

Basic Diffraction

T. Milster

College of Optical Sciences, University of Arizona

5.1 Introduction To Diffraction

Optical systems typically have limiting apertures through which the light beam must pass. Geometrically, the light beam is simply truncated by the aperture. In reality, interactions of the light beam and edges of the aperture produce disturbances in the transmitted beam and in the geometrical shadow. These effects occur with all types of illumination, but they are most striking when the aperture is illuminated with a long-coherence-length laser.

Diffraction theory is the study of light propagation effects that are not predicted by geometrical (ray-trace) models. For example, the light distribution behind a laser-illuminated hole in an otherwise opaque screen contains annular variations in irradiance, which are called Fresnel rings, that cannot be calculated geometrically.

The basic problem to be addressed in this chapter is illustrated in Fig. 5.1. An aperture is illuminated by a known electric field distribution. Our task is to determine the electric field amplitude and phase at some observation position r_0 behind the aperture. To solve the problem, we need to know:

- 1) Properties of the illumination field $U_s(\mathbf{r})$;
- 2) Properties of the aperture (transmission amplitude and phase); and
- 3) Location of the observation.

One approach to solving this problem is to assume that the illuminating wave in the aperture can be described by a collection of secondary sources, as suggested by Huygens [6.1]. The field at the observation point is found by integrating contributions from each source after propagation, as illustrated in Fig. 5.2. This basic concept is similar to a multiple-beam interference pattern, where the number of beams approaches infinity. A *free-space point spread function* (PSF) can be defined that describes the contribution from each secondary source over the obser-

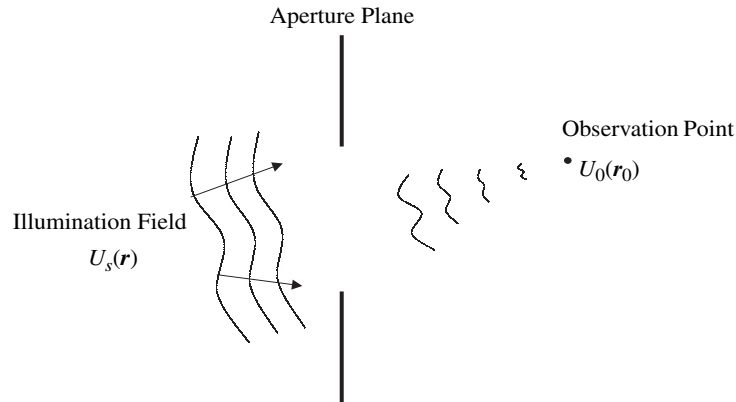


Fig. 5.1. Opaque aperture illuminated by a known electric field distribution. Observation point \mathbf{r}_0 is to the right of the aperture plane.

vation surface. Depending on the distance of the aperture from the observation, the size of the aperture and the wavelength, the free-space PSF can be approximated by functions that are straightforward to evaluate. The total diffraction pattern is found by integrating contributions from all secondary sources in the aperture. This approach yields intuition about diffraction phenomena, and it yields simple relationships that can be used to estimate diffraction patterns without significant computation.

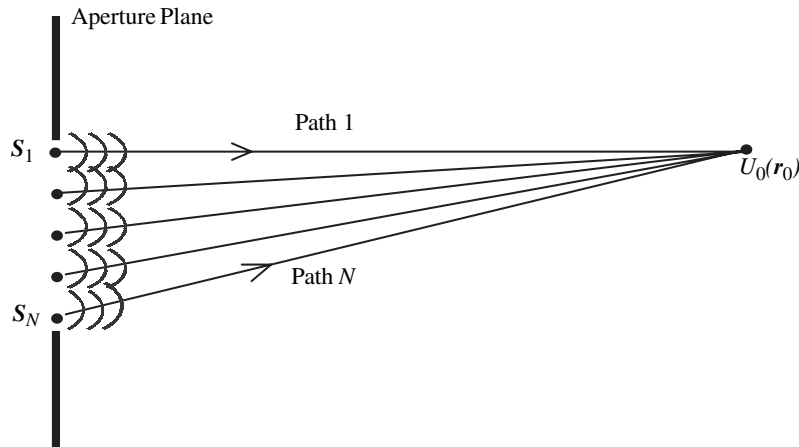


Fig. 5.2. Huygens wavelet construction. The field distribution at the aperture is composed of an infinite number of secondary sources from which light propagates and interferes at the observation point.

Fresnel diffraction is one of the free-space PSF approximations, where the observation plane is relatively close to the aperture. This situation is often encoun-

tered in laser engineering, where a laser beam is passed through an aperture and the transmitted light illuminates a target or detector. In addition, this chapter discusses *Fresnel zone plates*, which are lenses that use diffraction, rather than refraction, to form images.

Fraunhofer diffraction is a free-space PSF approximation that is valid when the separation between the aperture and the observation plane is very large. It can also be applied at the image plane of a lens. In Fraunhofer diffraction, the distribution of light in the aperture plane is related to the distribution of light in the observation plane by a simple Fourier transform.

An alternate approach is to consider diffraction as a linear operation and define a *transfer function* that describes propagation from the aperture to the observation plane. The transfer function is defined by taking the Fourier transform of the free-space PSF described above. The transfer-function approach leads to a very useful philosophical interpretation of the diffraction process in terms of the *plane-wave spectrum*. The *Talbot effect* is an application of the plane-wave spectrum for the re-imaging of periodic patterns without lenses.

Babinet's Principle is a result of the assumption that the diffraction problem is a linear operation. This principle is straightforward, but it is often misunderstood and misused. It can be used to simplify complicated diffraction patterns into smaller, well defined problems that can be easily solved. Then, the problem segments are recombined in the proper way in order to provide a solution to the larger problem.

A *diffraction grating* is an important optical element that can be used to measure the power spectrum of a source, provide feedback for control of spot position in an optical disk, and is useful in many other applications. It consists of many straight, parallel lines or slits from which the reflected or transmitted light has particular properties.

These and other interesting physical optics phenomena are discussed in the following sections. In Section 5.2, a basic mathematical background is developed for diffraction theory. Section 5.3 presents a detailed investigation of Fresnel diffraction, and Section 5.4 discusses Fraunhofer diffraction. Section 5.5 discusses the theory of gratings, and Section 5.6 provides a pictorial atlas of diffraction patterns in the near field.

5.2 Mathematical theory of diffraction

This extensive section contains a detailed scalar mathematical development for the theory of diffraction. The serious student of optics should appreciate these concepts, because they provide the basis for understanding many phenomena observed with laser systems. We will assume that the illuminating wavefront is an ideal monochromatic wave.

5.2.1 Overview

There are two mathematical approaches that equally well describe diffraction, as shown in Fig. 5.3. The free-space point spread function approach calculates a spatial description of the electric field at the observation plane resulting from each point in the aperture. Then, the total field in the observation plane is calculated by integrating individual responses. Depending on the distance of the observation plane from the aperture, different approximations can be used for the free-space PSF, which is in the form of wavelets arriving at the observation point. The wavelets are a complete Huygens wavelet, spherical wavelet, parabolic wavelet, or a planar wavelet. Alternatively, the transfer function approach utilizes the Fourier transform of the electric field at the aperture. This transform multiplied by the free-space transfer function is the Fourier transform of the electric field at the observation plane. The Huygens wavelet approach and the transfer function approach are both rigorous and without significant approximation. They are related by Weyl's integral, which is studied in Section 5.2.7.

5.2.2 Integral Theorem of Helmholtz and Kirchhoff

Our method for solving the diffraction problem is based on a well-known technique for solving inhomogeneous differential equations subject to boundary conditions. We start with *Green's Theorem*, which follows from the *Divergence Theorem* as described by Jackson [6.3]:

$$\iiint_V (G \nabla^2 U - U \nabla^2 G) dv = \iint_S \left(G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n} \right) ds. \quad (5.1)$$

The left hand side of Eq. (5.1) is an integral over closed volume V , as shown in Fig. 5.4. The right hand side is a surface integral surrounding the volume over surface S . Arbitrary scalar functions U and G , along with their first and second partial derivatives, are single valued and continuous scalar fields within and on S . U and G are both functions of position. By relating the volume integral to the surface integral, we should be able to solve for the scalar functions inside the volume V by knowing the properties of the functions on the surface. Other definitions are

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \quad (5.2)$$

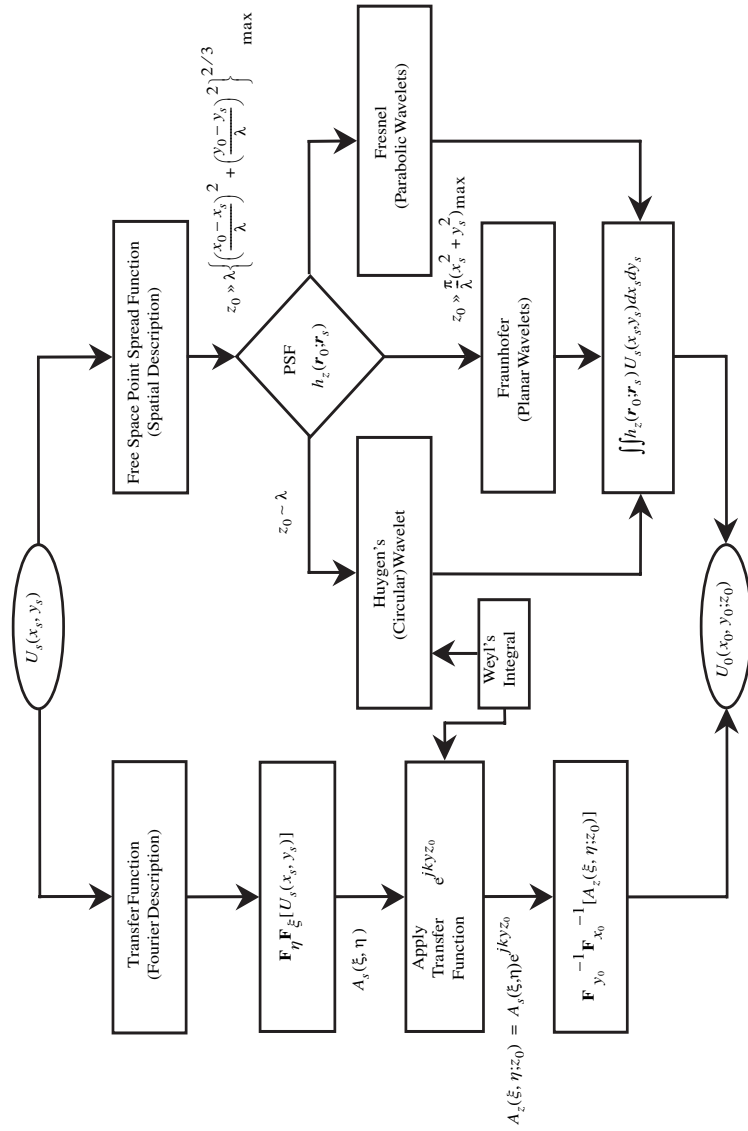


Fig. 5.3. Schematic description of free space point spread function and transfer function approaches to diffraction. The two approaches are related by Weyl's integral.

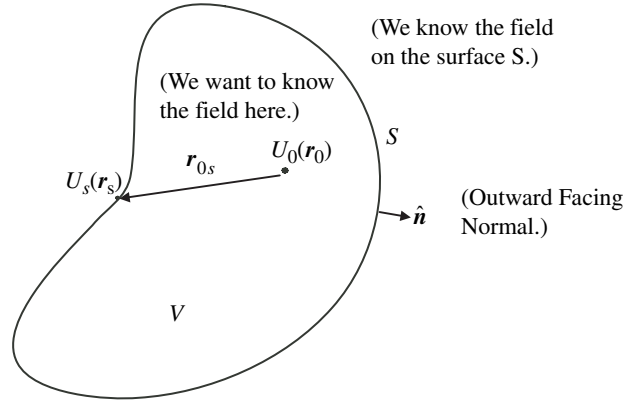


Fig. 5.4. Geometric illustration for Helmholtz–Kirchhoff integral theorem. The observation point is enclosed within a finite volume V with a surface area S .

and

$$\frac{\partial}{\partial n} = \hat{\mathbf{n}} \cdot \nabla, \quad (5.3)$$

where $\hat{\mathbf{n}}$ is the outward facing normal to surface S .

In applying Eq. (5.1) to solve linear diffraction problems, the differential equation of interest is the scalar Helmholtz equation for the harmonic electric field $Ue^{-j\omega t}$, such that

$$(\nabla^2 + k^2)U = 0, \quad (5.4)$$

where $k = 2\pi n/\lambda$, λ is the vacuum wavelength and n is the real refractive index of the homogeneous and non-absorbing medium in which the calculation takes place.¹ Equation (5.4) assumes that there are no sources inside volume V .

We are free to choose scalar function G for application in Eq. (5.1). However, a clever choice is to constrain G so that

$$(\nabla^2 + k^2)G = \delta(\|\mathbf{r} - \mathbf{r}_0\|), \quad (5.5)$$

where $\delta(\|\mathbf{r} - \mathbf{r}_0\|)$ is the Dirac delta function. In this form, G is the *Green's function* that is the solution to the Helmholtz equation for a unit impulse source at observation point $\mathbf{r}_0 = x_0\hat{\mathbf{x}} + y_0\hat{\mathbf{y}} + z_0\hat{\mathbf{z}}$. Notice that G is a function of both the

1. Take care not to confuse the term *homogeneous* that is associated with both a type of differential equation and a material property. Although the term is the same, the meaning is completely different.

location $\mathbf{r}_s = x_s\hat{\mathbf{x}} + y_s\hat{\mathbf{y}} + z_s\hat{\mathbf{z}}$ on the surface and observation point when applied to the surface integral.^{1,2}

Substitution of Eqs. (5.4) and Eq. (5.5) into Eq. (5.1) yields for the volume integral

$$\begin{aligned} \iiint_V (G\nabla^2 U - U\nabla^2 G) d\mathbf{v} &= \iiint_V \left\{ G[-k^2 U] - U[\delta(\|\mathbf{r} - \mathbf{r}_0\|) - k^2 G] \right\} d\mathbf{v} \\ &= \iiint_V \left\{ -k^2 GU + k^2 GU - U\delta(\|\mathbf{r} - \mathbf{r}_0\|) - k^2 G \right\} d\mathbf{v} \\ &= -U(\mathbf{r}_0). \end{aligned}$$

Therefore,

$$U_0(\mathbf{r}_0) = \iint_S \left[U_s(\mathbf{r}_s) \frac{\partial G(\mathbf{r}_0; \mathbf{r}_s)}{\partial n} - G(\mathbf{r}_0; \mathbf{r}_s) \frac{\partial U_s(\mathbf{r}_s)}{\partial n} \right] ds, \quad (5.6)$$

where explicit spatial dependencies are now associated with U and G . A simple choice for a Green's function that satisfies Eq. (5.5) is

$$G(\mathbf{r}_0; \mathbf{r}_s) = -\frac{e^{jkr_{0s}}}{4\pi r_{0s}}, \quad (5.7)$$

where $r_{0s} = \sqrt{(x_0 - x_s)^2 + (y_0 - y_s)^2 + (z_0 - z_s)^2}$.

Equation (5.7) is simply an expanding spherical wave centered on the observation point. Substitution of Eq. (5.7) into Eq. (5.6) yields

$$U_0(\mathbf{r}_0) = \frac{1}{4\pi} \iint_S \left[\frac{e^{jkr_{0s}}}{r_{0s}} \frac{\partial U_s(\mathbf{r}_s)}{\partial n} - U_s(\mathbf{r}_s) \frac{\partial}{\partial n} \left(\frac{e^{jkr_{0s}}}{r_{0s}} \right) \right] ds. \quad (5.8)$$

If the function U and its normal derivative are known over surface S , and if the normal derivative of G can be calculated over surface, the field $U_0(\mathbf{r}_0)$ inside V can be determined from integration. Equation (5.8) is known as the *Integral Theorem of Helmholtz and Kirchhoff*. For particular surfaces, it may be possible to

1. See [Barrett and Myers, 2004, pp. 467-476] for a more complete description of Green's functions.

2. Throughout this chapter, the diffracting surface S is labeled with s subscripts on the variables.

design G or its normal derivative to be zero over surface S , which can simplify the calculation. An example of specifying $G = 0$ on S is given in the next section.

5.2.3 Diffraction by a plane screen

Consider the geometry of Fig. 5.5, where an open aperture in an otherwise opaque screen is illuminated from the left. Media on both sides of the aperture are air with refractive index of $n = 1$. Our goal is to calculate the field amplitude and phase at point \mathbf{r}_0 . The surface S is divided into two parts. S_1 is a flat surface on the right side of and in contact with the aperture plane. S_2 is a large partial sphere centered on \mathbf{r}_0 and connecting with S_1 at the aperture plane.

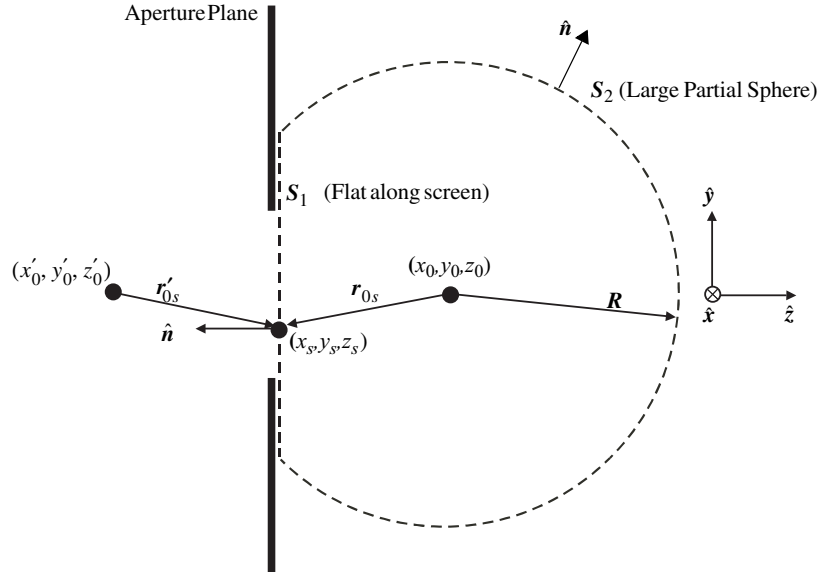


Fig. 5.5. Geometry for diffraction by a planar aperture. The aperture S is illuminated by a known field from the left. The observation point is to the right of the aperture, and $\hat{\mathbf{n}} = -\hat{\mathbf{z}}$.

The Green's function we choose is designed to have zero value on S_1 , thus simplifying Eq. (5.6), where

$$G(\mathbf{r}_0; \mathbf{r}_s) = -\frac{1}{4\pi} \left(\frac{e^{jkr_{0s}}}{r_{0s}} - \frac{e^{jkr'_{0s}}}{r'_{0s}} \right), \quad (5.9)$$

$$\text{with } r'_{0s} = \sqrt{(x'_0 - x_s)^2 + (y'_0 - y_s)^2 + (z'_0 - z_s)^2} = r_{0s},$$

has the normal derivative

$$\frac{\partial G(\mathbf{r}_0; \mathbf{r}_s)}{\partial n} = -2 \frac{\partial}{\partial z_s} \left(-\frac{1}{4\pi} \frac{e^{jkr_{0s}}}{r_{0s}} \right), \quad (5.10)$$

because the outward surface normal $\hat{\mathbf{n}} = -\hat{\mathbf{z}}$. The second expanding spherical wave at $\mathbf{r}'_0 = x'_0 \hat{\mathbf{x}} + y'_0 \hat{\mathbf{y}} + z'_0 \hat{\mathbf{z}}$ forces a zero value of $G(\mathbf{r}_0; \mathbf{r}_s)$ on S_1 . Since it is outside the volume V , it does not contribute to the volume integral in Eq. (5.1). Evaluation of Eq. (5.10) is shown below. The z axis origin is taken at plane S_1 .

$$\begin{aligned} \left. \frac{\partial}{\partial z_s} \frac{e^{jkr_{0s}}}{2\pi r_{0s}} \right|_{z_s=0} &= \left(\frac{\partial}{\partial z_s} e^{jkr_{0s}} \right) \frac{1}{2\pi r_{0s}} + \left(\frac{\partial}{\partial z_s} \frac{1}{2\pi r_{0s}} \right) e^{jkr_{0s}} \Big|_{z_s=0} \\ &= \left(-jk + \frac{1}{r_{0s}} \right) \frac{z_0}{r_{0s}} \frac{e^{jkr_{0s}}}{2\pi r_{0s}} \\ &= \left(-\frac{j}{\lambda} + \frac{1}{2\pi r_{0s}} \right) \gamma_z \frac{e^{jkr_{0s}}}{r_{0s}}, \end{aligned} \quad (5.11)$$

where $\gamma_z = z_0/r_{0s} = (\hat{\mathbf{n}} \cdot \mathbf{r}_{0s})$. On the large partial sphere (surface S_2) with radius vector \mathbf{R} ,

$$G(\mathbf{r}_0; \mathbf{r}_s) = -\frac{1}{4\pi} \left(\frac{e^{jkR}}{R} - \frac{e^{jkR'}}{R'} \right), \quad (5.12)$$

where

$$R' = \|\mathbf{R} + \mathbf{r}_0 - \mathbf{r}'_0\|, \quad (5.13)$$

Evaluation of Eq. (5.6) is made in two parts, one over surface S_1 and one over surface S_2 . We start with evaluation of the Green's function and its normal derivative over S_2 . With R large, the first term of the Green's function has the normal derivative

$$\frac{\partial}{\partial n} \left(-\frac{1}{4\pi} \frac{e^{jkR}}{R} \right) = -\frac{jk}{4\pi} \frac{e^{jkR}}{R}. \quad (5.14)$$

