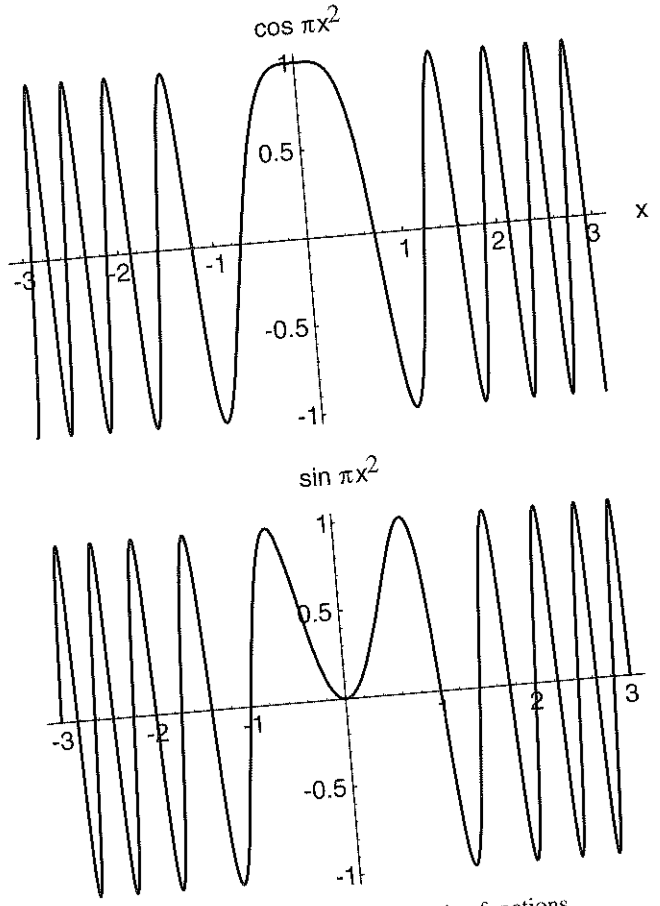

Figure 4.3 Quadratic phase cosine and sine functions.

Figure 4.4 shows the magnitude of the integral of a quadratic-phase exponential function,

$$\left|\int_{-X}^{X} \exp(j\pi x^2)\,dx\right| = \left|\sqrt{2}C\left(\sqrt{2}X\right) + j\sqrt{2}S\left(\sqrt{2}X\right)\right|$$

which has also been expressed in terms of the Fresnel integrals $C(z)$ and $S(z)$ mentioned in Section 2.2. As can be seen from the figure, the integral grows toward its asymptotic value of unity with increasing X. Note in particular that the integral first reaches unity when $X = 0.5$, and then oscillates about that value with diminishing fluctuations. We conclude that, to a reasonable approximation, the major contributions to a convolution of this function with a second function that is smooth and slowly varying will come from the range $-2 < X < 2$, due to the fact that outside this range the rapid oscillations of the integrand do not yield a significant addition to the total area.

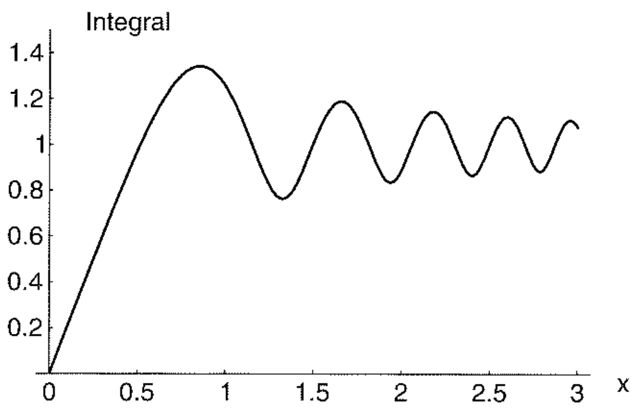

Figure 4.4 Magnitude of the integral of the quadratic-phase exponential function.

For the *scaled* quadratic-phase exponential of Eqs. (4-14) and (4-16), the corresponding conclusion is that the majority of the contribution to the convolution integral comes from a square in the (ξ, η) plane, with width $4\sqrt{\lambda z}$ and centered on the point $(\xi = x, \eta = y)$. This square grows in size as the distance z behind the aperture increases. In effect, when this square lies entirely within the open portion of the aperture, the field observed at distance z is, to a good approximation, what it would be if the aperture were not present. When the square lies entirely behind the obstruction of the aperture, then the observation point lies in a region that is, to a good approximation, dark due to the shadow of the aperture. When the square bridges the open and obstructed parts of the aperture, then the observed field is in the transition region between light and dark. The detailed structure within these regions may be complex, but the general conclusions above are correct. Figure 4.5 illustrates the various regions mentioned. For the case of a one-dimensional

