1

Foundations of the Diffraction Integral Method

This chapter introduces some fundamental electromagnetic concepts that can be obtained reasonably quickly from Maxwell's equations. These concepts will be useful in understanding much of the material on optics presented later in the text. Since Maxwell's equations form the basis of all classical optical phenomena, it is worthwhile spending some time with these relations to see how they can be applied in understanding optics.

Maxwell's equations define the field of study known as classical electromagnetics. This field unfortunately excludes many types of active optical phenomena and devices, including lasers and doped fiberoptic amplifiers. On the other hand, it still includes a vast array of theory and technology in optics. In fact, classical optics is even used in at least one aspect of laser design, that is, in the design of the resonator mirrors and external diffraction gratings used to confine the optical fields within the active laser medium.

Historically, the study of optics has not been approached from the point of view of Maxwell's equations and general electromagnetic theory. Many optical instruments were developed long before the consolidation of Maxwell's equations into their current form, so traditional means of understanding optical phenomena have developed in rather ad hoc ways. For example, the concept of *light rays* is quite ancient and certainly predates the modern understanding of ray theory as an asymptotic, high-frequency limit of Maxwell's equations. The concept of *Huygens sources* is also a historical notion, having since been replaced by the much more precisely defined concept of *Green's functions* (which is presented later in this chapter).

In this book, optical theory and technology will not be presented in the traditional, historical fashion, not because this approach is ineffective, but because it is simply not the best way to examine the subject. In addition, the historical concepts are rarely used today in the technical literature—in either optics or general electromagnetics. They have been replaced by something better, and that is a vector formulation of classical optics based on Maxwell's equations. This formulation is the subject of the present chapter.

1.1 MAXWELL'S EQUATIONS

For the purposes of this book then, Maxwell's equations will be regarded as being *given.* Every electromagnetic analyst knows Maxwell's equations by heart. If the reader hasn't yet committed them to memory, they're repeated below for reference. Faraday's law is

$$\nabla \times E = -\frac{\partial B}{\partial t} \tag{1.1}$$

Ampère's law is

$$\nabla \times H = \frac{\partial D}{\partial t} + J \tag{1.2}$$

Gauss' law for electricity is

$$\nabla \cdot D = \rho \tag{1.3}$$

Gauss' law for magnetism is

$$\nabla \cdot H = 0 \tag{1.4}$$

The continuity equation is

$$\nabla \cdot J = -\frac{\partial \rho}{\partial t} \tag{1.5}$$

and the constitutive (material) equations are

$$D = \epsilon E, \qquad B = \mu H \tag{1.6}$$

In the equations above,

E = electric field intensity (voltage/distance)

D = electric displacement (charge/area)

H = magnetic field intensity (current/distance)

B = magnetic induction (voltage/velocity/distance)

J = volume current density (current/area)

ρ = volume charge density (charge/volume)

ϵ, μ = permittivity, permeability tensors (capacitance/distance, inductance/distance, respectively)

Some readers may not be familiar with the continuity equation (1.5). This equation is shown in its common differential form, but Gauss' law allows it to be rewritten in integral form as

$$\iint J \cdot dS = -\frac{dQ}{dt}$$

This version of the equation is obtained almost immediately from the definition of electrical current, which states that current is equal to the quantity of electrical charge

passing across (i.e., passing *normal to*) a given plane in a given amount of time. If we imagine a cubic volume in three dimensions, with a normal component of current passing through each of its six faces, we may apply the definition of electrical current to all three sets of opposing faces of the cube. Doing so yields the integral form of the continuity equation above, namely, that the rate of charge accumulation between all three sets of opposing faces is equal to the integral of the normal component of current over all six faces.

The two field quantities that have definite physical significance are the electric field strength E (measured in voltage/distance) and the magnetic induction B (measured in terms of voltage/velocity/distance). These two fields exert a definite *mechanical* force on a charged particle (of charge q) that is given by the relation

$$F = q(E + v \times B) \tag{1.7}$$

where

$$v = \text{particle velocity vector}$$

The other two field quantities (D, H) are related to E, B via the material parameters (permittivity, ϵ and permeability, μ). D is a quantity that is related to electrical charge, and H is related to electrical current, both independent of the particular medium involved. Thus, both quantities are related to electrical charge, D to *static* charge and H to charge in *motion.*

In this book we'll be dealing with fields that are harmonic in time (single-frequency sinusoids having infinite temporal duration). This assumption has been implicit in the study of optical phenomena literally for centuries, yet only within the last three decades have laser light sources been available which can produce this type of *coherent radiation* in the optical regime. This has enabled longstanding mathematical models of optical fields to finally be in step with the real phenomena.

The assumption of sinusoidal time variation is generally not considered that restrictive since an arbitrary time-varying field may be expressed in terms of a Fourier transform superposition of single-frequency sinusoids. Thus, field solutions at many frequencies may (in principle) be combined to yield the solution for a complex time-dependent field. (In practice, such a superposition is rarely used to solve time-domain problems.)