The effect on integration over S_2 in Eq. (5.6) from the first term of the Green's function is

$$\begin{aligned}
& \iint_{S_2} \left[-\frac{jk}{4\pi R} e^{jkR} U_s(\mathbf{r}_s) + \frac{e^{jkR}}{4\pi R} \frac{\partial U_s(\mathbf{r}_s)}{\partial n} \right] ds \\
&= \frac{1}{4\pi} \iint_{S_2} \left[-jk U_s(\mathbf{r}_s) + \frac{\partial U_s(\mathbf{r}_s)}{\partial n} \right] \frac{e^{jkR}}{R} ds \\
&= \frac{1}{4\pi} \int_{\Omega} \left(\frac{\partial U_s(\mathbf{r}_s)}{\partial n} - jk U_s(\mathbf{r}_s) \right) e^{jkR} R d\omega,
\end{aligned} \tag{5.15}$$

where the last step in Eq. (5.15) is simply a transformation to spherical coordinates and integration over solid angle Ω , which is the angular subtense of S_2 . If it can be shown that Eq. (5.15) vanishes as $R \rightarrow \infty$, the calculation of Eq. (5.6) is greatly simplified. If the field U behaves as a spherical wave in the limit of large R ,

$$U_s(\mathbf{r}_s) \approx \frac{e^{jkR}}{R}, \tag{5.16}$$

and

$$\begin{aligned}
& \lim_{R \rightarrow \infty} \left\{ \left[\frac{\partial U_s(\mathbf{r}_s)}{\partial n} - jk U_s(\mathbf{r}_s) \right] e^{jkR} R \right\} \\
&= \lim_{R \rightarrow \infty} \left\{ \left[\left(jk - \frac{1}{R} \right) U_s(\mathbf{r}_s) - jk U_s(\mathbf{r}_s) \right] e^{jkR} R \right\} \\
&= \lim_{R \rightarrow \infty} \left\{ \left[-\frac{1}{R} U_s(\mathbf{r}_s) \right] e^{jkR} R \right\} \\
&= \lim_{R \rightarrow \infty} [-U_s(\mathbf{r}_s) e^{jkR}].
\end{aligned} \tag{5.17}$$

Equation (5.17) indicates that, if the field $U_s(\mathbf{r}_s)$ vanishes at least as fast as a spherical wave, the contribution to the integral over S_2 from the first term in the Green's function is zero. Since the second term in the Green's function behaves in a similar manner, the contribution from it is also zero. The assumption that U falls off as $R \rightarrow \infty$ at least as fast as a spherical wave is called the *Sommerfeld Radiation Condition*,¹ and it allows significant simplification to Eq. (5.6). Combination of the Sommerfeld Radiation Condition, the fact that the value of Eq. (5.9) is a zero on S_1 and Eq. (5.11) yields a simplified form of Eq. (5.6) as

1. A more complete justification of this radiation condition can be found in [Born and Wolf, 1980, p. 379].

$$U_0(\mathbf{r}_0) = \iint_{S_1} \left(-\frac{j}{\lambda} + \frac{1}{2\pi r_{0s}} \right) \gamma_z \frac{e^{jkr_{0s}}}{r_{0s}} U_s(\mathbf{r}_s) ds, \quad (5.18)$$

Evaluation of Eq. (5.18) requires only that the field $U_s(\mathbf{r}_s)$ is known on S_1 . If the boundary conditions are such that the field $U_s(\mathbf{r}_s)$ is undisturbed by the aperture and $U_s(\mathbf{r}_s)$ is zero in the shadow of the aperture plane,

$$U_0(\mathbf{r}_0) = \iint_{ap} \left(-\frac{j}{\lambda} + \frac{1}{2\pi r_{0s}} \right) \gamma_z \frac{e^{jkr_{0s}}}{r_{0s}} U_s(\mathbf{r}_s) ds, \quad (5.19)$$

where limits ap represent the open part of the aperture. If the observation distance $r_{0s} \gg \lambda$, Eq. (5.19) simplifies to

$$U_0(\mathbf{r}_0) = -\frac{j}{\lambda} \iint_{ap} \gamma_z \frac{e^{jkr_{0s}}}{r_{0s}} U_s(\mathbf{r}_s) ds. \quad (5.20)$$

Equations (5.19) and (5.20) are forms of the *Rayleigh-Sommerfeld diffraction formula*, and they are the basis for much of scalar diffraction theory.

Notice that Eqs. (5.19) and (5.20) can be put into the form

$$U_0(\mathbf{r}_0) = \iint_{ap} W_z(\mathbf{r}_0; \mathbf{r}_s) U_s(\mathbf{r}_s) \frac{e^{jkr_{0s}}}{r_{0s}} ds, \quad (5.21)$$

where, for Eq. (5.20),

$$W_z(\mathbf{r}_0; \mathbf{r}_s) = -\frac{j}{\lambda} \gamma_z, \quad (5.22)$$

is the weighting factor. Equation (5.21) is a formal statement of the *Huygens Principle*, which states that the field at the observation point is due to a weighted summation of spherical waves multiplied by the incident field. In Huygens's original work, the weighting factor $W_z(\mathbf{r}_0; \mathbf{r}_s)$ was not considered. [Huygens, 1690]

As developed in Eqs. (5.19) through (5.21), the boundary conditions where U is undisturbed in the aperture and $U = 0$ in the shadow of the aperture plane are commonly called *Dirichlet Boundary Conditions*. A clever choice of the Green's function in Eq. (5.9) allows calculation of the field at the observation point without approximation. There are other choices for boundary conditions and associated Green's functions that lead to different forms of the diffraction integral. The Dirichlet, Neuman and Cauchy (also called Kirchhoff) boundary conditions are

shown in Table 5.1.¹ A detailed discussion of these boundary conditions is given by Goodman. [Goodman, 1968] They differ primarily in the angular dependence of the weighting factor, which is γ_z in Eqs. (5.19) and (5.20). This angular dependence is called the *obliquity factor*.

5.2.4 Derivation of a Huygens Wavelet

In this section, a form of the free-space point spread function is derived that is called the *Huygens wavelet*. We start by rewriting Eq. (5.19) in the form of a superposition integral

$$U_0(\mathbf{r}_0) = \int \int_{ap} U_s(\mathbf{r}_s) h_z^H(\mathbf{r}_0; \mathbf{r}_s) ds, \quad (5.25)$$

where $U_s(\mathbf{r}_s)$ is the total field at the aperture plane and $h_z^H(\mathbf{r}_0; \mathbf{r}_s)$ is a point spread function given by

$$\begin{aligned} h_z^H(\mathbf{r}_0; \mathbf{r}_s) &= \frac{\partial}{\partial z_s} \frac{e^{jk r_{0s}}}{2\pi r_{0s}} \\ &= \left(-\frac{j}{\lambda} + \frac{1}{2\pi r_{0s}} \right) \gamma_z \frac{e^{jk r_{0s}}}{r_{0s}} \\ &= a e^{j\phi} \gamma_z e^{jk r_{0s}}. \end{aligned} \quad (5.26)$$

The constant a and phase ϕ are given by

$$a = \frac{1}{r_{0s}} \left| -\frac{j}{\lambda} + \frac{1}{2\pi r_{0s}} \right| = \frac{1}{r_{0s} \sqrt{\lambda^2 + \frac{1}{(2\pi r_{0s})^2}}} = \frac{1}{2\pi r_{0s}^2} \sqrt{1 + (k r_{0s})^2}, \quad (5.27)$$

and

$$\phi = -\tan^{-1}(k r_{0s}), \quad (5.28)$$

respectively. Combination of Eqs. (5.26) through (5.28) yields

1. The Neumann and Dirichlet boundary conditions are commonly applied to solutions of partial differential equations. The Dirichlet boundary conditions specify the value of a function on a surface, and the Neumann boundary conditions specify the normal derivative of a function on a surface. See [Arfken, 1985, pp. 502-504], for more information.

Table 5.1 Summary of boundary conditions, obliquity factors and diffraction integrals.

Boundary Conditions	$G(r_0; r_s)$	Across the aperture	In Shadow	Obliquity Factor	Diffraction Integral	Comment
Kirchhoff ¹ (Cauchy)	$-\frac{e^{jk_0 r_0}}{4\pi r_{0s}}$	U and $\frac{\partial U}{\partial n}$ same as without screen	$U=0$ $\frac{\partial U}{\partial n}=0$	$\frac{\hat{n} \bullet \hat{r}_{0s} - \hat{n} \bullet \hat{r}_{sc,s}}{2}$	$U_0(r_0) = \frac{jA}{\lambda} \iint_{ap} \frac{e^{jk(r_{sc,s} + r_{0s})}}{r_{sc,s} r_{0s}} \left[\frac{(\hat{n} \bullet \hat{r}_{0s}) - (\hat{n} \bullet \hat{r}_{sc,s})}{2} \right] ds$ (5.23)	Fresnel-Kirchhoff Diffraction Formula
Neumann	$-\frac{e^{jk_0 r_0}}{4\pi r_{0s}} - \frac{e^{jk_0 r_s}}{4\pi r_{0s}}$	$\frac{\partial U}{\partial n}$ same as without screen	$\frac{\partial U}{\partial n}=0$	contained in $\frac{\partial U}{\partial n}$	$U_0(r_0) = \frac{1}{4\pi ap} \iint \frac{\partial U_s(r_s)}{\partial n} \frac{2e^{jk_0 r_0}}{r_{0s}} ds$ (5.24)	
Dirichlet	$-\frac{e^{jk_0 r_0}}{4\pi r_{0s}} + \frac{e^{jk_0 r_s}}{4\pi r_{0s}}$	U same as without screen	$U=0$	$\gamma_z = \hat{n} \bullet \hat{r}_{0s}$	$U_0(r_0) = \iint_{ap} \left(\frac{-j}{\lambda} + \frac{1}{2\pi r_{0s}} \right) \gamma_z \frac{e^{jk_0 r_0}}{r_{0s}} U_s(r_s) ds$ (5.19)	Rayleigh-Sommerfeld Diffraction Formula

¹ Illuminated with a single spherical wave located at $r_{sc,s}$, to the left of the aperture plane. The spherical wave has amplitude A .

$$\begin{aligned}
 h_z^H(\mathbf{r}_0; \mathbf{r}_s) &= \frac{\partial}{\partial z_s} \frac{e^{jk r_{0s}}}{2\pi r_{0s}} \\
 &= \frac{\sqrt{1 + (k r_{0s})^2}}{2\pi r_{0s}^2} \gamma_z e^{j[k r_{0s} - \tan^{-1}(k r_{0s})]},
 \end{aligned} \tag{5.29}$$

which is an expression for the *Huygens wavelet*. Notice that the Huygens wavelet is not exactly a spherical wave. An obliquity factor and a phase factor produce slight differences from a spherical wave for small r_{0s} , as shown in Figs. 5.6 and 5.7. For example, the phase of Eq. (5.29) varies significantly when $r_{0s} < \lambda$. However, as r_{0s} increases, ϕ approaches the constant value $-\pi/2$.

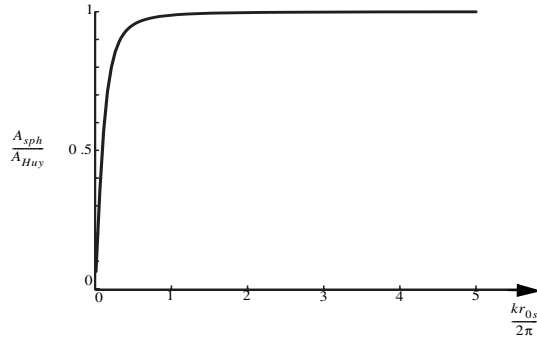


Fig. 5.6. Comparison of a Huygens wavelet and a spherical wave. The amplitude ratio of a spherical wave to a Huygens wave is plotted against wavelength-normalized distance from the source. If r_{0s} is small (if the observation point is near the aperture), the Huygens wave and spherical wave have very different amplitudes. The amplitude ratio varies significantly when $r_{0s} < \lambda$. After a distance of 2λ , the amplitudes are essentially the same.

The Huygens wavelet h_z^H in Eq. (5.29) is the *Huygens free-space point spread function* (PSF). There are no approximations in its derivation, other than those associated with the Dirichlet boundary conditions. Equation (5.29) can be used in Eq. (5.25) to find the field at \mathbf{r}_0 from field distribution $U(\mathbf{r}_s)$ in the aperture. It is useful when describing diffraction at distances close to the aperture. Also, as we will find in Section 5.2.7, the Fourier transform of the Huygens wavelet is the transfer function of free space, which provides a spatial-frequency-domain tool for diffraction analysis.

5.2.5 The Fresnel Approximation

The goal in this section is to derive the Fresnel point spread function $h_z^F(\mathbf{r}_0; \mathbf{r}_s)$, or *Fresnel wavelet*, which is an approximation of the Huygens wavelet

for the case where the distances involved in the diffraction geometry become relatively large. This approximation allows a straightforward calculation of diffraction patterns that is less complicated than using the Huygens wavelet as a free-space point spread function. It also provides the framework for a powerful, intuitive tool called Fresnel zones, which are used when analyzing diffraction from apertures. Fresnel zones are studied in detail in Section 5.3.

We start the derivation of the Fresnel wavelet with Eq. (5.26). In the Fresnel approximation, $r_{0s} \gg \lambda$, so the Huygens wavelet becomes

$$h_z^H(\mathbf{r}_0; \mathbf{r}_s) \approx \frac{-j\gamma_z \exp(jkr_{0s})}{\lambda r_{0s}}. \quad (5.30)$$

Equation (5.25) simplifies to

$$U_0(\mathbf{r}_0) = \frac{-j}{\lambda} \iint_{ap} U_s(\mathbf{r}_s) \gamma_z \frac{\exp(jkr_{0s})}{r_{0s}} ds. \quad (5.31)$$

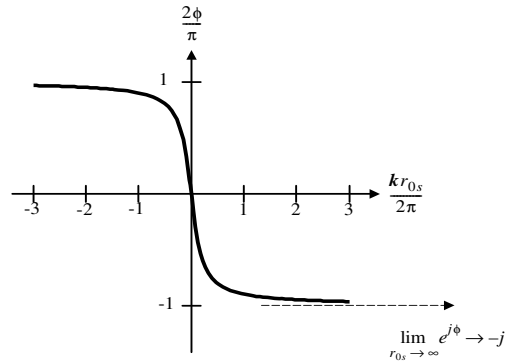


Fig. 5.7. Phase variation of the Huygens wavelet. The function $2\phi/\pi$ is plotted against wavelength-normalized distance from the source. As r_{0s} increases, the phase factor approaches the constant value $-j$. Normally, the observation distance is several wavelengths away from the aperture plane. This phase is also the phase difference between a simple expanding spherical wave and a Huygens wavelet.

With the substitution

$$\gamma_z = \frac{z_0}{r_{0s}}, \quad (5.32)$$

the diffraction formula becomes

$$U_0(\mathbf{r}_0) = \frac{-jz_0}{\lambda} \iint_{ap} U_s(\mathbf{r}_s) \frac{\exp(jkr_{0s})}{r_{0s}^2} ds. \quad (5.33)$$

The next step in the derivation of the Fresnel wavelet is to substitute appropriate approximations for r_{0s} in Eq. (5.33). The binomial expansion

$$\sqrt{1+b} \approx 1 + \frac{b}{2} - \frac{b^2}{8} + \dots \quad (5.34)$$

is used to expand r_{0s} so that

$$\begin{aligned} r_{0s} &= \sqrt{(x_0 - x_s)^2 + (y_0 - y_s)^2 + z_0^2} \\ &\approx z_0 \left\{ 1 + \frac{1}{2} \left[\frac{(x_0 - x_s)^2}{z_0^2} \right] + \frac{1}{2} \left[\frac{(y_0 - y_s)^2}{z_0^2} \right] \right\}, \end{aligned} \quad (5.35)$$

where the coordinates of the source point are $\mathbf{r}_s = (x_s, y_s, 0)$ and coordinates of the observation point are $\mathbf{r}_0 = (x_0, y_0, z_0)$. Equation (5.33) is not significantly affected by small changes in amplitude of the integrand, so the $1/r_{0s}^2$ term can be approximated by

$$\frac{1}{r_{0s}^2} \approx \frac{1}{z_0^2}. \quad (5.36)$$

However, more terms must be used to approximate r_{0s} in the phase of the exponential $\exp(jkr_{0s})$, since $k = 2\pi/\lambda$ is large. Hence,

$$\begin{aligned} \exp(jkr_{0s}) &\equiv \exp \left(jkz_0 \left\{ 1 + \frac{1}{2} \left[\frac{(x_0 - x_s)^2}{z_0^2} \right] + \frac{1}{2} \left[\frac{(y_0 - y_s)^2}{z_0^2} \right] \right\} \right) \\ &= e^{jkz_0} \exp \left\{ \frac{jk}{2z_0} \left[(x_0 - x_s)^2 + (y_0 - y_s)^2 \right] \right\}. \end{aligned} \quad (5.37)$$

With these additional approximations, the *Fresnel diffraction formula* becomes

$$U_0(\mathbf{r}_0) = \frac{-j\mathbf{e}^{jkz_0}}{\lambda z_0} \iint_{ap} U_s(\mathbf{r}_s) \exp\left\{\frac{jk}{2z_0}[(x_0 - x_s)^2 + (y_0 - y_s)^2]\right\} ds. \quad (5.38)$$

where the integration is open over the area of the aperture at $z_s = 0$.

The free-space point spread function in Eq. (5.38) is the *Fresnel wavelet*:

$$h_z^F(\mathbf{r}_0; \mathbf{r}_s) = \frac{-j\mathbf{e}^{jkz_0}}{\lambda z_0} \exp\left\{\frac{jk}{2z_0}[(x_0 - x_s)^2 + (y_0 - y_s)^2]\right\}. \quad (5.39)$$

In Eq. (5.39), the form of the phasefronts arriving at the observation plane are paraboloids. That is, surfaces of constant phase are found by setting the phase argument of the exponential equal to a constant, where

$$\frac{\pi}{\lambda z_0}[(x_0 - x_s)^2 + (y_0 - y_s)^2] = \text{constant} \quad (5.40)$$

describes parabolic surfaces. Furthermore, Eq. (5.39) is shift invariant, and we can write

$$h_z^F(\mathbf{r}_0; \mathbf{r}_s) = h_z^F(x_0 - x_s, y_0 - y_s). \quad (5.41)$$

The phase of the exponential depends only on the difference between coordinates in the aperture and observation planes—not on the individual values.

Equation (5.38) can be written into a form that is very convenient for computation. It is desirable to replace the integration limits over the aperture by limits of integration over all space, so that the equation is amenable to Fourier techniques. This modification is accomplished by replacing the incident field $U_s^-(\mathbf{r}_s)$ by $U_s^+(x_s, y_s)$, where

$$U_s^+(x_s, y_s) \equiv U_s^-(x_s, y_s) t_{ap}(x_s, y_s), \quad (5.42)$$

and $t_{ap}(x_s, y_s)$ is the *amplitude transmittance of the aperture* defined by

$$t_{ap}(x_s, y_s) = \begin{cases} 1, & \text{if } \mathbf{r}_s \text{ in the aperture} \\ 0, & \text{otherwise} \end{cases}. \quad (5.43)$$

When $U_s^+(x_s, y_s)$ is substituted into Eq. (5.40) and the integrand exponential is expanded, a useful and explicit formula is found for the scalar electric field at $\mathbf{r}_0 = (x_0, y_0, z_0)$, which has the limits of integration over the entire aperture plane. That is,

$$\begin{aligned}
U_0(x_0, y_0) &= -\frac{j e^{jkz_0}}{\lambda z_0} \exp\left[j \frac{k}{2z_0}(x_0^2 + y_0^2)\right] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_s^+(x_s, y_s) \exp\left[\frac{jk}{2z_0}(x_s^2 + y_s^2)\right] \\
&\quad \times \exp\left[\frac{-jk}{z_0}(x_0 x_s + y_0 y_s)\right] dx_s dy_s \\
&= -\frac{j e^{jkz_0}}{\lambda z_0} \exp\left[j \frac{k}{2z_0}(x_0^2 + y_0^2)\right] \mathbf{F}_\eta \mathbf{F}_\xi \left\{ U_s^+(x_s, y_s) \exp\left[\frac{jk}{2z_0}(x_s^2 + y_s^2)\right] \right\},
\end{aligned} \tag{5.44}$$

where $\xi = x_0/(\lambda z_0)$, $\eta = y_0/(\lambda z_0)$, and $\mathbf{F}_b \mathbf{F}_a$ is the forward two-dimensional Fourier transform operator taken with variables a and b . (Similarly, $\mathbf{F}_b^{-1} \mathbf{F}_a^{-1}$ is the inverse two-dimensional Fourier transform operator.) ξ and η are spatial frequency variables with units of inverse length. In Eq. (5.44), information concerning the finite aperture is encoded in the field $U_s^+(x_s, y_s)$ via the function $t_{\text{ap}}(x_s, y_s)$.

How valid is the Fresnel approximation? Examination of the binomial expansions in Eqs. (5.34) and (5.35) reveals that the second order term of the form $b^2/8$ must be much less than unity for the approximation to be accurate. If this constraint is expressed with respect to the variables in Eq. (5.38),

$$z_0 \gg \sqrt[3]{\frac{\pi}{4} \lambda \left[\left(\frac{x_0 - x_s}{\lambda} \right)^2 + \left(\frac{y_0 - y_s}{\lambda} \right)^2 \right]_{\text{max}}^{2/3}}. \tag{5.45}$$

When the aperture-observation distance satisfies Eq. (5.45), it is common to say that the observer is in the region of Fresnel diffraction or in the *near field*. Recently, the term *near-field optics* is often used to indicate aperture-observation distances much smaller than those specified by Eq. (5.45), even directly adjacent to and including the aperture.¹

It is often argued that the Fresnel approximation is valid over a larger range than the constraint in Eq. (5.45) implies. The *Principle of Stationary Phase* supports this claim, where the integrand in Eq. (5.38) is examined for certain characteristics. To illustrate the effect, consider a uniformly filled aperture, where $U_s(\mathbf{r}_s) = 1$, and consider a one-dimensional calculation over length L , where the complex exponential is expanded with Euler's identity. That is,

1. Solutions to these types of near-field diffraction problems require a more sophisticated development than is presented in this chapter.

$$\begin{aligned}
U_0(x_0) &= \frac{j e^{jkz_0}}{\lambda z_0} \int_{-\frac{L}{2}}^{\frac{L}{2}} e^{j \frac{k}{2z_0} (x_0 - x_s)^2} dx_s \\
&= \frac{j e^{jkz_0}}{\lambda z_0} \int_{-\frac{L}{2}}^{\frac{L}{2}} \cos \left[\frac{k}{2z_0} (x_0 - x_s)^2 \right] dx_s \\
&\quad + \frac{e^{jkz_0}}{\lambda z_0} \int_{-\frac{L}{2}}^{\frac{L}{2}} \sin \left[\frac{k}{2z_0} (x_0 - x_s)^2 \right] dx_s.
\end{aligned} \tag{5.46}$$

Since the sine and cosine terms in the integrands of Eq. (5.46) vary rapidly in the range where $(x_0 - x_s)^2$ is large, the net contribution to the integral there is small because the integral of the rapid oscillations averages to zero or relatively small values. Significant contributions to the integral are due to the regions near the *stationary points*, where $(x_0 - x_s)^2$ is small. It is fortunate that errors in the quadratic approximation of r_{0s} in the Fresnel diffraction formula occur for relatively large values of $(x_0 - x_s)^2$, where rapid phase oscillations in Eq. (5.46) minimize the impact on integration.