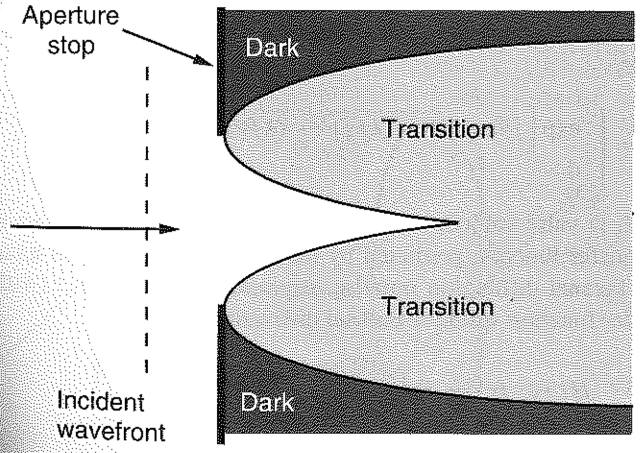

Figure 4.5 Light, dark, and transition regions behind a rectangular slit aperture.

rectangular slit, the boundaries between the light region and the transition region, and between the dark region and the transition region, can be shown to be parabolas (see Prob. 4-5).

Note that if the amplitude transmittance and/or the illumination of the diffracting aperture is *not* a relatively smooth and slowly varying function, the above conclusions may not hold. For example, if the amplitude of the field transmitted by the aperture has a high-spatial-frequency sinusoidal component, that component may interact with the high frequencies of the quadratic-phase exponential kernel to produce a nonzero contribution from a location other than the square mentioned above. Thus the restriction of attention to the square of width $4\sqrt{\lambda z}$ must be used with some caution. However, the idea is valid when the diffracting apertures do not contain fine structure and when they are illuminated by uniform plane waves.

If the distance z is allowed to approach zero, i.e., the observation point approaches the diffracting aperture, then the two-dimensional quadratic-phase function behaves in the limit like a delta function, producing a field $U(x, y)$ that is identical to the aperture field $U(\xi, \eta)$ in the aperture. In such a case, the predictions of geometrical optics are valid, for such a treatment would predict that the field observed behind the aperture is simply a geometrical projection of the aperture fields onto the plane of observation.

Our discussion above is closely related to the *principle of stationary phase*, a method for finding the asymptotic values of certain integrals. A good discussion of this method can be found in Appendix III of Ref. [29]. For other examinations of the accuracy of the Fresnel approximation, see Chapter 9 of Ref. [244] and also Ref. [290]. The general conclusions of all of these analyses are similar, namely, the accuracy of the Fresnel approximation is extremely good to distances that are very close to the aperture.

4.2.3 The Fresnel Approximation and the Angular Spectrum

It is of some interest to understand the implications of the Fresnel approximations from the point of view of the angular spectrum method of analysis. Such understanding can be developed by beginning with Eq. (3-74), which expresses the transfer function of propagation through free space,

$$H(f_X, f_Y) = \begin{cases} \exp\left[j2\pi \frac{z}{\lambda}\sqrt{1-(\lambda f_X)^2-(\lambda f_Y)^2}\right] & \sqrt{f_X^2+f_Y^2} < \frac{1}{\lambda} \\ 0 & \text{otherwise} \end{cases} . \qquad (4\text{-}20)$$

This result, which is valid subject only to the scalar approximation, can now be compared with the transfer function predicted by the results of the Fresnel analysis. Fourier transforming the Fresnel diffraction impulse response (4-16), we find (with the help of Table 2.1) a transfer function valid for Fresnel diffraction,

$$\begin{aligned} H(f_X, f_Y) &= \mathcal{F}\left\{\frac{e^{jkz}}{j\lambda z}\exp\left[j\frac{\pi}{\lambda z}\left(x^2+y^2\right)\right]\right\} \\ &= e^{jkz}\exp\left[-j\pi\lambda z\left(f_X^2+f_Y^2\right)\right]. \end{aligned} \qquad (4\text{-}21)$$

Thus in the Fresnel approximation, the general spatial phase dispersion representing propagation is reduced to a *quadratic* phase dispersion. The factor e^{jkz} on the right of this equation represents a constant phase delay suffered by all plane-wave components traveling between two parallel planes separated by normal distance z. The second term represents the different phase delays suffered by plane-wave components traveling in different directions.

The expression (4-21) is clearly an approximation to the more general transfer function (4-20). We can obtain the approximate result from the general result by applying a binomial expansion to the exponent of (4-20),

$$\sqrt{1-(\lambda f_X)^2-(\lambda f_Y)^2} \approx 1 - \frac{(\lambda f_X)^2}{2} - \frac{(\lambda f_Y)^2}{2}, \qquad (4\text{-}22)$$

which is valid provided $|\lambda f_X| \ll 1$ and $|\lambda f_Y| \ll 1$. Such restrictions on f_X and f_Y are simply restrictions to *small angles*. So we see that, from the perspective of the angular spectrum, the Fresnel approximation is accurate provided only small angles of diffraction are involved. It is for this reason that we often say that the Fresnel approximations and the *paraxial* approximation are equivalent.