Under the assumption of time-harmonic fields, we may write (for the electric field)

$$E(x, y, z, t) = E(x, y, z)\, e^{j\omega t}$$

In addition to specifying time-harmonic fields, we'll also limit ourselves to isotropic (scalar) media, for which the permittivity and permittivity tensors are simple scalar quantities. Under the joint assumptions of time-harmonic fields and scalar media, Maxwell's equations become (see Appendix B):

$$\nabla \times E = -j\omega\mu H \tag{1.8}$$

$$\nabla \times H = j\omega\epsilon E + J \tag{1.9}$$

$$\nabla \cdot E = \rho/\epsilon \tag{1.10}$$

$$\nabla \cdot H = 0 \tag{1.11}$$

$$\nabla \cdot J = -j\omega\rho \tag{1.12}$$

$$D = \epsilon E, \qquad B = \mu H \tag{1.13}$$

This is the form of Maxwell's equations that we'll use in this book. It is the form most often used in optics and electromagnetics. The fields in this format are no longer purely real quantities, and they are certainly no longer temporal quantities. They may now take on complex number values. These complex field values are known as *phasor* quantities, inasmuch as it is the *phase* of these quantities that distinguishes them from ordinary real-valued oscillatory temporal field quantities.

The magnitude part of the phasor is related to the energy density of the field (as in the non-time-harmonic case). The phase part of the phasor carries the temporal information (though in a somewhat coded form). *Time shifts* are translated into *phase shifts* (phase retardation for time delay and phase advance for forward time shifts). Time advances and delays may take on any numerical values, whereas phase shifts may only occupy the range from 0 to 2π.

For readers who are not already familiar with phasor quantities, we may readily relate time functions with their phasor counterparts. Let

$$E(t) = E(r, t), \text{ for some fixed point, } r$$

We'll use the Fourier transform pairs

$$E(t) = \int_{-\infty}^{\infty} E(\omega)\, e^{j\omega t}\, d\omega \tag{1.14a}$$

$$E(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} E(t)\, e^{-j\omega t}\, dt \tag{1.14b}$$

to relate the temporal and spectral representations of the field. According to these representations, the temporal dependence

$$E(t) = e^{j\omega_0 t} \tag{1.15a}$$

is given in the spectral domain as

$$E(\omega) = \delta(\omega - \omega_0) \tag{1.15b}$$

where we've used the identity below, which is obtained in Appendix B on Fourier analysis.

$$\delta(\omega - \omega_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{j(\omega-\omega_0)t}\, dt$$

A time-delayed field of the form

$$E(t - t_0)$$

will have a phasor (spectral) form obtained via Eq. (1.14b) as

$$E'(\omega) = e^{-j\omega t_0}\, E(\omega)$$

where $E(\omega)$ is given by (1.15). Thus, a time-delayed field produces a linearly phase-retarded spectrum. This is one way in which temporal information is *phase-coded* in the phasor domain. That is, a linear *phase* in the spectral domain relates to either a time delay or an advance. (In causal systems, only time *delays*—or phase lags—are permitted.) It's interesting to note that in electrical circuit theory, "linear-phase," or "constant group

delay'' filters are often designed for the purpose of producing a desired time delay for a signal consisting of many frequency components.

The reader may also verify directly from (1.14a) (using the product rule of integration, along with an assumption that $E(t) \rightarrow 0$ as $t \rightarrow \pm\infty$) that derivatives in the time domain correspond to multiplication by $j\omega$ in the spectral domain. This is one more way in which time-domain information is coded in the spectral domain via the phase of the phasor spectrum. Other relationships between time functions and their spectra are given in Appendix B.

1.2 THE WAVE EQUATIONS

The first step in mathematical refinement up from Maxwell's equations is given by the wave equations. Whereas Maxwell's equations contain the electric and magnetic field vectors all coupled together in an inconvenient fashion, the wave equations are a set of equations in which the electric and magnetic fields have been separated, or *decoupled.* This decoupling feature allows the electric and magnetic fields to be solved quickly in terms of the currents (a process we'll carry out in the next session).

The electric field wave equation is derived in about three lines. Take the curl of Faraday's law (1.8) and substitute Ampère's law (1.9) into it to obtain

$$\nabla \times \nabla \times E - \omega^2 \mu \epsilon E = -j\omega\mu J \tag{1.16}$$

It will be convenient for our purposes to remove the double curl operator using the following vector identity. (Later, we'll talk more about the meaning of the second term on the right-hand side.)

$$\nabla \times \nabla \times E = \nabla(\nabla \cdot E) - \nabla^2 E \tag{1.17}$$

to obtain

$$\nabla^2 E + k^2 E = j\omega\mu J + \frac{1}{\epsilon}\nabla\rho \tag{1.18}$$

where

$$k^2 = \omega^2 \mu \epsilon$$

and we've used Gauss' law (1.10) to obtain (1.18). Equation (1.18) is the electric field wave equation we sought. Note that only the electric field is present in this equation on the left-hand side. A similar process may be used to obtain a wave equation for the magnetic field. This is given as

$$\nabla^2 H + k^2 H = -\nabla \times J \tag{1.19}$$

The wave equations are linear partial differential equations for the fields. The forcing functions for the equations are the currents and charges. [*Note:* We could have expressed (1.18) entirely in terms of the electric current, by invoking the continuity condition (1.12).]