For example, Fig. 5.8 shows the exact and Fresnel approximations to r_{0s} , along with the cosine integrand of Eq. (5.46) versus x_s for $x_0 = 0$, $z_0 = 200\lambda$ and $L = 100\lambda$. The observation distance of $z_0 = 200\lambda$ is well below the limit specified by Eq. (5.45) of 428λ . As x_s increases, the error in r_{0s} also increases. However, oscillations in the cosine term also increase rapidly as the error in r_{0s} increases. These rapid oscillations tend to reduce error in the integral. In fact, if L increases for the same observation distance, error in the integral decreases, as shown in Fig. 5.9. This example illustrates that the Fresnel approximation can be applied to a much wider range than Eq. (5.45) indicates, due to the Principle of Stationary Phase. Of course, the variation in $U_s(x_s)$ must be much slower than the variation in the trigonometric terms for this condition to be true, and some geometries may exhibit larger error than others, as shown in the example of Fig. 5.9.

Applications of Fresnel diffraction are described in Section 5.3.

5.2.6 The Fraunhofer Approximation

Consider the Fresnel diffraction formula given in Eq. (5.44). If the aperture-to-observation distance z_0 is increased further, such that

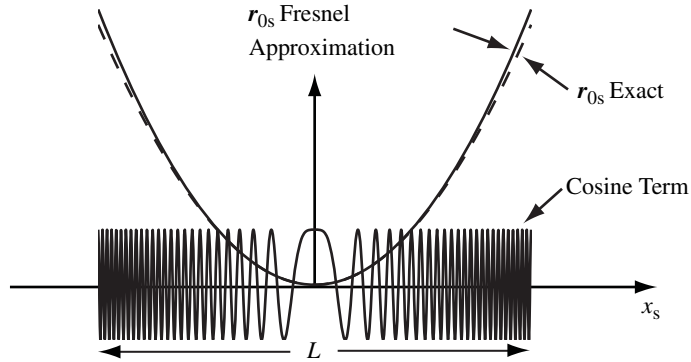


Fig. 5.8. Analysis of stationary phase for a one-dimensional example with $z_0 = 200\lambda$. The r_{0s} exact and Fresnel approximation values are shown versus x_s for $x_0 = 0$ and $L = 100\lambda$ in Eq. (5.46). Notice that, as the error in r_{0s} increases, so does the oscillation frequency of the cosine term.

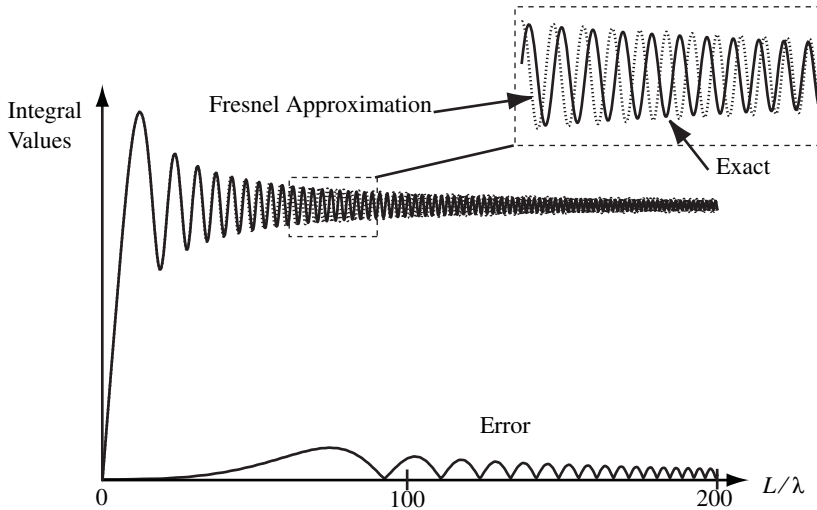


Fig. 5.9. Integral values illustrating the stationary phase effect. Integral values of Eq. (5.46) are shown versus aperture size L for $x_0 = 0$ and $z_0 = 200\lambda$. Notice that errors between the Fresnel approximation and exact values generally decrease as the aperture size increases.

$$z_0 \gg \frac{k}{2}(x_s^2 + y_s^2)_{\max}, \tag{5.47}$$

then

$$\exp\left[\frac{jk}{2z_0}(x_s^2 + y_s^2)\right] \rightarrow 1 \quad (5.48)$$

and

$$\begin{aligned} U_0(x_0, y_0) &= \frac{j e^{jkz_0}}{\lambda z_0} e^{j\frac{k}{2z_0}(x_0^2 + y_0^2)} \\ &\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_s^+(x_s, y_s) \exp\left[\frac{-jk}{z_0}(x_0 x_s + y_0 y_s)\right] dx_s dy_s \\ &= \frac{j e^{jkz_0}}{\lambda z_0} e^{j\frac{k}{2z_0}(x_0^2 + y_0^2)} \mathbf{F}_{\eta = \frac{y_0}{\lambda z_0}} \mathbf{F}_{\xi = \frac{x_0}{\lambda z_0}} [U_s^+(x_s, y_s)]. \end{aligned} \quad (5.49)$$

The approximation of Eq. (5.48) is the *Fraunhofer approximation*, and Eq. (5.49) is the *Fraunhofer diffraction formula*. The region of space for which the Fraunhofer diffraction is valid is called the *far-field* or *Fraunhofer region*.

Inspection of the Fraunhofer formula shows that the scalar electric field $U_0(x_0, y_0)$ is but a scaled two-dimensional Fourier transform of the field $U_s^+(x_s, y_s)$ in the aperture, evaluated at spatial frequencies

$$\xi = \frac{x_0}{\lambda z_0}, \quad \eta = \frac{y_0}{\lambda z_0}. \quad (5.50)$$

The form of the phasefronts reaching the observation plane from each point in the open aperture are planar, as indicated by

$$\frac{2\pi}{\lambda z_0}(x_0 x_s + y_0 y_s) = \text{constant}. \quad (5.51)$$

In fact, at the observation plane, direction cosines

$$\begin{aligned} \alpha &= \lambda \xi \\ \beta &= \lambda \eta \\ \gamma &= \sqrt{1 - \alpha^2 - \beta^2} \\ &= \sqrt{1 - (\lambda \xi)^2 - (\lambda \eta)^2}, \end{aligned} \quad (5.52)$$

determine the angular orientation of each plane wave. It is important to keep in mind that the wavelets are not planar at the aperture, but they attain an effectively planar shape after traversing the distance specified by the constraint of Eq. (5.47). Fraunhofer diffraction is discussed in more detail in Section 5.4.

Example 5.1: Regions of validity for approximations using a 1mm diameter aperture.

Consider a 1mm diameter aperture illuminated by a $\lambda = 500\text{nm}$ plane wave, as shown in Fig. 5.10. The Huygens wavelet described by Eq. (5.29) can be used to calculate the diffraction pattern at any distance from the aperture. However, The Fresnel wavelet described by Eq. (5.39) may lead to a simpler calculation for distances z_0 given by

$$z_0 \gg \sqrt[3]{\frac{\pi}{4}(0.5 \times 10^{-6}) \left(\frac{0.5 \times 10^{-3}}{0.5 \times 10^{-6}} \right)^{4/3}} = 0.0046 \text{ m}$$

from Eq. (5.45). If the distance z_0 is such that

$$z_0 \gg \frac{\pi}{0.5 \times 10^{-6}} (0.5 \times 10^{-3})^2 = 1.6 \text{ m},$$

from Eq. (5.47), the plane-wave Fraunhofer approximation of Eq. (5.49) may be used. Figure 5.10 illustrates that the regions of validity for the Fresnel and Fraunhofer approximations start at the distances calculated above. In practice, it may be necessary to increase the observation distances for negligible error.

5.2.7 Transfer function of free space

In this section, the Fourier description of propagation (left-side path in Fig. 5.3) is derived. That is, the field in the aperture first undergoes a Fourier transform. Then, an operator is applied that contains the effect of propagation. Finally, the resulting field undergoes an inverse transform to yield the field in the observation plane. The development includes a discussion of how the free-space point spread function method and the Fourier method are related through Weyl's integral. [Weyl, 1919]

If the plane $z = 0$ contains the aperture and plane $z = z_0$ describes the observation plane, the Huygens wavelet in free space given by Eq. (5.29) is linear and shift invariant, and we can write

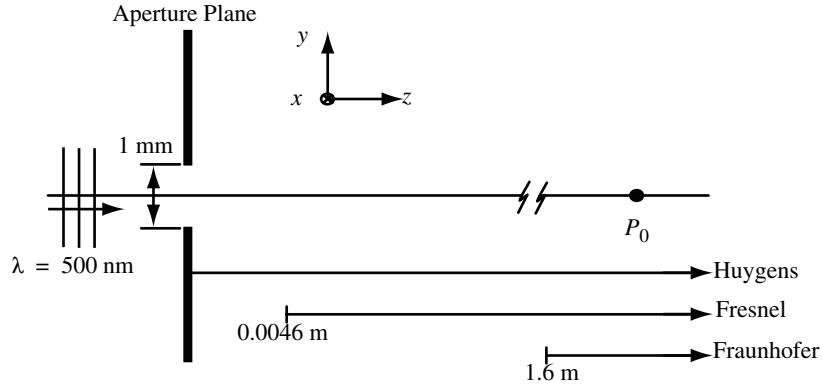


Fig. 5.10. A $\lambda = 0.5 \mu\text{m}$ plane wave illuminates a 1 mm diameter aperture. The Huygens calculation is valid over all values of z_0 , including right up to the aperture plane. The Fresnel approximation is valid for $z_0 \gg 0.0046 \text{ m}$, and sometimes closer, depending on the behavior of the stationary phase points. The Fraunhofer approximation is valid for $z_0 \gg 1.6 \text{ m}$. A Fourier transform of the aperture is observed in the far-field Fraunhofer region.

$$h_z^H(\mathbf{r}_0; \mathbf{r}_s) = h_z^H(x_0 - x_s, y_0 - y_s; z_0). \quad (5.53)$$

The point spread function $h_z^H(x_0 - x_s, y_0 - y_s; z_0)$ relates the aperture field $U_s^+(x_s, y_s)$ to the observation field $U_0(x_0, y_0)$ with a convolution as shown in Eq. (5.25), where

$$U_0(x_0, y_0) = U_s^+(x_s, y_s) ** h_z^H(x_s, y_s; z_0), \quad (5.54)$$

and $**$ represents two dimensional convolution.

What we seek in this section is the Fourier method, where

$$\mathbf{F}_\eta \mathbf{F}_\xi [U(\mathbf{r}_0)] = \mathbf{F}_\eta \mathbf{F}_\xi [U_s^+(\mathbf{r}_s)] \mathbf{F}_\eta \mathbf{F}_\xi [h_z^H(\mathbf{r}_0; \mathbf{r}_s)]. \quad (5.55)$$

We start with the plane-wave decomposition of an expanding spherical wave given by Weyl,

$$\frac{e^{jkr}}{-jkr} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{jk(ax + \beta y + \gamma z)} d\Omega. \quad (5.56)$$

The integration variable in Eq. (5.56) can be changed from solid angle to direction cosine by application of

$$d\Omega = \frac{d\alpha d\beta}{\gamma}, \quad (5.57)$$

which results in¹

$$\frac{e^{jkr}}{-jkr} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\gamma} e^{jk(ax + \beta y + \gamma z)} d\alpha d\beta. \quad (5.58)$$

Now the change of variables

$$\frac{\alpha}{\lambda} = \xi, \quad \frac{\beta}{\lambda} = \eta, \quad (5.59)$$

with

$$d\alpha = \lambda d\xi, \quad d\beta = \lambda d\eta, \quad (5.60)$$

results in

$$\frac{e^{jkr}}{-jkr} = \frac{\lambda^2}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\gamma} e^{jk\gamma z} e^{j2\pi(\xi x + \eta y)} d\xi d\eta. \quad (5.61)$$

Application of $\partial/\partial z$ to each side Eq. (5.61) results in

$$\begin{aligned} \frac{\partial}{\partial z} \left(\frac{e^{jkr}}{-jkr} \right) &= \frac{\lambda^2}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\gamma} \frac{\partial}{\partial z} (e^{jk\gamma z}) e^{j2\pi(\xi x + \eta y)} d\xi d\eta \\ &= \frac{\lambda^2}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\gamma} jk\gamma e^{jk\gamma z} e^{j2\pi(\xi x + \eta y)} d\xi d\eta \\ &= \frac{jk\lambda^2}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{jk\gamma z} e^{j2\pi(\xi x + \eta y)} d\xi d\eta \\ &= \frac{jk\lambda^2}{2\pi} \mathbf{F}_y^{-1} \mathbf{F}_x^{-1} [e^{jk\gamma z}]. \end{aligned} \quad (5.62)$$

1. Notice that the integration range of Eq. (5.58) extends from $-\infty$ to $+\infty$, which is required for the convergence of the integral. A translation of Weyl's original paper discusses the integration range of Eq. (5.56) in this way: "One does not thus only have to summarize all plane waves, whose propagation direction (α, β, γ) with the z axis forms an angle θ between 0 and $\pi/2$, but still another continuous series of complex angles of inclination θ ."

Rearranging Eq. (5.62) results in

$$\frac{\partial}{\partial z} \left(\frac{e^{jkz}}{2\pi r} \right) = \mathbf{F}_y^{-1} \mathbf{F}_x^{-1} [e^{jk\gamma z}]. \quad (5.63)$$

Since

$$h_z^H(\mathbf{r}_0; \mathbf{r}_s) = \frac{\partial}{\partial z_s} \left(\frac{e^{jkz_s}}{2\pi r_{0s}} \right) \quad (5.64)$$

from Eq. (5.29), substitution of Eq. (5.63) into Eq. (5.64) and then performing a Fourier transform yields

$$\mathbf{F}_\eta \mathbf{F}_\xi [h_z^H(\mathbf{r}_0; \mathbf{r}_s)] = e^{jk\gamma z_0} = H_z(\gamma; z_0), \quad (5.65)$$

where $H_z(\gamma; z_0) = \exp(jk\gamma z_0)$ is the *Transfer Function of Free Space*. $H_z(\gamma; z_0)$ relates the Fourier transform of an electric field distribution at $z = 0$ to the Fourier transform of the same electric field distribution after propagating distance z_0 . Note that there are no approximations in this scalar theory. Mathematically,

$$\begin{aligned} \mathbf{F}_\eta \mathbf{F}_\xi [U_0(\mathbf{r}_0)] &= \mathbf{F}_\eta \mathbf{F}_\xi [U_s^+(\mathbf{r}_s) ** h_z^H(\mathbf{r}_s)] \\ &= \mathbf{F}_\eta \mathbf{F}_\xi [U_s^+(\mathbf{r}_s)] \mathbf{F}_\eta \mathbf{F}_\xi [h_z^H(\mathbf{r}_s)] \\ &= \mathbf{F}_\eta \mathbf{F}_\xi [U_s^+(\mathbf{r}_s)] H_z(\gamma; z_0) \\ &= \mathbf{F}_\eta \mathbf{F}_\xi [U_s^+(\mathbf{r}_s)] e^{jk\gamma z_0}. \end{aligned} \quad (5.66)$$

The Fourier transform of $U_s^+(\mathbf{r}_s)$ is given special consideration, where

$$\begin{aligned} \mathbf{F}_\eta \mathbf{F}_\xi [U_s^+(\mathbf{r}_s)] &= \iint_{-\infty}^{\infty} U_s^+(x_s, y_s) e^{-j2\pi(\xi x_s + \eta y_s)} dx_s dy_s \\ &= A_s^+(\xi, \eta), \end{aligned} \quad (5.67)$$

and

$$\begin{aligned}
U_s^+(x_s, y_s) &= \iint_{-\infty}^{\infty} A_s^+(\xi, \eta) e^{j2\pi(\xi x_s + \eta y_s)} d\xi d\eta \\
&= \mathbf{F}_{y_s}^{-1} \mathbf{F}_{x_s}^{-1} [A_s^+(\xi, \eta)].
\end{aligned} \tag{5.68}$$

Fourier transform variables in Eq. (5.67) are ξ and η , which have units of inverse length. A physical interpretation of the transform variables is that they are spatial frequencies.¹ That is, a particular value of ξ corresponds to a spatial frequency component of the electric field in the x direction. For example, a single spatial frequency specified by $A_s^+(\xi, \eta) = \delta(\xi - \xi_0, \eta)$ corresponds to a complex electric field amplitude of $U_s^+(x_s, y_s) = \exp(j2\pi\xi_0 x_s)$. The physical electric field amplitude corresponding to this complex amplitude, after adding the $e^{-j\omega t}$ time dependence, is $\text{Re}[U_s^+(x_s, y_s)] = \cos(2\pi\xi_0 x_s - \omega t)$, which is a simple cosine traveling along the x_s axis with spatial frequency ξ_0 . Units of $A_s^+(\xi, \eta)$ are Vm.

In terms of the *spatial frequency spectrum* $A_s^+(\xi, \eta)$, propagation in free space is defined by

$$A_z(\xi, \eta; z_0) = A_s^+(\xi, \eta) e^{jk\gamma z_0}, \tag{5.69}$$

where $A_z(\xi, \eta; z_0)$ is the spatial frequency spectrum at plane z_0 . To find $U_0(x_0, y_0; z_0)$,²

$$\begin{aligned}
U_0(x_0, y_0; z_0) &= \mathbf{F}_{y_0}^{-1} \mathbf{F}_{x_0}^{-1} [A_z(\xi, \eta; z_0)] \\
&= \mathbf{F}_{y_0}^{-1} \mathbf{F}_{x_0}^{-1} [A_s^+(\xi, \eta) e^{jk\gamma z_0}].
\end{aligned} \tag{5.70}$$

5.2.8 Angular spectrum of plane waves

The exponential in the integrand of Eq. (5.68) has a familiar form. In fact, the exponential has the same mathematical structure as a plane wave if we add a time dependence and substitute $\xi = \alpha/\lambda$ and $\eta = \beta/\lambda$, where

$$e^{j[2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y) - \omega t]} = e^{j(\mathbf{k} \cdot \mathbf{r} - \omega t)} \Big|_{z=0}. \tag{5.71}$$

-
1. In some reference material, the spatial frequency variables (ξ, η, ζ) are written as $(\sigma_x, \sigma_y, \sigma_z)$. In this chapter, σ is used to denote $\sigma = (\alpha^2 + \beta^2)^{1/2}$ for direction cosines.
 2. Writing the observation-space field distribution as $U_0(x_0, y_0; z_0)$ is redundant, because (x_0, y_0) coordinates are in the z_0 plane. However, the field is written in this way here to emphasize the propagation over distance z_0 .

The (α, β, γ) terms in the propagation vector

$$\mathbf{k} = k(\alpha\hat{\mathbf{x}} + \beta\hat{\mathbf{y}} + \gamma\hat{\mathbf{z}}) = 2\pi[(\alpha/\lambda)\hat{\mathbf{x}} + (\beta/\lambda)\hat{\mathbf{y}} + (\gamma/\lambda)\hat{\mathbf{z}}] \quad (5.72)$$

are direction cosines of the plane wave, and $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are unit vectors of the Cartesian coordinate system. Therefore, the field transmitted through the aperture can be considered as a collection of plane waves, with weighting coefficients given by $A_s^+(\xi, \eta)$, where

$$U_s^+(x_s, y_s)e^{-j\omega t} = \iint_{\infty} A_s^+(\xi, \eta)e^{j(\mathbf{k} \cdot \mathbf{r} - \omega t)} \Big|_{z=0} d\xi d\eta . \quad (5.73)$$

Since, from Eq. (5.69) the spatial frequency spectrum at plane z_0 is given by the multiplication of $A_s^+(\xi, \eta)$ and $\exp(jk\gamma z_0)$, the complex electric field $U_0(x_0, y_0; z_0)e^{-j\omega t}$ can be written as

$$\begin{aligned} U_0(x_0, y_0; z_0)e^{-j\omega t} &= \iint_{\infty} A_s^+(\xi, \eta)e^{j2\pi\left(\frac{z_0}{\lambda}\right)} e^{j2\pi\left(\frac{\alpha}{\lambda}x_0 + \frac{\beta}{\lambda}y_0\right)} e^{-j\omega t} d\xi d\eta \\ &= \iint_{\infty} A_s^+(\xi, \eta)e^{j\left[\frac{2\pi}{\lambda}(\alpha x_0 + \beta y_0 + \gamma z_0) - \omega t\right]} d\xi d\eta \\ &= \iint_{\infty} A_s^+(\xi, \eta)e^{j(\mathbf{k} \cdot \mathbf{r} - \omega t)} d\xi d\eta . \end{aligned} \quad (5.74)$$

Equation (5.74) shows that field $U_0(x_0, y_0; z_0)e^{-j\omega t}$ is composed of a summation of plane waves, where the complex amplitude of each plane wave is specified by $A_s^+(\xi, \eta)$, $\xi = \alpha/\lambda$, and $\eta = \beta/\lambda$. The weighting coefficients of the *plane-wave spectrum* are the same as for $z = 0$, except for the z -dependent term $\exp(jk\gamma z_0)$ that naturally arises from the propagation of each plane-wave component. Normally, the time-dependent term is dropped from the discussion, but it is added here for clarity.