4.2.4 Fresnel Diffraction between Confocal Spherical Surfaces

Until now, attention has been focused on diffraction between two *planes*. An alternative geometry, of more theoretical than practical interest but nonetheless quite instructive, is diffraction between two confocal spherical surfaces (see, for example, [25], [26]). As shown in Fig. 4.6, two spheres are said to be confocal if the center of each lies on the surface of the other. In our case, the two spheres are tangent to the planes previously used, with the points of tangency being the points where the z axis pierces those planes. The distance r_{01} in our previous diffraction analysis is now the distance between the two spherical caps shown.

A proper analysis would write equations for the left-hand spherical surface and for the right-hand spherical surface, and then use those equations to find the distance r_{01} between

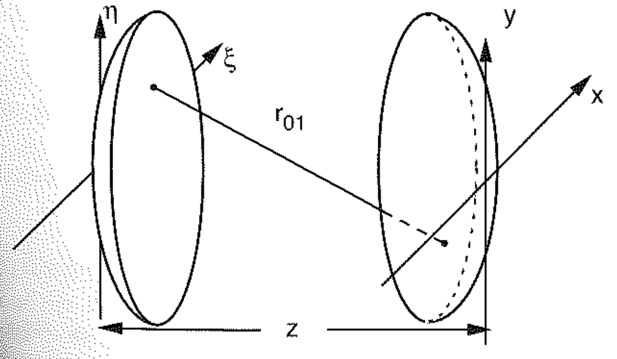

Figure 4.6 Confocal spherical surfaces.

the two spherical caps. In the process it would be helpful to simplify certain square roots by using the first two terms of their binomial expansions (i.e., to make *paraxial* approximations to the spherical surfaces). The result of such an analysis is the following simple expression for r_{01}, valid if the extent of the spherical caps about the z-axis is small compared with their radii:

$$r_{01} \approx z - x\xi/z - y\eta/z.$$

The Fresnel diffraction equation now becomes

$$U(x, y) = \frac{e^{jkz}}{j\lambda z} \iint_{-\infty}^{\infty} U(\xi, \eta) e^{-j\frac{2\pi}{\lambda z}(x\xi + y\eta)}\, d\xi\, d\eta, \tag{4-23}$$

which, aside from constant multipliers and scale factors, expresses the field observed on the right-hand spherical cap as the *Fourier transform* of the field on the left-hand spherical cap.

Comparison of this result with the previous Fourier-transform version of the Fresnel diffraction integral, Eq. (4-17), shows that the quadratic-phase factors in (x, y) and (ξ, η) have been eliminated by moving from the two planes to the two spherical caps. The two quadratic phase factors in the earlier expression are in fact simply paraxial representations of spherical phase surfaces, and it is therefore reasonable that moving to the spheres has eliminated them.

One subtle point worth mention is that, when we analyze diffraction between two spherical caps, it is not really valid to use the Rayleigh-Sommerfeld result as the basis for the calculation, for that result was explicitly valid only for diffraction by a planar aperture. However, the Kirchhoff analysis remains valid, and its predictions are the same as those of the Rayleigh-Sommerfeld approach provided paraxial conditions hold.

4.3 The Fraunhofer Approximation

Before presenting several examples of diffraction pattern calculations, we consider another more stringent approximation which, when valid, greatly simplifies the calculations. It was seen in Eq. (4-17) that, in the region of Fresnel diffraction, the observed field strength $U(x, y)$ can be found from a Fourier transform of the product of the aperture distribution $U(\xi, \eta)$ and a quadratic phase function $\exp\left[j(k/2z)(\xi^2 + \eta^2)\right]$. If in addition to the Fresnel approximation the stronger (Fraunhofer) approximation

$$z \gg \frac{k(\xi^2 + \eta^2)_{\max}}{2} \tag{4-24}$$

is satisfied, then the quadratic phase factor under the integral sign in Eq. (4-17) is approximately unity over the entire aperture, and the observed field strength can be found (up to a multiplicative phase factor in (x, y)) directly from a *Fourier transform* of the aperture distribution itself. Thus in the region of *Fraunhofer diffraction* (or equivalently, in the *far field*),

$$U(x, y) = \frac{e^{jkz} e^{j\frac{k}{2z}(x^2+y^2)}}{j\lambda z} \iint_{-\infty}^{\infty} U(\xi, \eta) \exp\left[-j\frac{2\pi}{\lambda z}(x\xi + y\eta)\right] d\xi\, d\eta. \tag{4-25}$$

Aside from multiplicative phase factors preceding the integral, this expression is simply the Fourier transform of the aperture distribution, evaluated at frequencies

$$\begin{aligned} f_X &= x/\lambda z \\ f_Y &= y/\lambda z. \end{aligned} \tag{4-26}$$

At optical frequencies, the conditions required for validity of the Fraunhofer approximation can be severe ones. For example, at a wavelength of 0.6 μm (red light) and an aperture width of 2.5 cm (1 inch), the observation distance z must satisfy

$$z \gg 1{,}600 \text{ meters}.$$

An alternative, less stringent condition, known as the "antenna designer's formula," states that for an aperture of linear dimension D, the Fraunhofer approximation will be valid provided

$$z > \frac{2D^2}{\lambda} \tag{4-27}$$

where the inequality is now $>$ rather than $\gg$. However, for this example the distance z is still required to be larger than 2,000 meters. Nonetheless, the required conditions are met in a number of important problems. In addition, Fraunhofer diffraction patterns can be observed at distances much closer than implied by Eq. (4-24) provided the aperture is illuminated by a spherical wave converging toward the observer (see Prob. 4-16), or if a positive lens is properly situated between the observer and the aperture (see Chapter 5).

Finally, it should be noted that, at first glance, there exists no transfer function that can be associated with Fraunhofer diffraction, for the approximation (4-24) has destroyed the space invariance of the diffraction equation (cf. Prob. 2-10). The secondary wavelets with parabolic surfaces (as implied by the Fresnel approximation) no longer shift laterally in the (x, y) plane with the particular (ξ, η) point under consideration. Rather, when the location of the secondary source shifts, the corresponding quadratic surface tilts in the (x, y) plane by an amount that depends on the location of the secondary source. Nonetheless, it should not be forgotten that since Fraunhofer diffraction is only a special case of Fresnel diffraction, the transfer function (4-21) remains valid throughout both the Fresnel and the Fraunhofer regimes. That is, it is always possible to calculate diffracted fields in the Fraunhofer region by retaining the full accuracy of the Fresnel approximation.

4.4 Examples of Fraunhofer Diffraction Patterns

We consider next several examples of Fraunhofer diffraction patterns. For additional examples the reader may consult the problems (see Probs. 4-7 through 4-10).

The results of the preceding section can be applied directly to find the complex field distribution across the Fraunhofer diffraction pattern of any given aperture. However, of ultimate interest, for reasons discussed at the beginning of this chapter, is the intensity rather than the complex field strength. The final descriptions of the specific diffraction patterns considered here will therefore be distributions of intensity.

4.4.1 Rectangular Aperture

Consider first a rectangular aperture with an amplitude transmittance given by

$$t_A(\xi,\eta) = \operatorname{rect}\left(\frac{\xi}{2w_X}\right)\operatorname{rect}\left(\frac{\eta}{2w_Y}\right).$$

The constants w_X and w_Y are the half-widths of the aperture in the ξ and η directions. If the aperture is illuminated by a unit-amplitude, normally incident, monochromatic plane wave, then the field distribution across the aperture is equal to the transmittance function t_A. Thus using Eq. (4-25), the Fraunhofer diffraction pattern is seen to be

$$U(x,y) = \frac{e^{jkz}\, e^{j\frac{k}{2z}(x^2+y^2)}}{j\lambda z}\,\mathcal{F}\{U(\xi,\eta)\}\Big|_{f_X=x/\lambda z,\quad f_Y=y/\lambda z}.$$

Noting that $\mathcal{F}\{U(\xi,\eta)\} = A\,\operatorname{sinc}(2w_X f_X)\,\operatorname{sinc}(2w_Y f_Y)$, where A is the area of the aperture ($A = 4w_X w_Y$), we find

$$U(x,y) = \frac{e^{jkz}\, e^{j\frac{k}{2z}(x^2+y^2)}}{j\lambda z}\,A\,\operatorname{sinc}\left(\frac{2w_X x}{\lambda z}\right)\operatorname{sinc}\left(\frac{2w_Y y}{\lambda z}\right),$$

and

$$I(x,y) = \frac{A^2}{\lambda^2 z^2}\operatorname{sinc}^2\left(\frac{2w_X x}{\lambda z}\right)\operatorname{sinc}^2\left(\frac{2w_Y y}{\lambda z}\right). \tag{4-28}$$

Figure 4.7 shows a cross section of the Fraunhofer intensity pattern along the x axis. Note that the width of the main lobe (i.e., the distance between the first two zeros) is

$$\Delta x = \frac{\lambda z}{w_X}. \tag{4-29}$$

Figure 4.8 shows a photograph of the diffraction pattern produced by a rectangular aperture with a width ratio of $w_X/w_Y = 2$.