Recall that a single spatial frequency component of the $A_s^+(\xi, \eta)$ spectrum corresponds to a traveling cosine wave of electric field amplitude in the $z = 0$ plane. That is, if $A_s^+(\xi, \eta) = \delta(\xi - \xi_0, \eta)$, the complex field at $z = 0$ is $U_s^+(x_s, y_s) = e^{j(2\pi\xi_0 x_s - \omega t)}$, with the real-valued electric field $U_s^+(x_s, y_s) = \text{Re} [e^{j(2\pi\xi_0 x_s - \omega t)}] = \cos(2\pi\xi_0 x_s - \omega t)$. The $\hat{\mathbf{k}}$'s of the plane-wave components are directed with an angle equal to or less than 90° with respect to the z axis if $\sigma = (\alpha^2 + \beta^2)^{1/2} \leq 1$. Low spatial frequencies exhibit $\hat{\mathbf{k}}$'s at small angles

with respect to the z axis, as shown in Fig. 5.11. $\hat{\mathbf{k}}$'s that correspond to higher spatial frequencies, exhibit $\hat{\mathbf{k}}$'s at larger angles. Since there is a one-to-one correspondence between the spatial frequency of a cosine component and the angle of a plane wave, $A_s^+(\xi, \eta)$ is also called the *angular spectrum of plane waves* for the field $U_s^+(\mathbf{r}_s)$.

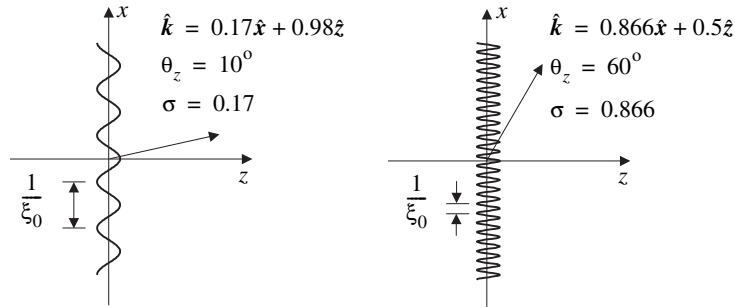


Fig. 5.11. Interpretation of the angular spectrum of plane waves at $t = 0$ and $z = 0$. On the left is shown a single spectral component (traveling cosine wave along the x axis) with a large period and correspondingly low spatial frequency. Consequently, the plane-wave angle is small. On the right is shown a single spectral component with a smaller period and correspondingly higher spatial frequency, which exhibits a larger diffraction angle.

Example 5.2: Angular spectrum of a cosine field distribution.

For example, a simple standing wave cosine field distribution in the aperture $U_s^+(x_s) = \cos[2\pi(a_0/\lambda)x_s] = \cos 2\pi\xi_0 x_s$ exhibits the angular spectrum

$$\begin{aligned} A_s^+(\xi, \eta) &= \int_{-\infty}^{\infty} \int \cos(2\pi\xi_0 x_s) e^{-j2\pi(\xi x_s + \eta y_s)} dx_s dy_s \\ &= \frac{1}{2} [\delta(\xi - \xi_0, \eta) + \delta(\xi + \xi_0, \eta)]. \end{aligned} \quad (5.75)$$

By applying Eq. (5.70), the complex electric field distribution at observation plane z_0 is given by¹

1. Note that the substitution $\gamma = [1 - (\lambda\xi)^2 - (\lambda\eta)^2]^{1/2}$ is used.

$$\begin{aligned}
U_0(x_0, y_0; z_0) &= \mathbf{F}_{y_0}^{-1} \mathbf{F}_{x_0}^{-1} [A_s^+(\xi, \eta) e^{jk\gamma z_0}] = \mathbf{F}_{y_0}^{-1} \mathbf{F}_{x_0}^{-1} \left\{ \frac{1}{2} [\delta(\xi - \xi_0, \eta) + \delta(\xi + \xi_0, \eta)] e^{jk\gamma z_0} \right\} \\
&= \frac{1}{2} \mathbf{F}_{y_0}^{-1} \mathbf{F}_{x_0}^{-1} \left[e^{jkz_0 \sqrt{1 - (\lambda\xi)^2 - (\lambda\eta)^2}} \delta(\xi - \xi_0, \eta) + e^{jkz_0 \sqrt{1 - (\lambda\xi)^2 - (\lambda\eta)^2}} \delta(\xi + \xi_0, \eta) \right] \\
&= \frac{1}{2} \left[e^{jkz_0 \sqrt{1 - (\lambda\xi_0)^2}} e^{j2\pi\xi_0 x_0} + e^{jkz_0 \sqrt{1 - (\lambda\xi_0)^2}} e^{-j2\pi\xi_0 x_0} \right] \\
&= \frac{1}{2} \left[e^{j2\pi\xi_0 x_0} + e^{-j2\pi\xi_0 x_0} \right] e^{jkz_0 \sqrt{1 - (\lambda\xi_0)^2}} \\
&= \cos(2\pi\xi_0 x_0) e^{jkz_0 \sqrt{1 - (\lambda\xi_0)^2}} = \cos(k\alpha_0 x_0) e^{jkz_0 \sqrt{1 - \alpha_0^2}},
\end{aligned} \tag{5.76}$$

which is simply a combination of two plane waves symmetrically propagating in the z direction with a difference angle of $2\cos^{-1}\alpha_0$. Irradiance is proportional to $\cos^2(k\alpha_0 x_0) = 1/2 + (1/2)\cos(k\alpha_0 x_0)$, which is simply the fringe pattern resulting from two plane waves, as described in Chapter 4.

Each delta function in the angular spectrum of Eq. (5.76) corresponds physically to a plane wave in a direction specified by the offset α_0 in angular spectrum direction cosine coordinates. A pictorial representation of the angular spectrum on the (α, β) plane is shown in Fig. 5.12 (a). The two delta functions are located symmetrically along the α axis. For any angular spectrum, all propagating plane waves are contained within or on a circle of radius $\sigma = 1$. Figure 5.12 (b) shows the proper interpretation in terms of the $\hat{\mathbf{k}}$'s on the (x, z) plane.

Example 5.2. Illustration of the procedure to follow when calculating diffraction from field $U_s^+(\mathbf{r}_s)$ to field $U_0(\mathbf{r}_0)$ on a plane at distance z_0 . Specifically, this procedure is:

- 1) Find the Fourier transform of $U_s^+(x_s, y_s)$ with respect to the frequency variables ξ and η . This result is the angular spectrum $A_s^+(\xi, \eta)$.
- 2) Multiply $A_s^+(\xi, \eta)$ by the transfer function of free space $H_z(\gamma, z) = \exp(jk\gamma z)$, where $\gamma = [1 - (\lambda\xi)^2 - (\lambda\eta)^2]^{1/2}$. This result is the angular spectrum $A_z(\xi, \eta; z_0)$. Simplify, if possible.

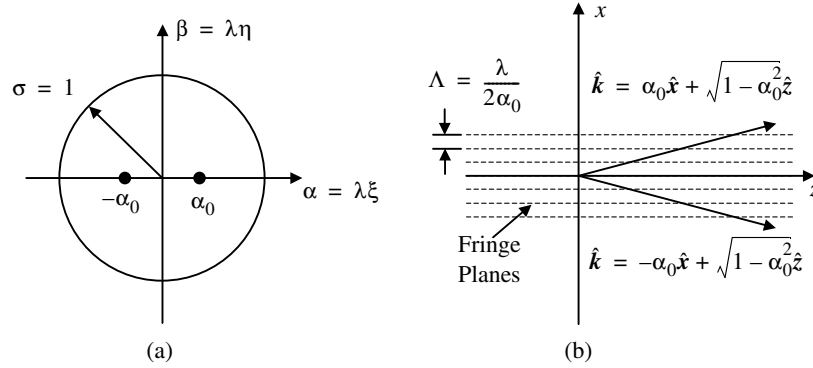


Fig. 5.12. Pictorial representation of the angular spectrum for a cosine field distribution. a) Representation of the angular spectrum on the (α, β) plane. Two delta functions are shown along the α axis; one at $-\alpha_0$ and one at $+\alpha_0$. For any angular spectrum, all propagating plane waves are contained within a circle of radius $\sigma = 1$; b) The resulting irradiance distribution is a cosine fringe pattern resulting from the two symmetrical plane-wave components.

- 3) Find the inverse Fourier transform of $A_z(\xi, \eta; z_0)$ with respect to spatial variables x_0 and y_0 . This result is the physical field on the observation plane at distance z_0 .

The phase of the transfer function $H_z(\gamma; z_0) = \exp(jkz_0)$ is $\psi = kz_0 = kz_0(1 - \sigma^2)^{1/2}$. For $|\sigma| \leq 1$, ψ is simply the phase shift of each plane wave that occurs between the aperture plane and the observation plane, as shown in Fig. 5.13. The phase is quadratic for small σ . The phase increases as z_0 increases. Maximum phase shift occurs for a plane wave traveling with $\sigma = 0$ ($\gamma = 1$, \hat{k} along the z axis), as shown in Fig. 5.13 (b), where the phase shift value is simply 2π multiplied by the number of wavelengths separating the source plane and the observation plane. Phase shift ψ decreases as propagation angle increases. In the limit of $\sigma = 1$ ($\gamma = 0$, $\hat{k} \perp$ to the z axis) there is no phase shift between the two planes, as shown in Fig. 5.13 (c). Graphs of the phase versus σ for different distances z_0 are shown in Fig. 5.14.

The magnitude of the transfer function is unity for $|\sigma| \leq 1$, as shown in Fig 5.15. That is, amplitude of spatial frequency components for which $|\sigma| \leq 1$ remain constant as they propagate from plane $z = 0$ to plane $z = z_0$. Because there is no reduction in amplitude, spatial frequency components for which $|\sigma| \leq 1$ are called *propagating components*.

Plane-wave components for which $|\sigma| > 1$ exhibit a different character than the propagating components. For $|\sigma| > 1$, transfer function magnitude falls off exponentially with increasing σ , where the decay constant linearly increases with z_0 , as shown in Fig. 5.15. That is, spatial frequency components for $|\sigma| > 1$ decay rapidly with increasing distance from the aperture plane to the observation plane.

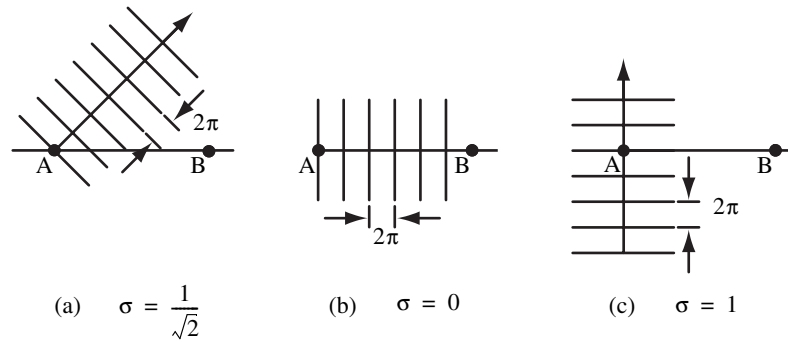


Fig. 5.13. Inclined plane-wave components of the angular spectrum traveling from A to B over distance z_0 . (a) One plane wave (propagating component) is shown traveling at a 45° angle from A to B, with $\sigma = 1/\sqrt{2}$. The phase between A and B is $\psi = kz_0/\sqrt{2}$. Note that there are 4 fringe crests between A and B; (b) Plane-wave component traveling at 0° angle from A to B, with $\sigma = 0$. The phase between A and B is $\psi = kz_0$. This is the angle at which the maximum phase shift is observed, *i.e.*, the maximum number of fringe crests between A and B; (c) Plane-wave component traveling at 90° angle from A to B, with $\sigma = 1$. When the wave propagates perpendicular to AB, the change in phase from A to B is zero ($\psi = 0$).

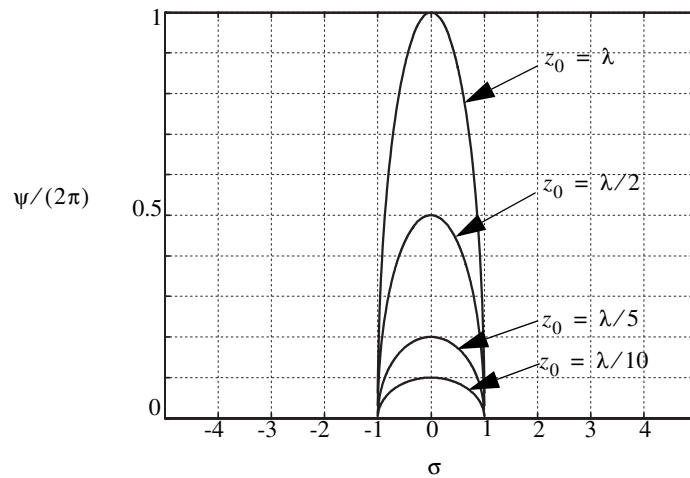


Fig. 5.14. Phase of the free-space transfer function versus σ . When $|\sigma| > 1$, the phase of the transfer function is zero. When $|\sigma| \leq 1$, the phase is a quadratic function of σ and z_0 .

Because these spatial frequency components decay rapidly, they are called *nonpropagating* or *evanescent* components. There is no phase factor in the transfer function associated with evanescent components.

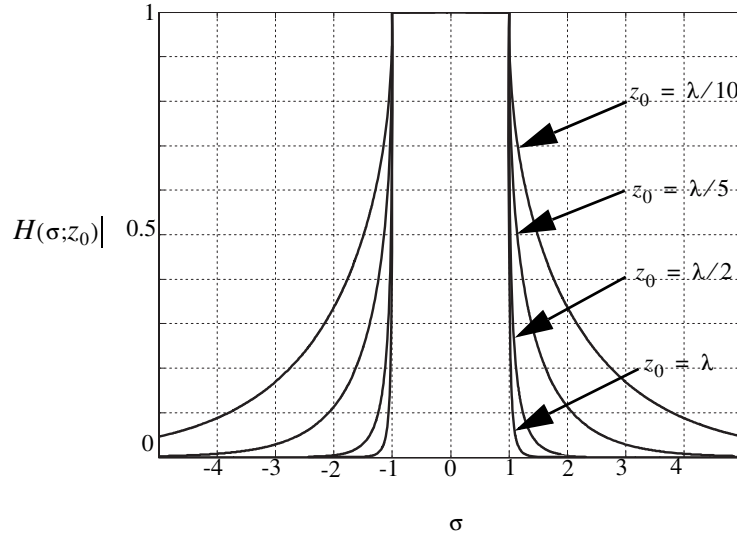


Fig. 5.15. Magnitude of the free-space transfer function versus σ . When $|\sigma| \leq 1$, the transfer function magnitude is unity. When $|\sigma| > 1$, the magnitude of the transfer function decays exponentially as a function of σ and z_0 . Spectral components in the region $|\sigma| > 1$ are called evanescent waves.

For example, consider again the angular spectrum

$$A_s^+(\xi, \eta) = \frac{1}{2} \left[\delta(\xi - \xi_0, \eta) + \delta(\xi + \xi_0, \eta) \right], \quad (5.77)$$

where $\lambda \xi_0 = \alpha_0 > 1$. Following a procedure similar to that used for Eq. (5.76) the observation-plane electric field at z_0 is given by

$$\begin{aligned} U_0(x_0, y_0; z_0) &= \frac{1}{2} \left[e^{jk(a_0 x_0 + z_0 \sqrt{1 - \alpha_0^2})} + e^{jk(-a_0 x_0 + z_0 \sqrt{1 - \alpha_0^2})} \right] \\ &= \frac{1}{2} \left[e^{jk a_0 x_0} + e^{-jk a_0 x_0} \right] e^{-k z_0 \sqrt{\alpha_0^2 - 1}} \\ &= \cos(k a_0 x_0) e^{-k z_0 \sqrt{\alpha_0^2 - 1}}, \end{aligned} \quad (5.78)$$

and the square of the electric field is proportional to

$$|U_0(x_0, y_0; z_0)|^2 = e^{-2k z_0 \sqrt{\alpha_0^2 - 1}} \left[\frac{1}{2} + \frac{1}{2} \cos(2k a_0 x_0) \right], \quad (5.79)$$

where now an exponential decay factor is associated with the distance z_0 from the source plane. If the time dependence is included, Eq. (5.78) can be interpreted as two counter-propagating evanescent waves along the x_0 axis, which produce an evanescent standing wave. There is no phase shift in the z_0 direction. That is, the phase is similar to a plane wave propagating at $\sigma = 1$, as shown in Fig. 5.13(c).

Example 5.3: Slit Diffraction

In addition to supplementing the conceptual understanding of diffraction, the concept of angular spectrum is very useful in the calculation of diffraction patterns, such as a collimated laser beam passing through an aperture. For example, consider the simple case of an ideal plane wave illuminating a one-dimensional slit aperture. The electric field amplitude immediately after the aperture is

$$U_s^+(x_s) = \text{rect}\left(\frac{x_s}{d}\right), \quad (5.80)$$

where d is the aperture width, as shown in Fig. 5.16. The angular spectrum in the source plane is

$$\begin{aligned} A_s^+(\xi, \eta) &= \iint_{-\infty}^{\infty} U_s^+(x_s, y_s) e^{-j2\pi(\xi x_s + \eta y_s)} dx_s dy_s \\ &= \iint_{-\infty}^{\infty} \text{rect}\left(\frac{x_s}{d}\right) e^{-j2\pi\xi x_s + \eta y_s} dx_s dy_s \\ &= d \text{sinc}(d\xi) \delta(\eta). \end{aligned} \quad (5.81)$$

Significant energy is present only in the range of angles where $|\alpha| < \lambda/d$. That is, the diffraction pattern does not spread rapidly with increasing z_0 if d is large.

Equation (5.81) is calculated in closed form. For more complicated electric field distributions, it is common to use computer techniques, such as the fast-Fourier transform (FFT).¹

1. [Bracewell, 1978, Ch. 18].

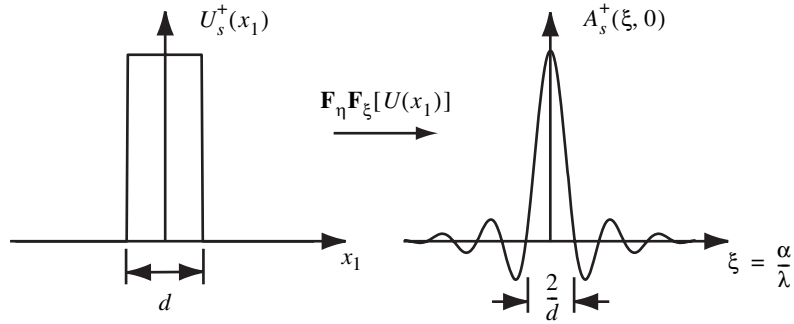


Fig. 5.16. Plane-wave spectrum for slit diffraction. The aperture width d as well as the wavelength of incident light has a direct effect on the diffraction pattern after the slit. Most of the energy in the diffracted wave becomes contained within $|\alpha| < \lambda/d$, and thus the diffraction pattern does not spread rapidly if d is large.

Example 5.4: Small slit diffraction

With respect to the slit aperture of Example 5.3, notice that $A_s^+(\xi, \eta)$ can contain significant amplitude past $|\sigma| = 1$ if $d < \lambda$. Evanescent components of the angular spectrum for $|\sigma| > 1$ decay exponentially away from the source (aperture) plane. Propagation of the angular spectrum to plane z_0 is simply the multiplication of Eq. (5.81) with the transfer function of free space. That is,

$$A_z(\xi, \eta; z_0) = d \operatorname{sinc}(d\xi) \delta(\eta) e^{jkz_0 \sqrt{1 - (\lambda\xi)^2}}. \quad (5.82)$$

For example, the fields for $d = 2\lambda$ are shown in Fig. 5.17 at the $z = 0$ aperture plane and in Fig. 5.18 for an observation plane with $z_0 = 3\lambda$. Notice that the aperture plane exhibits a large amount of evanescent energy that decays rapidly as the observation plane distance increases. Essentially all energy at the $z_0 = 3\lambda$ observation plane is propagating energy. The propagating energy fills the range $|\sigma| \leq 1$, which implies that the diffraction pattern changes rapidly with z_0 due to the large amount of energy in the high angles. Notice how the diffraction pattern is changed compared to how it appears in the aperture plane. Within the range of propagating plane waves, this change is due to the phase factor $\psi = kz(1 - \sigma^2)^{1/2}$ in the angular spectrum.

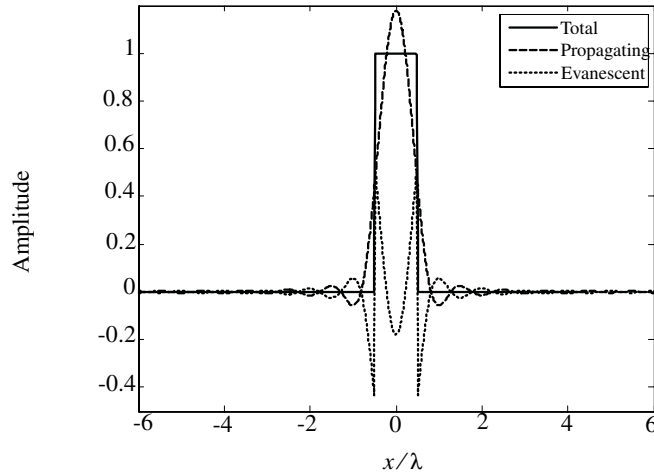


Fig. 5.17. Propagation of a uniform electric field through a $d = 2\lambda$ slit for $z_0 = 0$. The propagating wave is derived from the $|\sigma| \leq 1$ portion of the angular spectrum, while the evanescent wave is derived from the $|\sigma| > 1$ portion.