4.4.2 Circular Aperture

Consider a diffracting aperture that is circular rather than rectangular, and let the radius of the aperture be w. Thus if q is a radius coordinate in the plane of the aperture, then

$$t_A(q) = \operatorname{circ}\left(\frac{q}{w}\right).$$

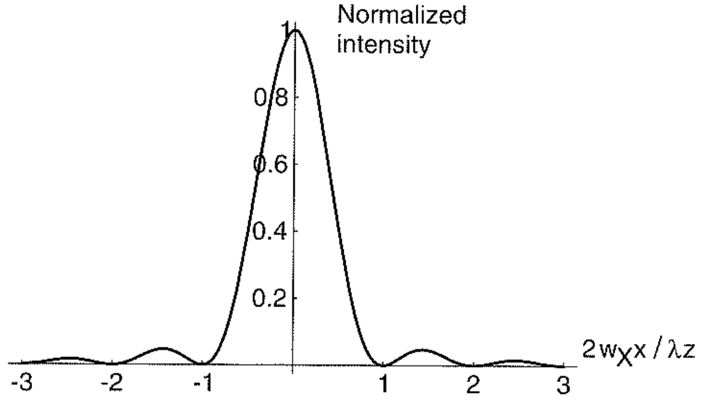

Figure 4.7 Cross section of the Fraunhofer diffraction pattern of a rectangular aperture.

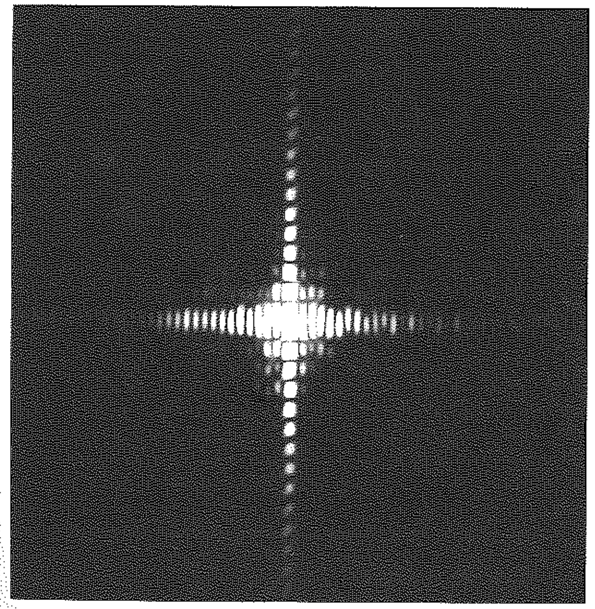

Figure 4.8 The Fraunhofer diffraction pattern of a rectangular aperture ($w_X/w_Y = 2$).

The circular symmetry of the problem suggests that the Fourier transform of Eq. (4-25) be rewritten as a Fourier-Bessel transform. Thus if r is the radius coordinate in the observation plane, we have

$$U(r) = \frac{e^{jkz}}{j\lambda z}\exp\left(j\frac{kr^2}{2z}\right)\mathcal{B}\{U(q)\}\Big|_{\rho=r/\lambda z}, \tag{4-30}$$

Table 4.1: Locations of maxima and minima of the Airy pattern.

x	$\left[2\frac{J_1(\pi x)}{\pi x}\right]^2$	max, min
0	1	max
1.220	0	min
1.635	0.0175	max
2.233	0	min
2.679	0.0042	max
3.238	0	min
3.699	0.0016	max

where $q = \sqrt{\xi^2 + \eta^2}$ represents radius in the aperture plane, and $\rho = \sqrt{f_X^2 + f_Y^2}$ represents radius in the spatial frequency domain. For unit-amplitude, normally incident plane-wave illumination, the field transmitted by the aperture is equal to the amplitude transmittance; in addition,

$$\mathcal{B}\left\{\text{circ}\left(\frac{q}{w}\right)\right\} = A\frac{J_1(2\pi w\rho)}{\pi w\rho},$$

where $A = \pi w^2$. The amplitude distribution in the Fraunhofer diffraction pattern is seen to be

$$U(r) = e^{jkz}e^{j\frac{kr^2}{2z}}\frac{A}{j\lambda z}\left[2\frac{J_1(kwr/z)}{kwr/z}\right],$$

and the intensity distribution can be written

$$I(r) = \left(\frac{A}{\lambda z}\right)^2\left[2\frac{J_1(kwr/z)}{kwr/z}\right]^2. \tag{4-31}$$

This intensity distribution is referred to as the *Airy pattern*, after G.B. Airy who first derived it. Table 4.1 shows the values of the Airy pattern at successive maxima and minima, from which it can be seen that the width of the central lobe, measured along the x or y axis, is given by

$$d = 1.22\frac{\lambda z}{w}. \tag{4-32}$$

Figure 4.9 shows a cross section of the Airy pattern, while Fig. 4.10 is a photograph of the Fraunhofer diffraction pattern of a circular aperture.

4.4.3 Thin Sinusoidal Amplitude Grating

In the previous examples, diffraction was assumed to be caused by apertures in infinite opaque screens. In practice, diffracting objects can be far more complex. In accord with

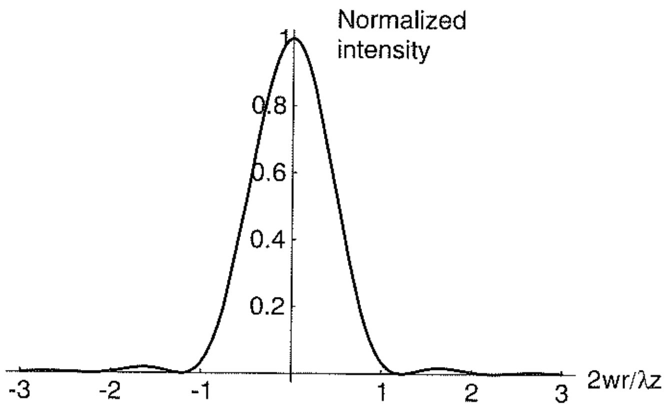

Figure 4.9 Cross section of the Fraunhofer diffraction pattern of a circular aperture.

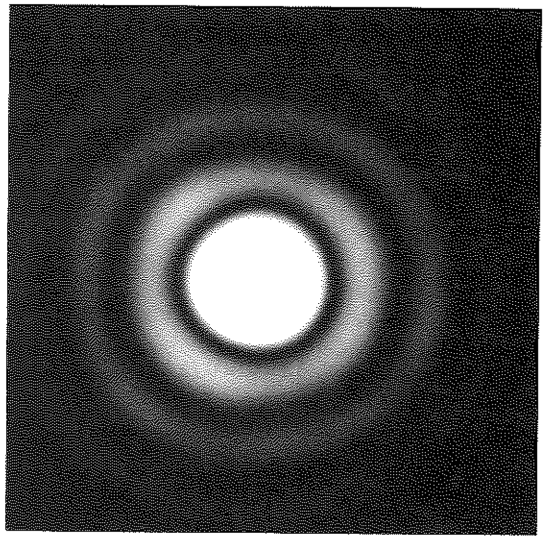

Figure 4.10 Fraunhofer diffraction pattern of a circular aperture.

our earlier definition (3-70), the amplitude transmittance $t_A(\xi, \eta)$ of a screen is defined as the ratio of the complex field amplitude immediately behind the screen to the complex amplitude incident on the screen. Until now, our examples have involved only transmittance functions of the form

$$t_A(\xi, \eta) = \begin{cases} 1 & \text{in the aperture} \\ 0 & \text{outside the aperture.} \end{cases}$$

It is possible, however, to introduce a prescribed amplitude transmittance function within a given aperture. Spatial attenuation can be introduced with, for example, an absorbing photographic transparency, thus allowing real values of t_A between zero and unity to be realized. Spatial patterns of phase shift can be introduced by means of transparent plates of varying thickness, thus extending the realizable values of t_A to all points within or on the unit circle in the complex plane.

As an example of this more general type of diffracting screen, consider a *thin sinusoidal amplitude grating* defined by the amplitude transmittance function

$$t_A(\xi,\eta) = \left[\frac{1}{2} + \frac{m}{2}\cos(2\pi f_0 \xi)\right] \operatorname{rect}\left(\frac{\xi}{2w}\right) \operatorname{rect}\left(\frac{\eta}{2w}\right) \tag{4-33}$$

where for simplicity we have assumed that the grating structure is bounded by a square aperture of width $2w$. The parameter m represents the peak-to-peak change of amplitude transmittance across the screen, and f_0 is the spatial frequency of the grating. The term *thin* in this context means that the structure can indeed be represented by a simple amplitude transmittance. Structures that are not sufficiently thin cannot be so represented, a point we shall return to in a later chapter. Figure 4.11 shows a cross section of the grating amplitude transmittance function.

If the screen is normally illuminated by a unit-amplitude plane wave, the field distribution across the aperture is equal simply to t_A. To find the Fraunhofer diffraction pattern, we first Fourier transform that field distribution. Noting that

$$\begin{aligned}\mathcal{F}\left\{\frac{1}{2} + \frac{m}{2}\cos(2\pi f_0\xi)\right\} &= \frac{1}{2}\delta(f_X, f_Y) \\ &\quad + \frac{m}{4}\delta(f_X + f_0, f_Y) + \frac{m}{4}\delta(f_X - f_0, f_Y)\end{aligned} \tag{4-34}$$

and

$$\mathcal{F}\left\{\operatorname{rect}\left(\frac{\xi}{2w}\right)\operatorname{rect}\left(\frac{\eta}{2w}\right)\right\} = A \operatorname{sinc}(2wf_X)\operatorname{sinc}(2wf_Y),$$

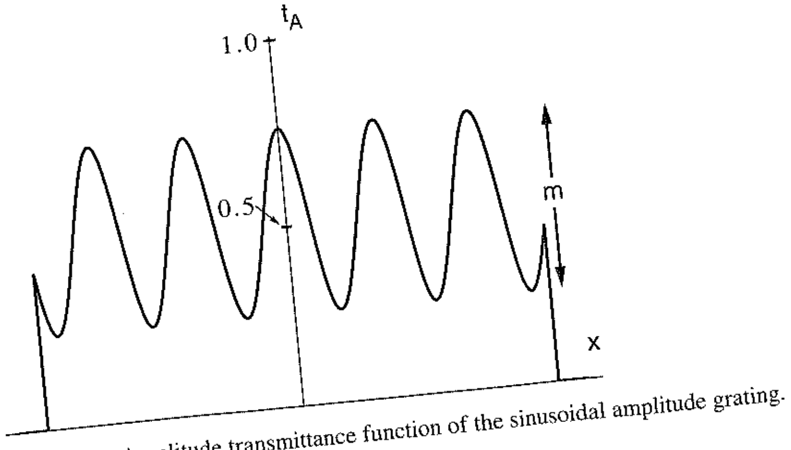

Figure 4.11 Amplitude transmittance function of the sinusoidal amplitude grating.