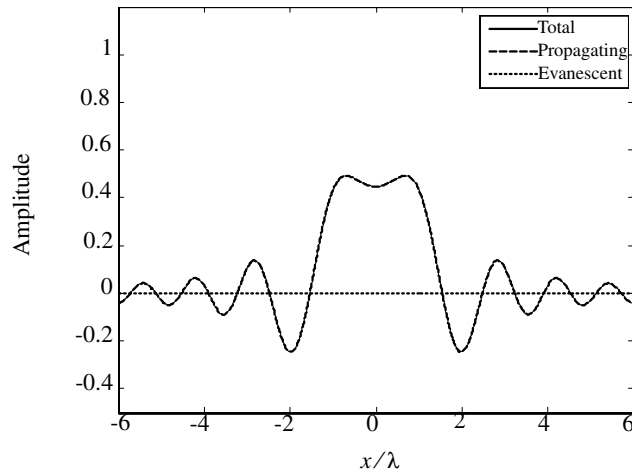


Fig. 5.18. Propagation of a uniform electric field through a $d = 2\lambda$ slit for $z_0 = 3\lambda$. This figure displays the total, propagating, and evanescent waves for the wave described in Fig. 5.17 with an additional propagation distance. In this case, we see the effect of propagation. The evanescent wave has virtually disappeared at a distance of $z_0 = 3\lambda$, and the propagating wave has begun take on the form of a crude sinc function. Because the evanescent wave provides little contribution, the total wave is for all intents and purposes the same as the propagating wave. As the distance z_0 increases to the Fraunhofer region, this propagating wave becomes more like a sinc function.

5.2.9 Talbot Effect

The Talbot effect is a curious diffraction phenomenon that occurs for periodic objects illuminated with laser light. If the period of the object is large compared to λ , the maximum extent of the object's angular spectrum is limited to small angles where $\sigma \ll 1$. The object is periodic in one dimension, say x_s , and infinite in the orthogonal dimension. These restrictions allow a one-dimensional simplification of the free-space transfer function to

$$e^{jkz_0\gamma} \approx e^{jkz_0} e^{-j\frac{kz_0}{2}\alpha^2} = e^{jkz_0} e^{-j\frac{kz_0}{2}(\xi\lambda)^2}. \quad (5.83)$$

Consider transmission of a weak phase grating in the aperture plane illuminated by an on-axis plane wave. The transmitted field is

$$U_s^+(x_s) = e^{j\phi_0 \cos(2\pi x_s/T)} \approx 1 + j\phi_0 \cos(2\pi x_s/T), \quad (5.84)$$

where the Taylor-series expansion in Eq. (5.84) is valid for $\phi_0 \ll 1$. The angular spectrum of Eq. (5.84) is

$$\begin{aligned} A_s^+(\xi, \eta) &= \iint_{-\infty}^{\infty} [1 + j\phi_0 \cos(2\pi x_s/T)] e^{-j2\pi(\xi x_s + \eta y_s)} dx_s dy_s \\ &= \left\{ \delta(\xi) + \frac{1}{2}j\phi_0 [\delta(\xi - \xi_0) + \delta(\xi + \xi_0)] \right\} \delta(\eta), \end{aligned} \quad (5.85)$$

where $\xi_0 = 1/T$. The form of Eq. (5.85) is that of three plane waves, as shown in Fig. 5.19. One plane wave is on axis, and the other two are symmetrical about the z axis. The resulting electric field is a slightly modified version of Eq. (5.76), where the addition of an on-axis plane wave modifies field and irradiance distributions. This condition is called *three-beam diffraction*. It differs from the simple two-plane wave interference pattern discussed in Chapter 4 by the addition of the on-axis plane wave.

Application of the free-space transfer function in Eq. (5.83) yields

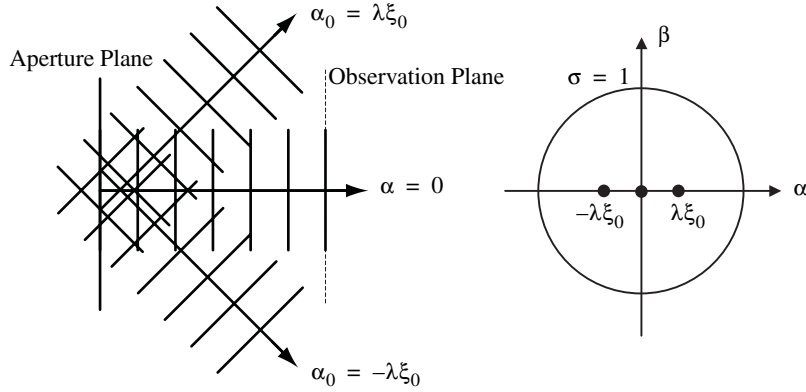


Fig. 5.19. Talbot imaging illustrated as three-beam diffraction. The angular spectrum of a laser beam transmitted through a weak phase grating consists of three plane waves, which combine to form periodic replications of the original phase distribution along the z axis.

$$\begin{aligned}
 A_z(\xi, \eta; z_0) &= e^{jkz_0} \left\{ \delta(\xi) + \frac{1}{2} j \phi_0 e^{-j \frac{kz_0}{2} (\lambda \xi_0)^2} \left[\delta(\xi - \xi_0) + \delta(\xi + \xi_0) \right] \right\} \delta(\eta) \\
 &= e^{jkz_0} \left\{ \delta(\xi) + \frac{1}{2} j \phi_0 e^{j\psi} \left[\delta(\xi - \xi_0) + \delta(\xi + \xi_0) \right] \right\} \delta(\eta). \quad (5.86)
 \end{aligned}$$

Note that if the $\psi = -(kz_0/2)(\lambda \xi_0)^2$ phase factor inside the brackets of Eq. (5.86) is $\psi = -2\pi p$, where $p \in \{0, \pm 1, \pm 2, \dots\}$, the exponential term is unity and the propagated angular spectrum is simply the angular spectrum of the original wave in Eq. (5.85) multiplied by an on-axis phase factor $\exp(jkz_0)$. That is, periodic z_0 distances given by

$$z_0 = \frac{2p}{\lambda \xi_0^2} = \frac{2\lambda p}{\alpha_0^2}, \quad p \in \{0, \pm 1, \pm 2, \dots\}, \quad (5.87)$$

where $\alpha_0 = \lambda/T$ produce exact replicas of the aperture-plane angular spectrum $A_s^+(\xi, \eta)$, except for the additional leading phase constant. Now consider planes where $\psi = -\pi(2p + 1)$ and the value of the bracketed exponential term is -1 . In this case, the observation plane wavefront is

$$U_0(x_0; z_0) = e^{jkz_0} e^{-jm \cos(2\pi x_0/T)}, \quad (5.88)$$

which is the phase conjugate of Eq. (5.84) multiplied by a phase constant. These conjugate planes occur at

$$z_0 = \frac{2p+1}{\lambda\xi_0^2} = \frac{\lambda(2p+1)}{\alpha_0^2}, \quad p \in \{0, \pm 1, \pm 2, \dots\}, \quad (5.89)$$

which are midway between planes specified by Eq. (5.87). If $\psi = -\pi(2p+1/2)$ and the value of the bracketed exponential is $-j$, the reconstructed wavefront in the observation plane is

$$U_0(x_0; z_0) = e^{jkz_0} [1 + m \cos(2\pi x_0/T)], \quad (5.90)$$

which is a purely amplitude modulated field in the x_0 direction. Note that the leading phase term does not modify irradiance. These purely amplitude modulated planes occur at

$$z_0 = \frac{2p + \frac{1}{2}}{\lambda\xi_0^2} = \frac{\lambda\left(2p + \frac{1}{2}\right)}{\alpha_0^2}, \quad p \in \{0, \pm 1, \pm 2, \dots\}, \quad (5.91)$$

which are located periodically just after the true phase reconstructions. If $\psi = -\pi(2p+3/2)$ and the value of the bracketed exponential is j , the reconstructed wavefront is

$$U_0(x_0; z_0) = e^{jkz_0} [1 - m \cos(2\pi x_0/T)], \quad (5.92)$$

which is similar to Eq. (5.90), except that the contrast is reversed. These reversed contrast planes occur at

$$z_0 = \frac{2p + \frac{3}{2}}{\lambda\xi_0^2} = \frac{\lambda\left(2p + \frac{3}{2}\right)}{\alpha_0^2}, \quad p \in \{0, \pm 1, \pm 2, \dots\}, \quad (5.93)$$

which are located after the conjugate phase reconstruction planes. The planes periodically specified by Eqs. (5.87), (5.89), (5.91) and (5.93) occur periodically. They are separated by a base distance z_{base} specified by

$$z_{\text{base}} = \frac{1}{2\lambda\xi_0^2} = \frac{\lambda}{2\alpha_0^2} = \frac{T^2}{2\lambda}. \quad (5.94)$$

The sequence of planes described by Eqs. (5.87) through (5.94) is shown in Fig. 5.20. The object described by Eq. (5.94) is a periodic phase modulation. The object and its pure phase reconstructions are labeled as A planes. B planes are locations where the phase modulation becomes amplitude modulation. C planes are the conjugate phase reconstructions, and D planes are the reversed contrast amplitude reconstructions. Notice that the grating object is completely reconstructed at the Talbot distance $z_{\text{Talbot}} = 4z_{\text{base}}$, or

$$z_{\text{Talbot}} = \frac{2}{\lambda\xi_0^2} = \frac{2\lambda}{\alpha_0^2} = \frac{2T^2}{\lambda}. \quad (5.95)$$

The mathematics specified by Eqs. (5.86) through (5.95) apply equally well to a cosine amplitude grating. In this case, the source plane now starts at *D*, and the periodic reconstruction planes occur in the sequence *DABCDABCD*, etc.

An example of a Talbot pattern, which is sometimes called a *Talbot carpet*, formed by illuminating a rectangular grating with a Gaussian laser beam is shown in Fig. 5.21.¹ The near-field irradiance close to the grating clearly shows periodic variations predicted by Eq. (5.95), even though the grating modeled in Fig. 5.21 is a binary amplitude grating. Due to the finite width of illumination, the Talbot pattern loses visibility and diverges after a few Talbot cycles. When the same grating is illuminated with a converging laser beam, as shown in the right portion of Fig. 5.21, the Talbot pattern changes scale according to the marginal angle of the focus beam. Far-field diffraction orders at beam focus are also clearly observed.

1. [McMorran and Cronin, 2008]

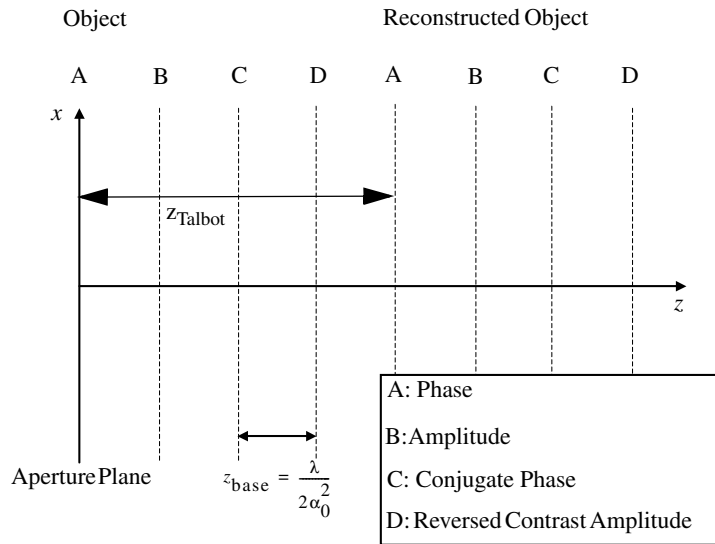


Fig. 5.20. Periodic spacing of observation planes associated with the Talbot effect. Due to the Talbot effect, various reconstructions of a phase grating can be observed at regular distances away from the aperture. The various observation planes are spaced by $z_{\text{base}} = \lambda/2\alpha_0^2$. Each type of observation is periodic with a period of $z_{\text{Talbot}} = 4z_{\text{base}} = 2\lambda/\alpha_0^2$.

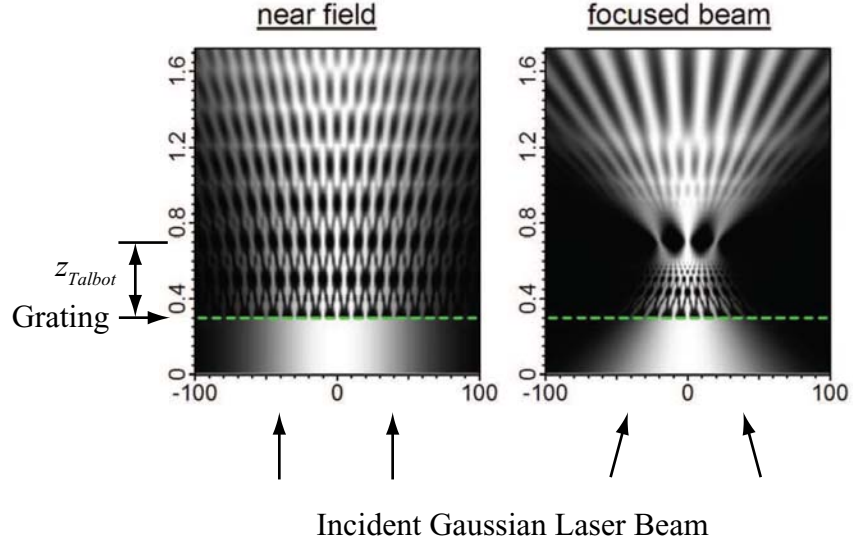


Fig. 5.21. Simulation of a Talbot irradiance pattern (Talbot carpet) behind a rectangular grating illuminated by a $\lambda = 500\text{nm}$ laser beam with a beam diameter of $100\mu\text{m}$ at the grating. The grating is a binary transmission grating with a period $d = 10\mu\text{m}$ and having $4\mu\text{m}$ wide open slits. In the left portion of the figure, the beam waist is position at the grating. In the right portion of the figure, the beam is focused 0.7mm behind the grating.

5.2.10 Babinet's Principle

From Eq. (5.25), we know that scalar diffraction is a linear process. Therefore, fields transmitted through the aperture can be divided into several geometrically separate parts $U_{sA}^+, U_{sB}^+, U_{sC}^+ \dots$ where

$$U_s^+ = U_{sA}^+ + U_{sB}^+ + U_{sC}^+ \dots, \quad (5.96)$$

and U_s^+ the total field transmitted through the aperture. Mathematically, scalar diffraction is a linear operation \mathbf{L} , and the diffraction calculation applied to U_s^+ can be separated into parts such that

$$\mathbf{L}(U_s^+) = \mathbf{L}(U_{sA}^+) + \mathbf{L}(U_{sB}^+) + \mathbf{L}(U_{sC}^+) + \dots \quad (5.97)$$

Equation (5.97) is, in essence, Babinet's Principle. The diffracted field from the aperture can be decomposed into components and propagated separately. The separate results are then combined to yield the equivalent calculation of propagating U_s^+ . This concept, while conceptually appealing, is deceptively simple. Consider apertures A , B and C , as shown in Fig. 5.22, which are displayed in the $z = 0$ plane. Hatched areas are opaque, and clear areas are transparent. Performing $\mathbf{L}(U_{sA}^+)$ directly is certainly possible, but it is usually easier to break apart the

complicated aperture A into the simpler apertures B and C , because known solutions exist for them. The problem then reduces to $\mathbf{L}(U_{sA}^+) = \mathbf{L}(U_{sC}^+) - \mathbf{L}(U_{sB}^+)$. The procedure of finding equivalent solutions in terms of simpler apertures is called *aperture algebra*.

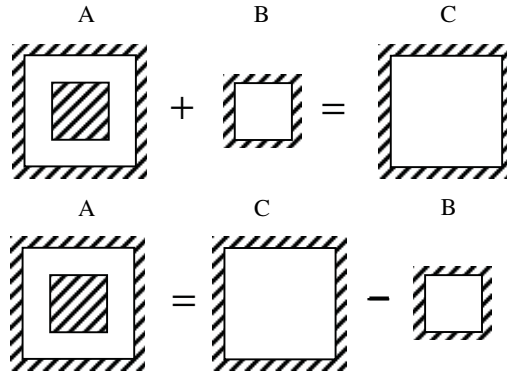


Fig. 5.22. Aperture algebra. The linear property of scalar diffraction leads to the Babinet principle, which is used to reform diffraction problems in terms of simpler, known distributions. The example above shows how to use aperture algebra to reduce the complicated problem of solving for the diffraction from aperture A into two separate propagations with apertures B and C . Then, the separate results are added to produce the equivalent result.

A second important concept related to Babinet's principle is the *Principle of Complementary Apertures*. Consider the diffraction from two apertures D and E that are illuminated by field U_s^- , as shown in Fig. 5.23. The apertures are complementary, in that opaque regions of D are clear in E and vice versa. For example, since apertures D and E in Fig. 5.23 are complementary, it is a common fallacy to assume that they produce the same diffraction pattern. In fact, when aperture algebra is applied, summation of the aperture fields produces the illumination field, U_s^- . The proper interpretation is $\mathbf{L}(U_{sD}^+) = \mathbf{L}(U_s^-) - \mathbf{L}(U_{sE}^+)$. A much different irradiance pattern results from the proper interpretation. An example of using Babinet's Principle to solve a diffraction problem is presented in Section 5.3.3

5.3 Fresnel Diffraction

Section 5.2 provides a rigorous introduction to scalar diffraction. In this section, these concepts are applied to solving particular diffraction problems where the observation plane is relatively close to the diffracting aperture.

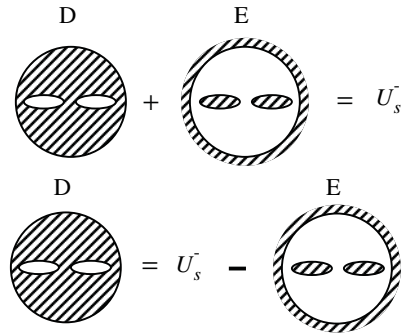


Fig. 5.23. Complementary Apertures. Apertures D and E are complementary. Consider an experiment where the D aperture is placed in the illumination field and the transmission is measured. The D aperture is removed and replaced by the E aperture. Again, the transmitted is measured. When the fields transmitted by the apertures are added together, it equals the incident field. Application of a linear diffraction operation to the aperture algebra distribution shows that the diffraction pattern observed for apertures separately are not the same, due to the diffraction pattern from the incident field.

5.3.1 Fresnel Zones

Of particular interest in working with laser beams is the property of light transmitted through circular apertures. For example, it common to observe *Fresnel Rings* behind an aperture that is illuminated with a uniform laser beam, as shown in Fig. 5.24. The rings can be problematic if the optical system is sensitive to variations in laser power across the beam. Portions of the detailed mathematical development presented in Section 5.2 are now applied to this problem. The starting point is developing an understanding of the on-axis behavior of the light field, which exhibits alternate maxima and minima depending on the wavelength, illumination conditions and aperture diameter. In this discussion, the concept of *Fresnel zones* is defined, which is a geometrical analysis tool that is extremely useful in the prediction and interpretation of diffraction patterns.

5.3.1.1 Application of the Rayleigh-Sommerfeld diffraction formula to a circular aperture

An approximation of Rayleigh-Sommerfeld diffraction formula in Eq.(5.20) is now evaluated in detail for a single point source illuminating the aperture. The geometry used in this calculation is shown in Fig. 5.25, where an on-axis point source P_{src} at distance z_{src} with amplitude A illuminates a circular aperture. The observation point P_0 is on axis a distance z_0 from the aperture. The light in the aperture re-radiates as a collection of secondary spherical wavelets that are weighted in amplitude and phase according to the field reaching the aperture plane

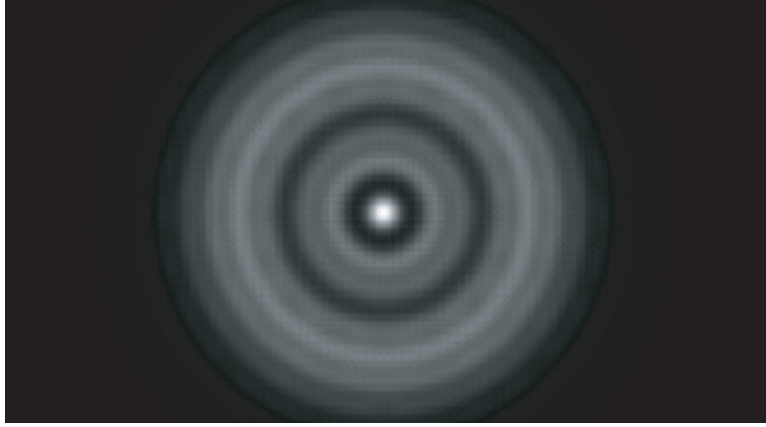


Fig. 5.24. The irradiance observed behind a circular aperture illuminated with a uniform laser beam commonly exhibits Fresnel Rings, which dramatically alter the irradiance profile. In this figure, a uniform and collimated $\lambda = 0.5\mu\text{m}$ laser beam illuminates a 1 mm diameter round aperture. The observation is made at a distance of 100mm. Three bright rings are distinctly observed in the pattern, and the pattern has a bright central spot. The diameter of the outermost ring is nearly equal to the diameter of the aperture.

from P_{src} . One of the secondary radiators at point Q is shown in Fig. 5.25, which is a radial distance ρ_s from the axis.

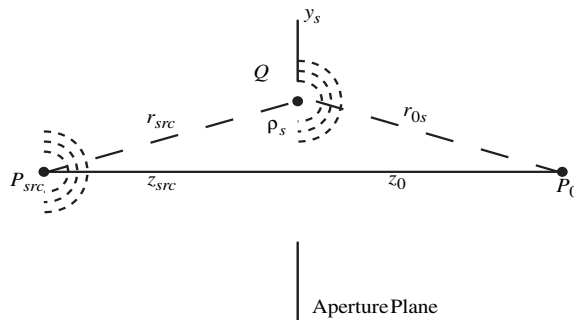


Fig. 5.25. Geometry for discussion of Fresnel diffraction developed with the Rayleigh-Sommerfeld diffraction formula.

At P_0 , the contributions from all of the reradiated waves are summed in an integral. Phases of individual reradiated waves are extremely important, because in-phase radiators contribute strongly to the field amplitude at P_0 , while out-of-phase radiators do not. The phase of the light field from any radiator Q arriving at point P_0 is determined by the path length $\overline{P_{src}QP_0}$.

The source and observation points are sufficiently far from the aperture that the obliquity term γ_z in the Rayleigh-Sommerfeld diffraction formula can be ignored.

We also assume that any amplitude terms in the integrand can be approximated by their axial values. That is, $1/r_{src} \approx 1/z_{src}$ and $1/r_{0s} \approx 1/z_0$. These approximations do not exclude near-field effects of the phase in the integrand, so a nearly accurate prediction of the light field behavior is obtained, even close to the aperture. After substitution of the point-source field $U_{src}(r_{src}) = A/r_{src} \exp(jkr_{src})$ and the above approximations into Eq. (5.20), the appropriate diffraction integral is

$$U_0(P_0) = \frac{-jA}{\lambda z_{src} z_0} \iint_{ap} e^{jk(r_{src} + r_{0s})} dx_s dy_s, \quad (5.98)$$

where distances r_{src} and r_{0s} correspond to distances $\overline{P_{src}Q}$ and $\overline{QP_0}$, respectively. The integration is performed in the plane of the aperture with coordinates (x_s, y_s) over its open portion.

Equation (5.98) is now modified for a radially-symmetric coordinate system in order to simplify the discussion. With the change of variables

$$\rho_s = \sqrt{x_s^2 + y_s^2}, \quad (5.99)$$

Eq. (5.98) for a circular aperture of radius a becomes

$$U_0(P_0) = \frac{-jA}{\lambda z_{src} z_0} \int_0^a e^{jk(r_{src} + r_{0s})} 2\pi\rho_s d\rho_s. \quad (5.100)$$

Next, we add an additional phase term that corresponds to the axial phase shift from the source to the observation point. The conjugate term is added inside the integral in order to balance the equation. Lastly, an amplitude term

$$L = \frac{z_{src} z_0}{z_{src} + z_0}, \quad (5.101)$$

is added. The result is

$$\begin{aligned} U_0(P_0) &= \frac{-jA}{\lambda z_{src} z_0} L e^{jk(z_{src} + z_0)} \int_0^a e^{jk(r_{src} + r_{0s})} e^{-jk(z_{src} + z_0)} \frac{2\pi\rho_s}{L} d\rho_s \\ &= \frac{-jA}{\lambda z_{src} z_0} L e^{jk(z_{src} + z_0)} \int_0^a e^{jk[r_{src} + r_{0s} - (z_{src} + z_0)]} \frac{2\pi\rho_s}{L} d\rho_s. \end{aligned} \quad (5.102)$$

Note that the term $r_{src} + r_{0s} - (z_{src} + z_0)$ corresponds to the *optical path difference* (OPD) between the axial ray and light through point Q along $\overline{P_{src}QP_0}$. Notice also that the leading term in Eq. (5.102) is the field amplitude at P_0 from the point source that is observed if the aperture is removed, with the exception of a $-j$ factor. We call this field value $U_\infty(P_0)$. Written explicitly in these terms, Eq. (5.102) becomes

$$U_0(P_0) = \frac{-jU_\infty(P_0)}{\lambda} \int_0^a e^{jk\text{OPD}(\rho_s)} \frac{2\pi\rho_s}{L} d\rho_s. \quad (5.103)$$

Equation (5.103) can be evaluated exactly if the function $\text{OPD}(\rho_s)$ is known.¹ However, it is convenient to approximate $\text{OPD}(\rho_s)$ with a quadratic expansion in our ideal example, such that

$$\text{OPD}(\rho_s) \approx \frac{\rho_s^2}{2L}, \quad (5.104)$$

in order to understand the basic characteristics of the observed diffraction patterns. In practice, this approximation yields surprisingly accurate results.

It is possible that the observation point P_0 lies on the left-hand side of the aperture. This situation can effectively be realized, for example, with an additional lens system. In this case, the value of z_0 is negative, and Eq. (5.104) can be either positive or negative, depending on the value of z_0 .

5.3.1.2 Definition of Fresnel zones

If we consider in detail the interference generated between light emitted from different locations in the aperture, the necessary condition for constructive interference (bright spot) at the observation point is $\text{OPD} = m\lambda$, where m is an integer. Conversely, the condition for destructive interference (dark spot) is $\text{OPD} = (m + 1/2)\lambda$. If we start from the center of the aperture at the axis, OPD increases as ρ_s increases. We define ρ_1 where $\text{OPD}(\rho_1) = \lambda/2$. Light from the on-axis radiator interferes destructively with light from secondary radiators at ρ_1 when observed at P_0 . Radius ρ_1 corresponds to the maximum radial boundary of the first *Fresnel zone*. Likewise, the aperture can be divided into multiple Fresnel zones, each corresponding to successive increments of $\lambda/2$ in OPD. The number of Fresnel zones across a circular aperture of radius a is called the *Fresnel number* N_f and, in the quadratic approximation of Eq. (5.104), is given by

1. For example, $\text{OPD}(\rho_s)$ can be calculated in an optical system with ray-tracing techniques.

$$N_f = \frac{a^2}{\lambda L}. \quad (5.105)$$

The radius of the m^{th} Fresnel zone is given by

$$\rho_m = \sqrt{m\lambda L}, \quad (5.106)$$

and the area of each Fresnel zone is

$$Area_f = \pi\rho_{m+1}^2 - \pi\rho_m^2 = \pi\lambda. \quad (5.107)$$

Notice the curious result that areas of all Fresnel zones are equal. An aperture with $N_f = 5$ with $z_{src} \rightarrow \infty$ is shown in Fig. 5.26, where the odd zones (m odd) are colored white and even zones (m even) are colored black.

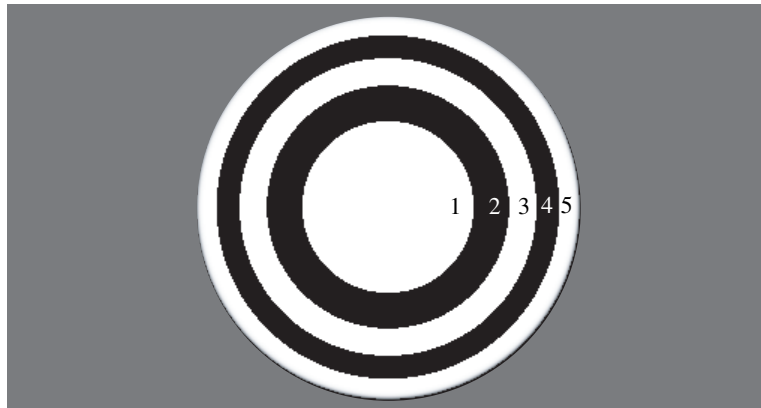


Fig. 5.26. An aperture with five Fresnel zones. Odd numbered zones are shown in white, and even zones are in black.

5.3.2 Fresnel diffraction from apertures

In this section, the results of Section 5.3.1 are used to describe the diffraction pattern observed behind a circular aperture. Sections 5.3.2.1 and 5.3.2.2 discuss the axial irradiance, Section 5.3.2.3 describes the off-axis irradiance, and Section 5.3.2.4 discusses applying the results to converging and diverging wave fields.

5.3.2.1 Diffraction behind a circular aperture.

A change of variables in Eq. (5.103) modifies the diffraction problem in terms of the Fresnel number. If the variable q is defined as the number of half-waves of OPD, where

$$q = \frac{\text{OPD}(\rho_s)}{\lambda/2}, \quad (5.108)$$

and the quadratic approximation to $\text{OPD}(\rho_s)$ in Eq. (5.104) is used, Eq. (5.103) becomes

$$\begin{aligned} U_0(P_0) &= -j\pi U_\infty(P_0) \int_0^{N_f} e^{j\pi q} dq \\ &= -\frac{j\pi}{j\pi} U_\infty(P_0) (e^{j\pi N_f} - 1) \\ &= -U_\infty(P_0) (e^{j\pi N_f} - 1). \end{aligned} \quad (5.109)$$

The extremely simple form of Eq. (5.109) has only one variable, the Fresnel number, other than the $U_\infty(P_0)$ term. That is, *the axial field behind the aperture depends only on the Fresnel number of the aperture*. This result written in terms of the axial irradiance becomes

$$\begin{aligned} I_0(P_0) &= CU(P_0)U^*(P_0) \\ &= 4CI_\infty(P_0)\sin^2\left(\frac{\pi N_f}{2}\right) \end{aligned} \quad (5.110)$$

where $C = 1/2c\epsilon_0$ in vacuum. Notice that the axial irradiance varies periodically as a function of the Fresnel number. *For odd Fresnel numbers ($N_f = 1, 3, \dots$), the axial irradiance is maximum and is equal to four times the irradiance without any aperture present*. For laser systems, this high peak irradiance can be problematic if the peak coincides with an optical surface that is sensitive to damage. *For even Fresnel numbers ($N_f = 2, 4, 6, \dots$), the axial irradiance is zero*. If P_0 lies on the left-hand side of the aperture (z_0 is negative) and L is negative, the interpretation of Eq. (5.105) is that m and N_f are negative.

5.3.2.2 Physical interpretation of the axial irradiance behind a circular aperture.

The result described by Eq. (5.110) has a physical interpretation by applying the principle of interference. We start with an observation point P_0 such that there are two Fresnel zones in the aperture, as shown in Fig. 5.27.¹ Differential regions are defined such that δA_1 and δA_2 have the same differential areas. Area δA_1 is

1. Note that the aperture of Fig. 5.27 could physically be the same aperture as that shown in Fig. 5.26 with the observation point moved farther from the aperture.

located at the axis inside the first Fresnel zone. δA_2 is located at and just beyond the zone boundary between Fresnel zones. Since light waves from δA_1 and δA_2 are $\lambda/2$ out of phase, they destructively interfere. Since they have equal areas, the cancellation is complete, and there is a net zero contribution to the field at P_0 from δA_1 and δA_2 . The next set of matched differential areas beyond δA_1 and δA_2 have slightly larger radius, but the OPD between them is again $\lambda/2$, and the net contribution to the field at P_0 is zero. The same argument can be made for all matched differential areas within the two zones. Since the zones have equal total areas, the total field observed at P_0 has a zero value.

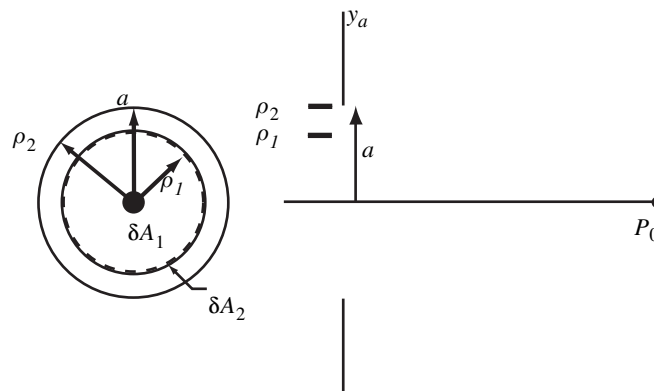


Fig. 5.27. An aperture with two Fresnel zones. Differential areas δA_1 and δA_2 are $\lambda/2$ out of phase, so they produce zero field amplitude at the observation point P_0 . A similar argument is applied to all matched areas within the two zones to explain the net zero field at the on-axis observation point.

Now the observation point P_0 is moved closer to the aperture. Radii of first and second Fresnel zones decrease, as shown in Fig. 5.28. Light from first and second zones still cancel, but light from the partial third zone does not. The result is an increasing axial irradiance as the observation point is moved closer to the aperture. Irradiance increases until the point where the third zone boundary is equal to the aperture radius. Beyond that point, portions of the fourth Fresnel zone are within the aperture, and some cancellation occurs with the third zone. Therefore, irradiance decreases again until it becomes zero at the point where the fourth zone boundary is equal to the aperture diameter. This argument can be extended until the observation point is very close to the aperture, where assumptions used in Eqs. (5.98) and (5.104) are no longer valid.

Note that there are several ways to change the Fresnel number for a particular source and observation point. For example, either the source or observation point can move, which change L . The size of the aperture can be increased or decreased,

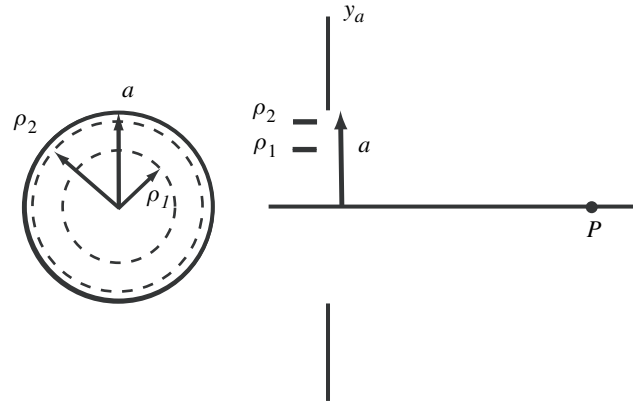


Fig. 5.28. An aperture with Fresnel number $2 < N_f < 3$. Light from the first two Fresnel zones cancel, but light from the partial third zone does not. The result is an increasing axial irradiance as the number of zones increases to $N_f = 3$.

which changes the number of zones transmitted through it. Lastly, the wavelength of the source can be changed.

This geometrical analysis of diffraction from a circular aperture is a very powerful tool in describing the axial irradiance of diffraction patterns. For example, consider again the geometry with two Fresnel zones in the aperture, as shown in Fig. 5.27. If the first zone is covered with an opaque mask, such that no light is transmitted in that zone, only light from the second Fresnel zone passes to the observation point. The irradiance at the observation point changes from zero to its maximum value $4I_\infty(P)$.

Lastly, application of the Fresnel-zone analysis to systems in which the aperture is not circular is discussed. For example, the aperture (colored in gray) shown in Fig. 5.29 is an irregular shape. A calculation of OPD between the source point and the observation point reveals that the aperture transmits a series of partial Fresnel zones, where odd zones are colored white in Fig. 5.29 and even zones are colored black. By geometrically determining the difference in the transmitted even and odd zone areas, we can determine that the net contribution is a bright spot on axis at the observation point.

The analysis of the preceding paragraphs suggests a simple tool to determine the irradiance at the observation point. The steps in this procedure are:

- 1) Geometrically divide the aperture into Fresnel zones based on the OPD between the source point and the observation point.
- 2) Geometrically sum the area of the odd transmitted zones.
- 3) Compare the area in (2) to the sum of the areas of the transmitted even Fresnel zones by subtracting the two areas.

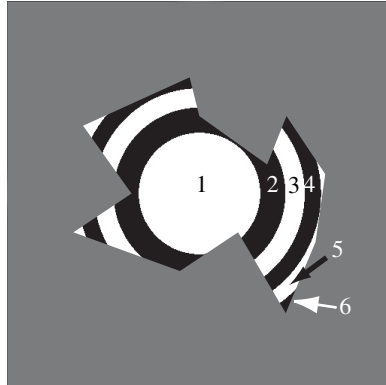


Fig. 5.29. An irregularly shaped aperture and the associated Fresnel zone pattern. The observation-point irradiance is a result of the balance between odd and even zones that are transmitted through the aperture.

- 4) Divide the result by the complete area of a Fresnel zone.
- 5) Multiply the square of this fraction by $4I_{\infty}(P)$. This value is the approximate irradiance at the observation point.

5.3.2.3 Off-axis irradiance behind a circular aperture

Division of the aperture into Fresnel zones and determination of the axial irradiance is a useful tool in diffraction analysis, and now we extend the geometrical model to calculation of the off-axis irradiance.

Consider a slightly off-axis observation point P_0' , as shown in Fig. 5.30. Calculation of OPD transmitted through the aperture results in a shifted Fresnel-zone pattern, where the center of the pattern is the intersection of the aperture plane and the line $P_{src}P_0'$. The shifted pattern for a $N_f = 5$ geometry is shown in Fig. 5.31a. Following the procedure outlined in Section 5.3.2.2, we find that, for small shifts of the observation point, the area of the fifth zone immediately starts to decrease, while the sixth zone enters the aperture. Therefore, the net balance of zone areas decrease, irradiance reduces, and the on-axis observation exhibits a local maximum. Shifting the observation point further off axis, the decrease of irradiance continues until the seventh zone enters the aperture, as shown in Fig. 5.31b. At that point, the irradiance begins to increase again. The increase continues to a maximum at the point where the eighth zone just starts to enter the aperture. However, this secondary maximum is not as large as the central maximum. This alternating observation of maxima and minima continues whenever a zone boundary crosses the edge of the aperture. For example, the next maximum occurs for the shift shown in Fig. 5.31c, where the first zone just crosses the edge. After this final maximum, the irradiance gradually decreases to zero, with no other significant peaks. The zone pattern for a very large shift is shown in

Fig. 5.31d, where the large number of zones in the transmitted section sum to nearly zero. Therefore, there are five maxima in the cross section of the irradiance pattern across the full aperture.

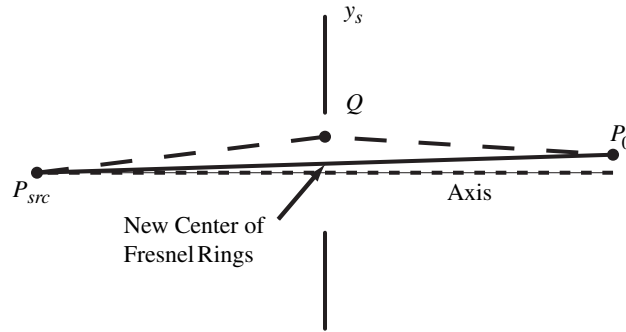


Fig. 5.30. An off-axis observation point is analyzed by considering the shifted Fresnel-zone pattern in the aperture with symmetry around $P_{src}P'_0$.

By recording the irradiance value determined by the geometrical procedure, a trace of the irradiance cross section can be determined. A circular expansion is then used to estimate the two-dimensional pattern. The result of this exercise for $N_f = 5$ is shown in Fig. 5.32. Notice that the pattern shows a bright central peak and two bright rings. A cross section of the profile exhibits five primary peaks, as shown in Fig. 5.33. In fact, we observe that *the number of primary peaks in the cross section of the irradiance pattern is equal to the Fresnel number for $N_f \geq 1$* . This useful fact allows quick estimation of diffraction patterns without complicated calculations, and it is a handy back-of-the-envelope engineering trick. Notice that the geometrical technique produces a result that is nearly identical to the complete calculation displayed in Fig. 5.24 that is found using the angular spectrum technique outlined in Section 5.2.8.

5.3.2.4 Fresnel zones in diverging, collimated and converging wave fields

The analysis of Sections 5.3.1.1 through 5.3.2.3 does not assume any special conditions about the axial position of the source point P_{src} . It could be on either side of the aperture plane or at an infinite conjugate. The position of the source point depends on the type of illumination, in terms of diverging, collimated or converging wave fields in the aperture. A diverging wave field is produced from a point source on the left-hand side of the aperture, as shown in Fig. 5.25. A collimated wave field is produced if $z_{src} \rightarrow \infty$, which can be accomplished by placing the source at the front focal point of a good lens before the aperture. A converging wave field produces an effective source point on the right-hand side of the aperture plane, where the position of z_{src} is at the wavefield's center of curvature.

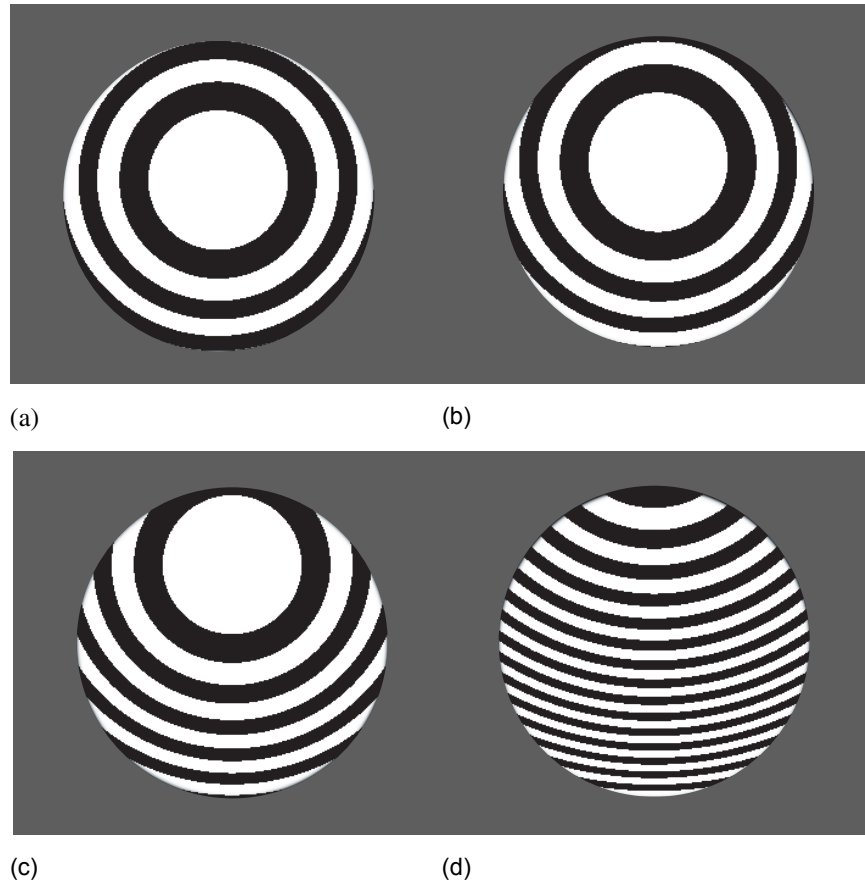


Fig. 5.31. (a)-(d) Shifted Fresnel-zone patterns in the aperture for increasing off-axis distance of the observation point.

First, consider the case as shown in Fig. 5.25, which is a diverging wavefront at the aperture plane. In this case, z_{src} and z_0 are positive, and Fresnel rings are observed in the diverging wavefront after the aperture. Of course, widths of the rings expand as the beam expands.

Next, consider the case where a collimated beam illuminates the aperture. If $z_{src} \rightarrow \infty$, $L \rightarrow z_0$ and the Fresnel number becomes

$$N_f = \frac{a^2}{\lambda z_0}. \quad (5.111)$$



Fig. 5.32. Geometrical estimation of the diffraction pattern from an observation point with $N_f = 5$.