the convolution theorem can be used to write

$$\begin{aligned}\mathcal{F}\{U(\xi,\eta)\} = \frac{A}{2}\operatorname{sinc}(2wf_Y)\Big\{&\operatorname{sinc}(2wf_X) + \frac{m}{2}\operatorname{sinc}\left[2w(f_X + f_0)\right] \\ &+ \frac{m}{2}\operatorname{sinc}\left[2w(f_X - f_0)\right]\Big\},\end{aligned}$$

where A signifies the area of the aperture bounding the grating. The Fraunhofer diffraction pattern can now be written

$$\begin{aligned}U(x,y) = \frac{A}{j2\lambda z} e^{jkz} e^{j\frac{k}{2z}(x^2+y^2)} \operatorname{sinc}\left(\frac{2wy}{\lambda z}\right)\Big\{&\operatorname{sinc}\left(\frac{2wx}{\lambda z}\right) \\ &+ \frac{m}{2}\operatorname{sinc}\left[\frac{2w}{\lambda z}(x + f_0\lambda z)\right] + \frac{m}{2}\operatorname{sinc}\left[\frac{2w}{\lambda z}(x - f_0\lambda z)\right]\Big\}.\end{aligned} \tag{4-35}$$

Finally, the corresponding intensity distribution is found by taking the squared magnitude of Eq. (4-35). Note that if there are many grating periods within the aperture, then $f_0 \gg 1/w$, and there will be negligible overlap of the three sinc functions, allowing the intensity to be calculated as the sum of the squared magnitudes of the three terms in (4-35). The intensity is then given by

$$\begin{aligned}I(x,y) \approx \left[\frac{A}{2\lambda z}\right]^2 \operatorname{sinc}^2\left(\frac{2wy}{\lambda z}\right)\Big\{&\operatorname{sinc}^2\left(\frac{2wx}{\lambda z}\right) \\ &+ \frac{m^2}{4}\operatorname{sinc}^2\left[\frac{2w}{\lambda z}(x + f_0\lambda z)\right] + \frac{m^2}{4}\operatorname{sinc}^2\left[\frac{2w}{\lambda z}(x - f_0\lambda z)\right]\Big\}.\end{aligned} \tag{4-36}$$

This intensity pattern is illustrated in Fig. 4.12. Note that some of the incident light is absorbed by the grating, and in addition the sinusoidal transmittance variation across

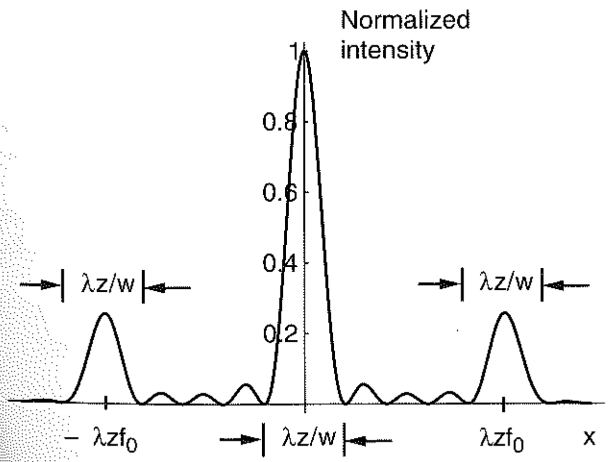

Figure 4.12 Fraunhofer diffraction pattern for a thin sinusoidal amplitude grating.

the aperture has deflected some of the energy out of the central diffraction pattern into two additional side patterns. The central diffraction pattern is called the *zero order* of the Fraunhofer pattern, while the two side patterns are called the *first orders*. The spatial separation of the first orders from the zero order is $f_0\lambda z$, while the width of the main lobe of all orders is $\lambda z/w$.

Another quantity of some practical interest in both holography and optical information processing is the *diffraction efficiency* of the grating. The diffraction efficiency is defined as the fraction of the incident optical power that appears in a single diffraction order (usually the +1 order) of the grating. The diffraction efficiency for the grating of interest can be deduced from Eq. (4-34). The fraction of power appearing in each diffraction order can be found by squaring the coefficients of the delta functions in this representation, for it is the delta functions that determine the power in each order, not the sinc functions that simply spread these impulses. From this equation we conclude that the diffraction efficiencies $\eta_0, \eta_{+1}, \eta_{-1}$ associated with the three diffraction orders are given by

$$\begin{aligned}\eta_0 &= 0.25\\ \eta_{+1} &= m^2/16\\ \eta_{-1} &= m^2/16.\end{aligned} \tag{4-37}$$

Thus a single first diffraction order carries at most $1/16 = 6.25\%$ of the incident power, a rather small fraction. If the efficiencies of the three orders are added up, it will be seen that only $1/4 + m^2/8$ of the total is accounted for. The rest is lost through absorption by the grating.