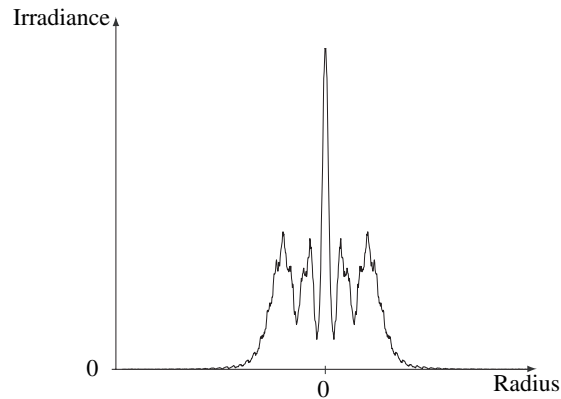


Fig. 5.33. Profile of the diffraction pattern shown in Fig. 5.32 for an observation point with $N_f = 5$. Note that the profile has five primary peaks.

Notice that the Fresnel number in Eq. (5.111) is very similar to Eq. (5.105), except that L is replaced with z_0 . That is, the diffraction pattern observed for a collimated beam is exactly the same as for a diverging or converging wavefront, except that the converging and diverging patterns occur at the *effective propagation distance* L and they are scaled in the transverse dimension. This observation leads to an extremely straightforward method for calculating near-field irradiance behind the aperture for any type of illuminating laser beam. That is,

- 1) Calculate the effective propagation distance L .

- 2) Use the angular spectrum technique described in Section 5.2.8 to calculate the field observed a distance L behind an aperture illuminated with a collimated wavefront.
- 3) Apply the scaling factor $\frac{z_0 + z_{src}}{z_{src}}$ to the transverse dimensions of the field. The scaling factor can be understood by tracing the diameter of the marginal rays from the source point through the aperture. The diffraction pattern scales as the ratio of the marginal ray diameter to the aperture diameter.
- 4) Scale the total energy in the pattern to the total energy at the aperture plane.

Example 5.5: Diffraction from a circular aperture illuminated with collimated laser light.

Consider the example of a unit-amplitude plane wave illuminating a 1mm diameter hole in a round screen. The on-axis irradiance from $z_0 = 0.05$ mm ($N_f = 10$) to $z_0 = 0.5$ mm ($N_f = 1$) is shown in Fig. 5.34. A cross section of the irradiance, as calculated with the angular spectrum method of Section 5.2.8, is shown in Fig. 5.35(a), and the associated phase is shown in Fig. 5.35(b).

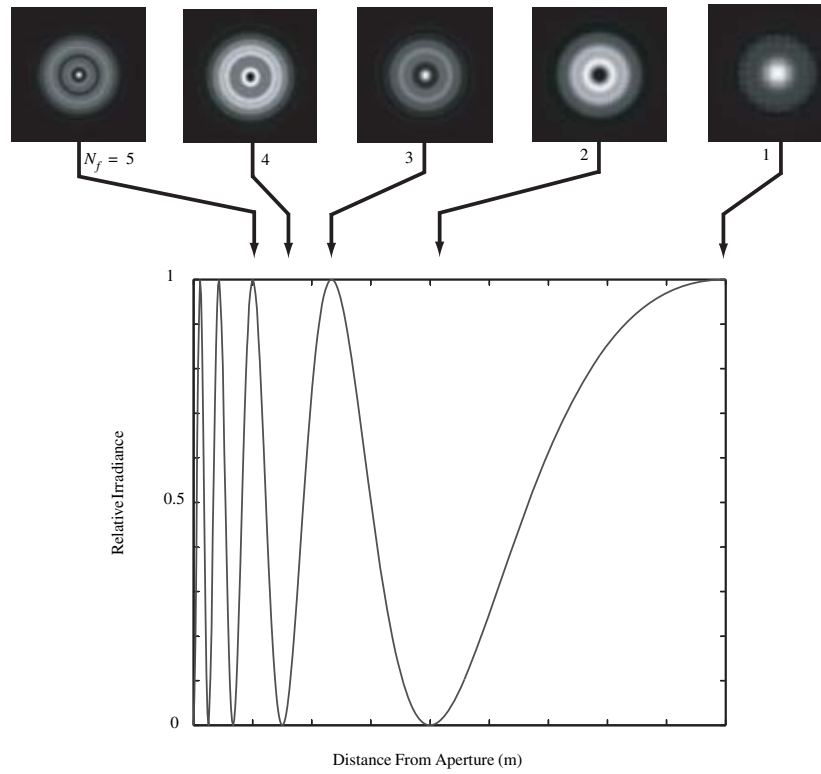


Fig. 5.34. Axial irradiance from a 1.0mm diameter circular aperture illuminated with a collimated and uniform $\lambda = 0.5 \mu\text{m}$ laser beam. Transverse irradiance distributions are shown for $N_f = 1$ through 5.

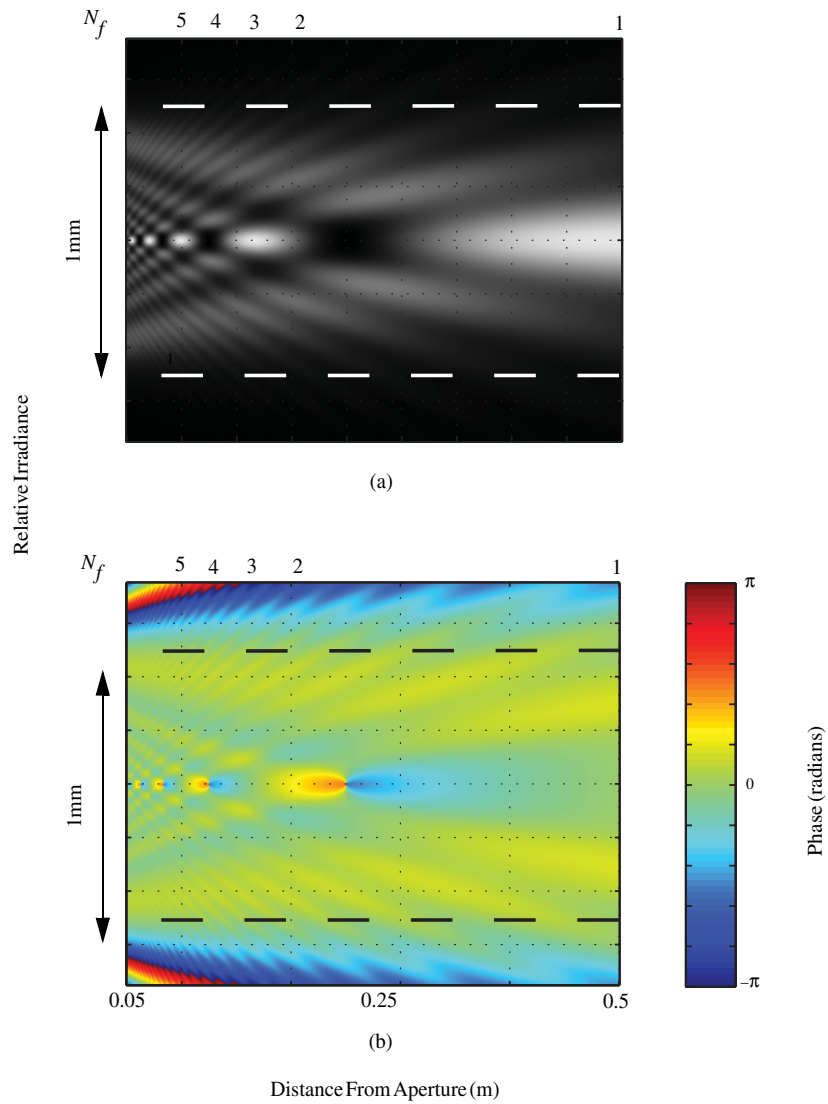


Fig. 5.35. Cross sections of the irradiance (a) and phase (b) diffracted from a 1.0mm diameter aperture illuminated with a collimated and uniform $\lambda = 0.5 \mu\text{m}$ laser beam. The phase distribution in (b) is the difference between the diffracted field and a plane wave. Dashed horizontal lines indicate the boundary of the geometrical shadow.

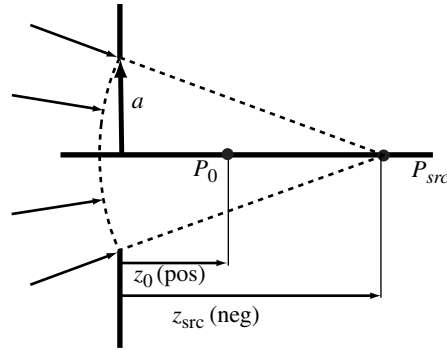


Fig. 5.36. The geometry for converging illumination on an aperture. In this case z_{src} is negative and z_0 is positive.

An important case is for converging illumination, as shown in Fig. 5.36. As a first approximation, this geometry can be used to simulate the focusing properties of a microscope objective or other lens from the exit pupil to a region close to the focus.¹ Notice that, in this case, z_{src} is negative and the Fresnel ring patterns get smaller as the observation plane at z_0 approaches the focus

Example 5.6: Irradiance in the focus cone of a lens illuminated by a laser.

Consider a 5 mm diameter lens with a focal length of 50 mm illuminated by a collimated and uniform $\lambda = 0.5 \mu\text{m}$ laser beam. Assume that the exit pupil is at the lens and has a 5 mm diameter ($a = 2.5 \text{ mm}$). The lens transforms the laser beam into a converging cone of light that passes through the exit pupil. Diffraction from the exit pupil causes Fresnel rings near the focus. Since the object conjugate for the lens is at infinity, the effective source position is $z_{src} = -50 \text{ mm}$. A diagram of the focus cone for this system is shown in Fig. 5.37. Equation (5.105) can be used to calculate the positions of the Fresnel patterns in the focus cone. Notice that Fresnel patterns similar to those in Fig. 5.34 are observed, but they are much smaller and are close to the focus region. Because the leading term $4I_{\infty}(P_0)$ in Eq. (5.110) increases quadratically as the observation point approaches focus, the on-axis irradiance increases quadratically also, unlike Fig. 5.34. Figure 5.38 shows yz -plane views of the irradiance and phase near focus.

1. At and very near the focus plane, the Fraunhofer approximation described in Section 5.2.6 can be used. Section 5.4 expands on the discussion of this approximation.

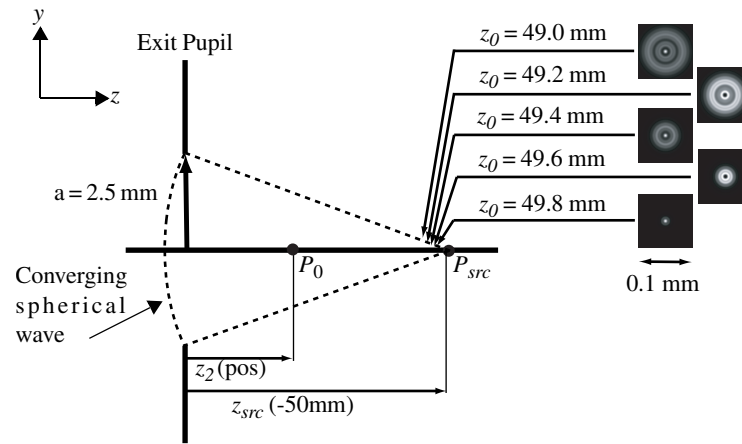


Fig. 5.37. Geometry for a laser beam converging through an aperture. $\lambda = 0.5\mu\text{m}$, $f = 50\text{mm}$ and $a = 2.5\text{mm}$.

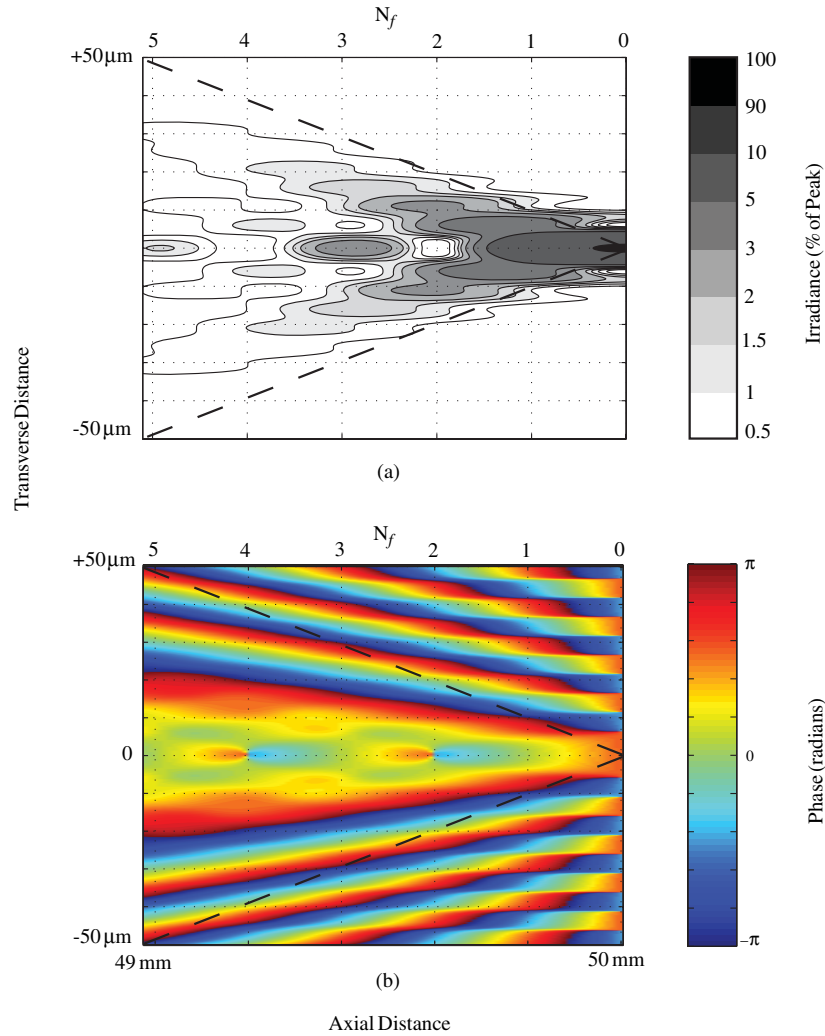


Fig. 5.38. yz -plane views of the irradiance (a) and phase (b) near focus for a laser beam converging through an aperture. $\lambda = 0.5 \mu\text{m}$, $f = 50 \text{ mm}$ and $a = 2.5 \text{ mm}$. The phase in (b) is the difference between the diffracted field and a converging spherical wave. The marginal rays of the focus cone are shown as dashed lines.

5.3.3 Poisson's Spot

A curious phenomenon occurs when a plane wave illuminates a circular disk, as shown in Fig. 5.39. The axial irradiance behind the disk is of interest. At first, one might expect a dark shadow, or perhaps an oscillatory behavior, like that observed with the circular aperture in Fig. 5.34. However, neither of these expectations is correct. Instead, a bright spot called *Poisson's Spot* is observed on axis. The correct analysis involves application of Babinet's Principle.

Applied to this geometry, Babinet's Principle and the resulting aperture

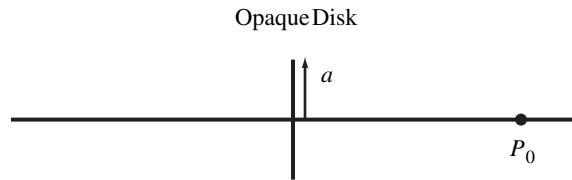


Fig. 5.39. An opaque disk is illuminated by laser beam U_s^- . The observation point P_0 is on axis a distance z_0 from the disk

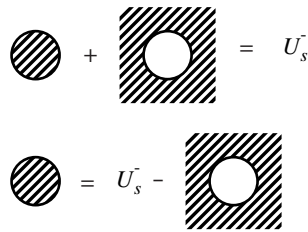


Fig. 5.40. Aperture algebra for solution of the opaque-disk problem. First, the complementary aperture (a circular aperture of the same radius as the disk) is added to equal the light in the illuminating wave. Then, the light from the circular aperture is subtracted from each side.

algebra take the form shown in Fig. 5.40. At the observation plane, addition of the field passed around the disk $U_{\text{disk}}(P_0)$ and the field passing through a circular aperture of the same diameter $U_{\text{aperture}}(P_0)$ is equal to the field of the illuminating wave U_s^- propagated to the observation point, which is $U_\infty(P_0)$ from Eq. (5.103). If the field from the circular aperture is subtracted from each side, the result is the field from the disk. Notice that the linear operation of propagation to the observation plane is assumed in the development. Mathematically, aperture algebra translates into

$$\begin{aligned}
 U_{\text{disk}}(P_0) &= U_{\infty}(P_0) - U_{\text{aperture}}(P_0) \\
 &= U_{\infty}(P_0) - U_{\infty}(P_0)(1 - e^{j\pi N_f}) \\
 &= U_{\infty}(P_0)e^{j\pi N_f},
 \end{aligned}
 \tag{5.112}$$

where Eq. (5.109) is substituted for $U_{\text{aperture}}(P_0)$. The axial field is simply the propagated illumination wave multiplied by a phase term. The on-axis irradiance is

$$I_{\text{disk}}(P_0) = I_{\infty}(P_0), \tag{5.113}$$

because the irradiance does not depend on the phase term. Surprisingly, the axial irradiance with the disk has the same value as the irradiance without the disk. Note that the axial irradiance does not oscillate, as observed with the circular aperture. Experimentally, Poisson's spot looks like the experimental result shown in Fig. 5.41. If a plane wave illuminates the disk, the axial irradiance behind the disk is not a function of z .

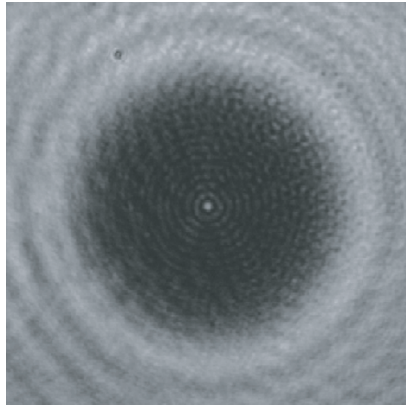


Fig. 5.41. An experimentally observed Poisson's spot.

A famous story involves the mathematicians Poisson and Fresnel in the 19th century. When Fresnel presented his new wave theory before Poisson, Poisson discounted the theory as impossible, because it would predict a bright spot behind an opaque disk. Of course, the experiment showed exactly this result. Thereafter, the phenomenon was known as Poisson's spot.¹

1. Poisson's spot is commonly called the spot of Arago, who was the scientist in charge of the experiment.

In practice, the axial irradiance approaches zero at distances very close to the aperture, due to the limited extent of the illuminating wave and the approximations used in the development of Eq. (5.109). However, Poisson's spot is easily observed with a high-quality ball bearing carefully attached to a microscope slide. The disk can also be used to form an image of a small, quasimonochromatic object, as shown in Fig. 5.42.

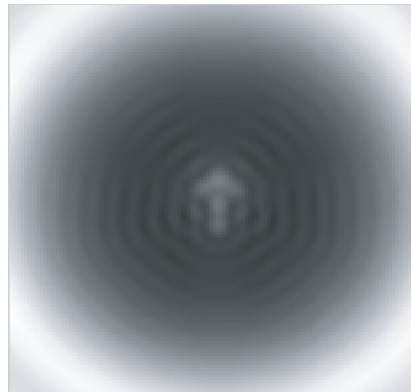


Fig. 5.42. An image of an arrow created with an opaque disk.

5.3.4 Fresnel Zone Plates

The *Fresnel zone plate* is a useful example of how a diffractive object can exhibit lens-like properties. An understanding of the Fresnel zone plate provides a foundation for understanding more sophisticated structures that can be used for optical testing, holography, wavelength compensation and diffractive optical elements. In this section, we first provide a historical perspective of zone-plate analysis, because reference literature is sometimes incomplete. Next, we extend the geometrical technique of Section 5.3.2 to understand on-axis and off-axis zone-plate behavior. The analysis is based on a binary-amplitude Fresnel zone plate, which is the easiest type of zone plate to understand. Then, the geometric lens properties of a zone plate are explained. Several other types of zone plates are described, along with a few applications.

5.3.4.1 Historical perspective of zone-plate analysis

The Fresnel zone plate has been a subject of study for over 130 years. It is a unique device that can be used to demonstrate near-field and far-field diffraction, both analytically and experimentally. Because of its focusing properties, the geometrical optics aspects are neatly embodied as well. In fact, R.W. Wood [Wood, 1988, p.38] made a low-power telescope using two simple zone plates, one

drawn by hand, and the other made with a reduced-size photograph. Its diffraction effects provide a useful tool to demonstrate a variety of approximations to diffraction equations. Due to its rotational grating structure, the Fresnel zone plate exhibits multiple axial foci. Many papers have been written describing the axial properties, including an insightful paper published by Zapata-Rodriguez *et al.* [Zapata-Rodriguez, 1999] However, the axial irradiance is often confused with the power contained within the first Airy disc. The axial irradiance is roughly constant for each of these foci, while the power drops off as the square of the diffraction order. The most widely referenced work which correctly describes the axial behavior within the Fresnel approximation was published by Boivin [Boivin, 1952], though a more elegant solution was proposed by C.T. Lane [Lane, 1930]. Another excellent resource is *Optical Data Processing* by Shulman [Shulman, 1970], where an entire chapter is devoted to the basic analysis of zone-plate theory.

5.3.4.2 Axial irradiance of the Binary Amplitude Fresnel Zone Plate

The binary amplitude Fresnel zone plate is usually constructed on a transparent glass plate, where an opaque coating is applied to block either the odd or even Fresnel zones. In Fig. 5.43, an opaque zone-plate mask is applied to the even zones of an aperture with $N_f = 5$, where the even zones are colored gray. If we follow the interference argument of Section 5.3.2.2, light from differential area δA_1 near the axis no longer interferes destructively with a small differential area δA_2 at the boundary between the first and second zones, because it is blocked by the mask. Instead, light from δA_1 interferes *constructively* with light from differential area δA_3 , which is at the boundary between the second and third zones, because $\text{OPD} = \lambda$ between them. Likewise, light from all differential areas inside the first zone interfere constructively with light from corresponding $\text{OPD} = \lambda$ differential areas in the third zone. Therefore, the on-axis irradiance results from a *coherent addition* of light from all open zones. If the first zone contributes $4I_\infty(P)$ to the on-axis irradiance, an aperture and mask with N_{FZ} open Fresnel zones produces a total on-axis irradiance of

$$I_0(P_0) = 4N_{\text{FZ}}^2 I_\infty(P_0). \quad (5.114)$$

As the number N_{FZ} of open Fresnel zones increases, the on-axis irradiance increases quadratically. Notice that the same result is achieved if the odd zones in the mask are blocked.