For a further discussion of gratings and their orders, see Appendix D.

4.4.4 Thin Sinusoidal Phase Grating

As a final example of Fraunhofer diffraction calculations, consider a *thin sinusoidal phase grating* defined by the amplitude transmittance function

$$t_A(\xi, \eta) = \exp\left[j\frac{m}{2}\sin(2\pi f_0\xi)\right] \operatorname{rect}\left(\frac{\xi}{2w}\right) \operatorname{rect}\left(\frac{\eta}{2w}\right) \tag{4-38}$$

where, by proper choice of phase reference, we have dropped a factor representing the average phase delay through the grating. The parameter m represents the peak-to-peak excursion of the phase delay.

If the grating is illuminated by a unit-amplitude, normally incident plane wave, then the field distribution immediately behind the screen is given precisely by Eq. (4-38). The analysis is simplified by use of the identity

$$\exp\left[j\frac{m}{2}\sin(2\pi f_0\xi)\right] = \sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right)\exp(j2\pi q f_0\xi)$$

where J_q is a Bessel function of the first kind, order q. Thus

$$\mathcal{F}\left\{\exp\left[j\frac{m}{2}\sin(2\pi f_0\xi)\right]\right\} = \sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right)\delta(f_X - qf_0, f_Y) \tag{4-39}$$

and

$$\begin{aligned}\mathcal{F}\{U(\xi, \eta)\} &= \mathcal{F}\{t_A(\xi, \eta)\}\\ &= [A\operatorname{sinc}(2wf_X)\operatorname{sinc}(2wf_Y)] \otimes \left[\sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right)\delta(f_X - qf_0, f_Y)\right]\\ &= \sum_{q=-\infty}^{\infty} A J_q\left(\frac{m}{2}\right)\operatorname{sinc}[2w(f_X - qf_0)]\operatorname{sinc}(2wf_Y).\end{aligned}$$

Thus the field strength in the Fraunhofer diffraction pattern can be written

$$U(x, y) = \frac{A}{j\lambda z} e^{jkz} e^{j\frac{k}{2z}(x^2+y^2)} \sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right)\operatorname{sinc}\left[\frac{2w}{\lambda z}(x - qf_0\lambda z)\right]\operatorname{sinc}\left(\frac{2wy}{\lambda z}\right). \tag{4-40}$$

If we again assume that there are many periods of the grating within the bounding aperture ($f_0 \gg 1/w$), there is negligible overlap of the various diffracted terms, and the corresponding intensity pattern becomes

$$I(x, y) \approx \left(\frac{A}{\lambda z}\right)^2 \sum_{q=-\infty}^{\infty} J_q^2\left(\frac{m}{2}\right)\operatorname{sinc}^2\left[\frac{2w}{\lambda z}(x - qf_0\lambda z)\right]\operatorname{sinc}^2\left(\frac{2wy}{\lambda z}\right). \tag{4-41}$$

The introduction of the sinusoidal phase grating has thus deflected energy out of the zero order into a multitude of higher orders. The peak intensity of the qth order is $[A\,J_q(m/2)/\lambda z]^2$, while the displacement of that order from the center of the diffraction pattern is $qf_0\lambda z$. Figure 4.13 shows a cross section of the intensity pattern when the peak-to-peak phase delay m is 8 radians. Note that the strengths of the various orders are symmetric about the zero order.

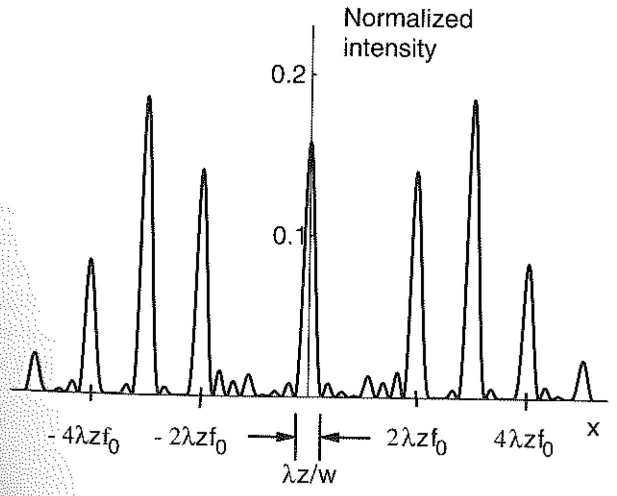

Figure 4.13 Fraunhofer diffraction pattern for a thin sinusoidal phase grating. The ±1 orders have nearly vanished in this example.