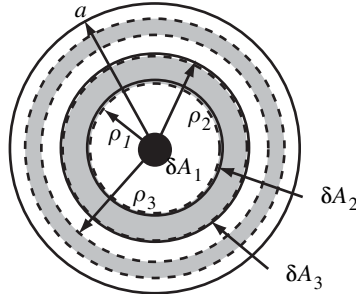


Fig. 5.43. A Fresnel-zone-plate mask in the aperture for a $N_f = 5$ system. Opaque annular regions are placed over the even zones. The number of open zones is $N_{FZ} = 3$. Light from differential area δA_1 is not canceled by light from differential area δA_2 , like in Fig. 5.27. Instead, since light from the even zones is blocked, the axial irradiance is nine times higher than without the mask.

Example 5.7: Relative on-axis irradiance of a Fresnel zone plate with ten open zones.

For a Fresnel zone plate with $N_{FZ} = 10$, the on-axis irradiance is $I_0(P_0) = 10^2 \times 4I_\infty(P_0) = 400I_0(P_0)$. Compared to an open aperture the same diameter as the zone plate, the zone plate on-axis irradiance is $400/4 = 100$ times more intense.

The intense light spot created with a zone plate occurs at the specific observation distance z_0 corresponding to one Fresnel zone overlapping each open area of the mask. Now, consider decreasing the observation distance by moving the observation point toward the plate to P'_0 , as shown in Fig. 5.44, which shows a side view of the blocked zones and the central open area of the mask. We analyze the on-axis behavior by examining the first open zone in detail and extrapolating the results to include the other zones.

Figure 5.44 also shows the location of the new zone boundary for the first Fresnel zone, where the new boundary ρ'_1 between the first and second zones falls within the radius a_{ZP1} of the first open area. Now, some of the light from the first Fresnel zone interferes destructively with light from the second Fresnel zone, so the on-axis irradiance contribution from the first open area decreases. When the boundary between the second and third Fresnel zones is at radius a_{ZP1} , the light from the first and second zones cancel, and zero irradiance contribution (a dark spot) is observed on axis. Further movement of the observation point toward the plate introduces a portion of the third zone inside radius a_{ZP1} , so the on-axis irradiance contribution increases. The irradiance contribution increases until it reaches a maximum where the boundary between the third and fourth Fresnel

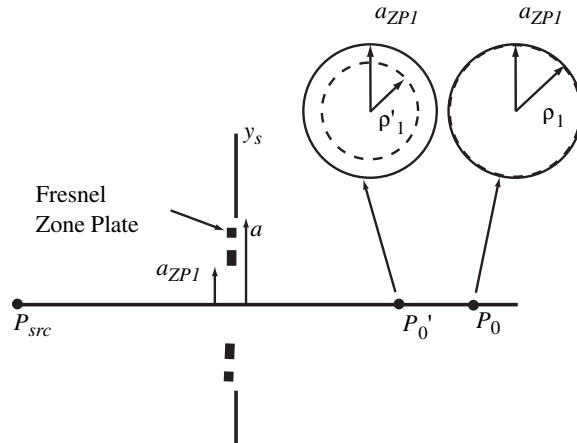


Fig. 5.44. Analysis of a Fresnel zone plate is accomplished by understanding the behavior of light diffracted from the first open area of the mask. If the observation point is moved toward the aperture, some of the light from the next even Fresnel zone enters through it, thereby reducing the irradiance. Similar behavior is exhibited from other open areas, so thereby reducing the total irradiance at the observation point.

zones is at radius a_{ZP1} . The irradiance contribution at this secondary maximum, due to the area inside radius a_{ZP1} only, is $4I_{\infty}(P_0)$. This pattern of alternating maxima and zeros in the on-axis irradiance contribution continues as the observation point is moved toward the plate. Notice that the axial position and irradiance values of the maxima and zeros created by the first open area of the Fresnel zone plate are identical to the maxima and zeros observed from an aperture with radius a_{ZP1} .

The same behavior is observed for each of the open areas in the Fresnel zone plate. That is, when the first open area contains two Fresnel zones, so will the next open area. Therefore, *the axial irradiance observed due to diffraction from a Fresnel zone plate exhibits alternating maxima and zeros, where the locations of the maxima and minima are identical to those observed from an open aperture equal in radius to the radius of the first area of the zone plate. The intense on-axis irradiance at each maximum is given by Eq. (5.114).*

Notice that collimated illumination with $z_{src} \rightarrow \infty$ exhibits $I(P_0)$ that does not vary with observation position, so *a Fresnel zone plate illuminated with collimated laser light produces equally intense maxima along the axis.*

Mathematically, Eq. (5.109) can be applied to the problem of calculating the on-axis irradiance by separating the integral into a sum of contributions from transmitted zones. That is,

$$\begin{aligned}
U_0(P_0) &= -j\pi U_\infty(P_0) \int_0^{N_f} e^{j\pi q} dq \\
&= -j\pi U_\infty(P_0) \left(\int_0^1 e^{j\pi q} dq + \int_2^3 e^{j\pi q} dq + \int_{N_f-1}^{N_f} e^{j\pi q} dq \right) \\
&= [2U_\infty(P_0)N_f]/2 \\
&= 2U_\infty(P_0)N_{FZ}, \tag{5.115}
\end{aligned}$$

and

$$I_0(P_0) = 4N_{FZ}^2 I_\infty(P_0) \tag{5.116}$$

which is identical to the result obtained from the intuitive argument leading to Eq. (5.114).

5.3.4.3 Off-axis behavior of the Binary Amplitude Fresnel Zone Plate

The off-axis behavior of a Fresnel zone plate can be understood by following an analysis similar to Section 5.3.2.3. For each off-axis observation position, the shifted Fresnel zone pattern is calculated based on OPD. The transmitted light through the zone-plate mask is divided into odd and even zones. The total geometrical area of each zone type is subtracted, and the absolute value of the result is proportional to the irradiance of the light at the off-axis observation position. Figure 5.45 displays the transmitted zone pattern at several off-axis observation positions for a zone plate with $N_{FZ} = 3$. Other parameters of the simulation are $a = 0.5$ mm, $\lambda = 0.5$ μ m, $z_{src} \rightarrow \infty$ and $a_{ZP1} = 0.2231$ mm. The on-axis observation reference is the point where the first Fresnel zone fills the first open area of the zone plate at $z_0 = 100$ mm. Notice that, unlike the case of the open aperture, equalization of the transmitted zone areas occurs very quickly as the observation point is moved off axis. The irradiance profile shown in Fig. 5.45 shows that the energy is tightly confined around the axis, with a width w of approximately

$$w = \frac{\lambda z_0}{a} = 0.1 \text{ mm.} \tag{5.117}$$

When the zone plate is illuminated with collimated laser light, w is equal to the width of the outermost Fresnel zone. That is, when the observation point is moved such that the outermost zone passes through the outermost open area of the mask, the irradiance is almost at a minimum. Except for minor sidelobes, the background irradiance is insignificant compared to the on-axis irradiance. Results of a calcu-

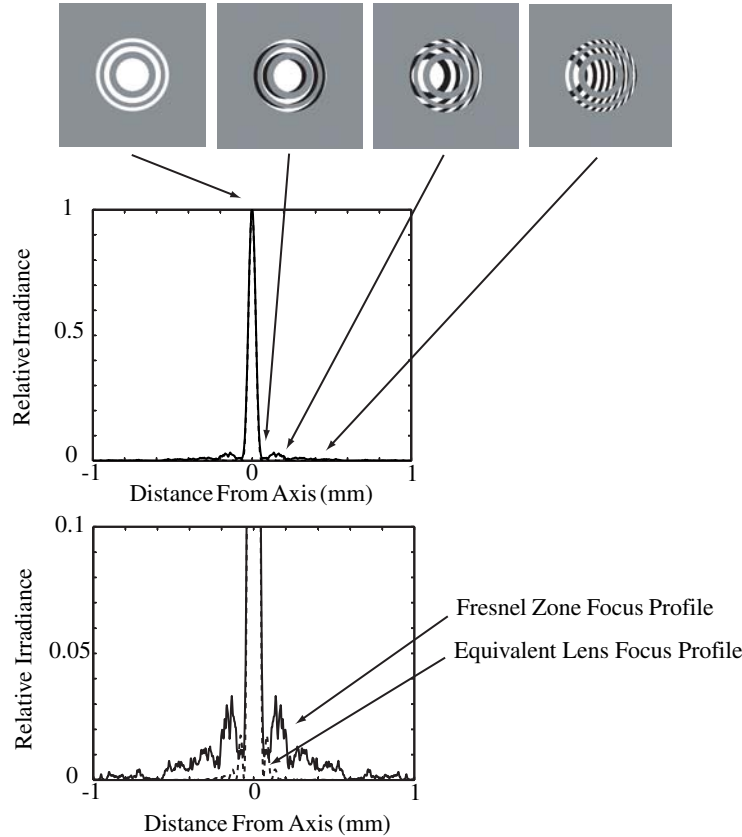


Fig. 5.45. The transmitted zone pattern and irradiance profile at several off-axis positions for the zone plate with $N_{FZ} = 3$. The irradiance is tightly confined around the optical axis. A lower irradiance background is also observed.

lation showing an enhanced background are displayed as an image in Fig. 5.46, where the central peak is 1160 times more intense than the brightest rings shown in the figure. Therefore, *the Fresnel zone plate produces a tightly confined region of high irradiance near the axis at the observation point corresponding to where the first Fresnel zone fills the first area of the zone plate. A lower irradiance background is also observed.*

Similar behavior is observed in transverse planes corresponding to the other axial maxima, except the w decreases in Eq. (5.117) due to the smaller value of z_0 . In regions where zero on-axis irradiance is observed, only the background irradiance is present, as shown in Fig. 5.47.

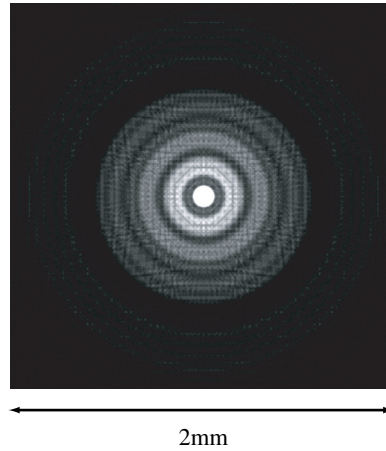


Fig. 5.46. An enhanced image of the zone-plate focus for the irradiance shown in Fig. 5.45, where the central peak is saturated. The lower irradiance background is circularly symmetric.

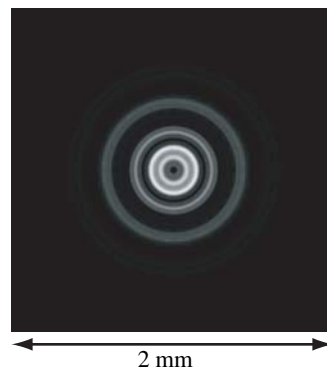


Fig. 5.47. In observation planes away from the focus points, only the lower-irradiance background is observed.

5.3.4.4 The Fresnel zone plate as a geometric lens

The concentration of energy from the Fresnel zone plate is similar to the action of a refractive positive lens, in that a point source is imaged into a point-like image. The position of the point-like image is the same as if an ideal thin lens with focal length $f = L$ replaced the zone-plate mask. That is, the first-order lens-law relationship is

$$\frac{1}{f} = \frac{1}{z_{src}} + \frac{1}{z_0}, \quad (5.118)$$

where z_{src} and z_0 are object and image distances, respectively, and L defined in Section 5.3.1.1 can be written as

$$\frac{1}{L} = \frac{1}{z_{src}} + \frac{1}{z_0}, \quad (5.119)$$

where z_{src} and z_0 are the source and observation distances, respectively. If we set L to correspond with the first axial maximum behind the zone plate,

$$L = \frac{a_{ZP1}^2}{\lambda}. \quad (5.120)$$

Changes in the source position z_{src} will affect the ‘image’ position of the first axial maximum in the same way as an ideal thin lens. Therefore, *the position of the first maximum of the zone-plate diffraction pattern corresponds to the position of the image from a thin lens with $f = a_{ZP1}^2/\lambda$.*

Interestingly, the size of the tightly confined pattern from the zone plate is equivalent to what is observed with a thin lens. As shown in the analysis of Fraunhofer diffraction patterns in Section 5.4.3, the width of the Airy disc between zeros in a thin-lens system is

$$w = 1.22 \frac{\lambda z_0}{a}. \quad (5.121)$$

Profiles of diffraction patterns from an equivalent thin lens and a Fresnel zone plate are shown in Fig. 5.45. The patterns are essentially identical. This diffraction pattern is the major component of the *point-source response* of the Fresnel zone plate at the first axial maximum. The second part of the point-source response is the small background irradiance. Therefore, *the point-source responses of an equivalent thin lens and a Fresnel zone plate are nearly identical, except for a small background observed with the zone plate.*

If we consider a single off-axis source point used with a Fresnel zone plate, the diffraction pattern at the observation plane will also shift, due to the shift of the Fresnel zone pattern in the aperture. The same tightly confined pattern and background are observed, except at an off-axis position. A simple geometrical argument can be used to show that the observation-plane shift due to an off-axis source point is the same as what would be observed with an equivalent thin lens. Since the Fresnel zone plate is a linear device, *a collection of source points will be imaged onto the plane corresponding to the first maximum in the same way as an equivalent thin lens.*

At this point in the analysis, we understand that *a Fresnel zone plate forms a lens-like image at the axial location of the first maximum.* If the observation point

is moved toward the plate such that the axial position corresponds to the next maximum at $N_f = 3$ inside radius a_{ZP1} , another lens-like image is formed, but with reduced magnification. A third image is found where $N_f = 5$ inside radius a_{ZP1} , and so on. The image positions correspond to a collection of equivalent thin lenses, where

$$f_n = \frac{a_{ZP1}^2}{n\lambda}, \quad (5.122)$$

and $n = 1, 3, 5, \dots$ corresponds to the odd-integer number of Fresnel zones inside radius a_{ZP1} . The presence of additional zone-plate foci suggests that the background irradiance observed at a focal point is the combination of out-of-focus images from the other focal points. The out-of-focus images spread their optical power over a much wider area than the concentrated focus energy, so the background irradiance is low. These multiple focal points are called the *positive diffracted orders* of the zone plate, as shown in Fig. 5.48.

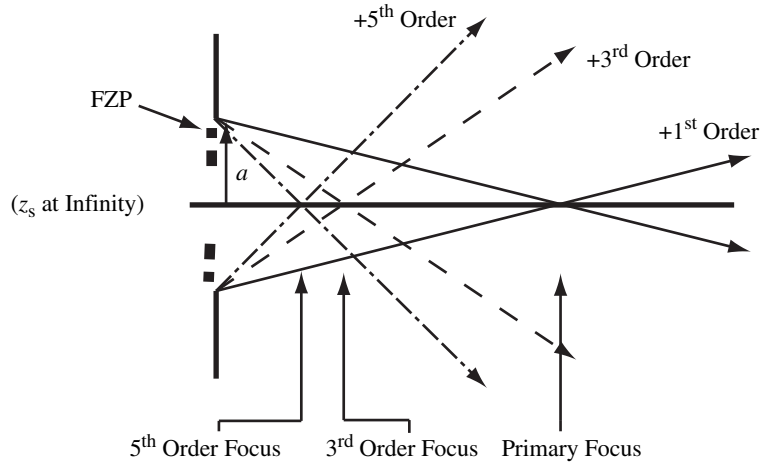


Fig. 5.48. Multiple focus points are created by the positive diffracted orders of the zone plate.

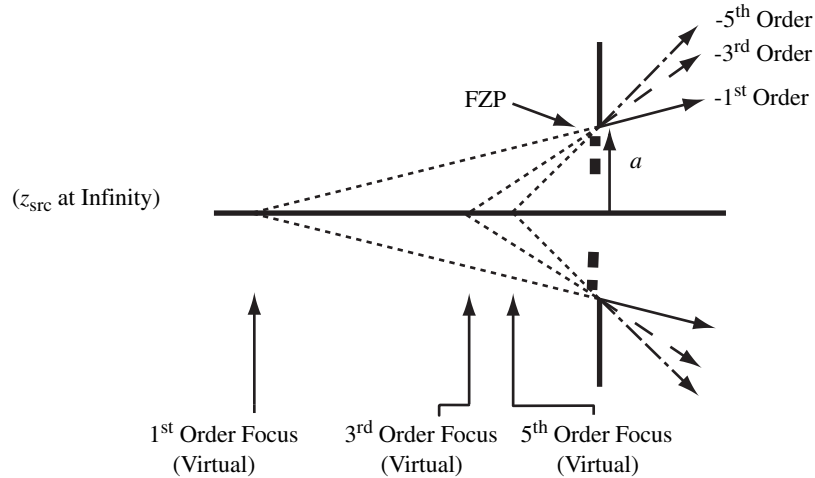


Fig. 5.49. Multiple virtual focus points are created by the negative diffracted orders of the zone plate.

Notice that an observation point on the left-hand side of the zone plate can also produce an effective focal point with negative n . Although not directly observable without an auxiliary lens system, each negative effective focal point produces diverging light in the same way as a negative lens. These multiple focal points are called the *negative diffracted orders* of the zone plate, as shown in Fig. 5.50. Therefore, *the Fresnel zone plate produces positive and negative diffracted orders, which act as a collection of positive and negative lenses with focal lengths given by Eq. (5.122), where $n = \pm 1, \pm 3, \pm 5, \dots$* A pictorial representation of this phenomenon is displayed in Fig. 5.51.

A curious effect occurs at the positive foci of a zone plate illuminated with collimated laser light. According to Eq. (5.114), the axial irradiance at each focus should be equal. In fact, the axial irradiances are equal, as shown in Fig. 5.52, but the width of each focal spot gets smaller as the order increases. The decreasing width can be interpreted as due to the increasing numerical aperture a/z_0 of the focus cones in Eq. (5.117). Therefore, laboratory experiments meant to measure the axial value must be careful to use an extremely small pinhole, so that the encircled energy, which decreases due to the width of the spot, is not mistaken for the axial value.

5.3.4.5 Other types of Fresnel zone plates

Variations of the binary amplitude Fresnel zone plate have certain advantages. For example, instead of blocking the even zones, the binary phase zone plate uses a modified glass plate, where the even zones are delayed or advanced in phase by π compared to the odd zones. The modification can be accomplished by etching a

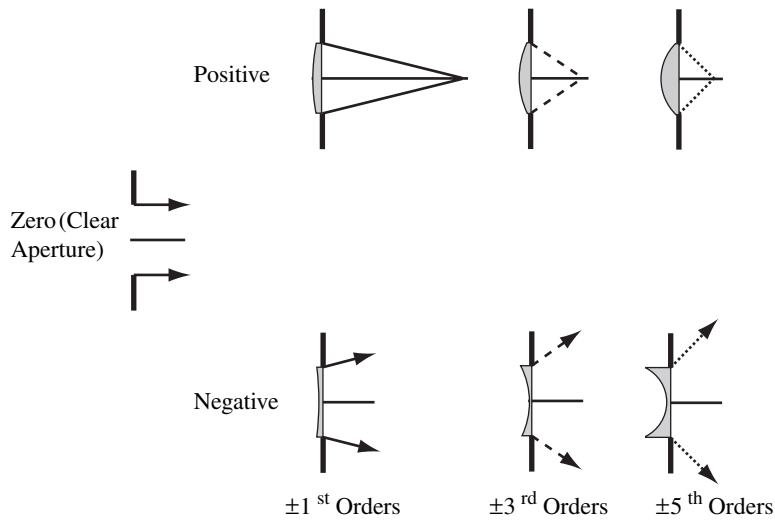


Fig. 5.50. The Fresnel zone plate can be thought of as a collection of positive and negative lenses, each of which produces a focus.

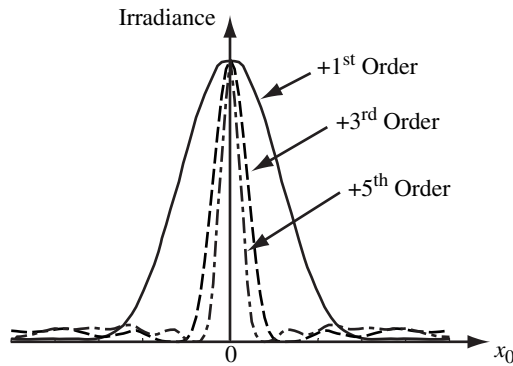


Fig. 5.51. Irradiance distribution of the light near several zone-plate foci when the plate is illuminated with collimated laser light. Notice that the peak irradiance at each focus is constant, but the width decreases for foci closer to the plate.

flat glass plate to the appropriate depth in the areas corresponding to even Fresnel zones. Alternatively, photoresist can be deposited, exposed and developed in the appropriate pattern. This modification results in double the axial field amplitude given by Eq. (5.115) and an irradiance amplification at the focal points by a factor

of four. A problem with the binary phase zone plate is that it is wavelength sensitive.

One disadvantage of the Fresnel zone plate is that it creates several positive and negative diffracted orders. If only one focus point is desired, the light directed to other foci reduces the efficiency of the lens. One way to increase the diffraction efficiency is to make smooth phase transitions in each zone, rather than an abrupt discontinuity at a zone boundary. A picture of a smooth-zone diffractive lens is shown in Fig. 5.52.

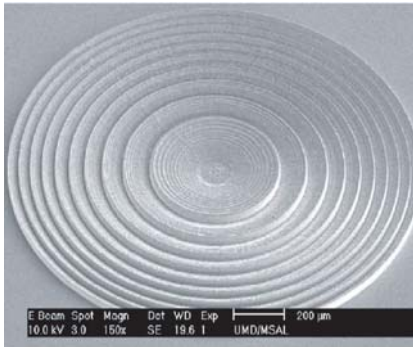


Fig. 5.52. A scanning electron microscope picture of a smooth-zone diffractive lens that was fabricated for use with an x-ray microscope.

5.3.5 Edge diffraction

5.3.6