Equation (1.17) may be regarded as the defining equation for the *vector Laplacian* [it's the second term on the RHS of (1.17)]. According to (1.17), the vector Laplacian is related to the ''curl curl'' operator and the ''gradient divergence'' operator. The vector

Laplacian defined by (1.17) is of little use in most coordinate systems. It is really useful only in rectangular coordinates, where it takes on an exceedingly simple form (see Appendix A). In the next section, we'll make use of the simple form of the vector Laplacian in rectangular coordinates to show how the wave equations above may be solved quickly.

1.3 SCALAR AND VECTOR POTENTIALS

In this section, we'll solve the two wave equations and in so doing introduce the concept of the scalar and vector potentials. Most readers are undoubtedly familiar with the ordinary scalar potential from electrostatic theory. In electrodynamics, however, there is a second potential—the vector potential—which takes on far greater importance than the scalar potential. This is because fields generated via the vector potential may radiate in space, whereas the fields generated via the scalar potential are "bound" to electrical charges and cannot radiate.

The reader is probably familiar with Huygens' principle from elementary optics. This principle provides a graphical means for determining the field produced by an optical wavefront. According to Huygens' principle, a wavefront may be divided up into an infinite number of point source radiators. Each of these point radiators may be regarded as a source of spherical waves; therefore, the radiated field may be determined graphically by moving a compass along the wavefront and drawing circles about a series of closely spaced points on the front. The intersections of the various circles so drawn are tangent to the new wavefront and therefore indicate its contour.

In this section, we're going to do the same sort of thing mathematically—it's really not that hard to do—in order to obtain a mathematical version of Huygens' graphical construction. The mathematical version we'll be looking at is based on the same principle as Huygens' construction. That is, we'll regard the source distribution as being composed of an infinite number of point source radiators and then sum up (actually, *integrate*) the contributions due to this distribution of point sources.

In retrospect, Huygens' discovery of this principle (in 1678!) is quite remarkable. Today, this same idea is used in numerous branches of engineering and physics (from electromagnetic scattering to digital signal processing), though now it generally goes by the name, *the superposition principle.* The concept simply states that a time-varying signal (or spatial distribution of charge, mass, etc.) may be regarded as the superposition integral of an infinite number in infinitesimally short impulses (or, for spatial distributions, point sources). And if the response of an electronic circuit to one impulse (or the response of an electromagnetic system to a single point charge, or a mechanical system to a point mass) is known, then it is possible (in principle) to integrate over the entire source distribution to get the total response of the system (due to the entire distribution). This is the principle we'll explore in this section.

The principle of superposition applies only in connection with *linear* differential equations. (Both the wave equations above are linear, since the fields only appear to the first power, e.g., there are no terms of the form E^2 present.) We'll describe the superposition principle briefly here (in connection with time signals) before applying it to the electromagnetic wave equations.

Say we want to find the solution to the linear differential equation

$$\frac{d^2}{dt^2} S(t) + k^2 S(t) = f(t) \tag{1.20}$$

where $S(t)$ is some unknown signal response function and $f(t)$ is a specified forcing function. By the sifting property of the impulse function, we may just as well write $f(t)$ in the form

$$f(t) = \int_{-\infty}^{\infty} f(t')\delta(t - t')\, dt' \tag{1.21}$$

that is, as a superposition of impulse functions, weighted by the value of the forcing function. With (1.21) in hand, it's now evident that all we have to do in order to solve (1.20) is to solve the equation

$$\frac{d^2}{dt^2} S(t) + k^2 S(t) = \delta(t - t') \tag{1.22}$$

and then add up all of the individual responses due to each of the various weighted impulse function inputs. So, say the solution to (1.22) is given by $h(t - t')$. This is called the *impulse response* of the differential operator on the LHS of (1.20). The solution to (1.20) is now just a matter of adding up all the impulse *responses* due to each of the impulse *inputs* in (1.21), that is,

$$S(t) = \int_{-\infty}^{t} f(t)h(t - t')\, dt' \tag{1.23}$$

So, all we have to do to solve (1.20) is to find the impulse response solution to (1.22). Once that is known, the response to an arbitrary forcing function may be obtained by superposition of impulse responses, using the same technique Huygens used. [*Note:* Huygens assumed that the field from a point source radiator—his impulse response—was a perfectly spherical wave. In this section, we'll verify that assumption.]

Note that the upper limit to the integral in (1.23) is set at t, since inputs after time t do not contribute to the output of the system at t. When the independent variables are spatial rather than time coordinates, a two-sided infinite range of integration may be employed.

With this brief introduction to the superposition principle, let's now apply it to the solution of the magnetic field wave equation (1.19). We begin by seeking a solution to the equation

$$\nabla^2 H + k^2 H = -\hat{u}\delta(r - r_0) \tag{1.24}$$

where

$$\hat{u} = \hat{x}, \hat{y}, \hat{z}$$

The delta function in three dimensions is defined by the usual equations

$$\delta(r - r_0) = \infty \qquad \text{when } r = r_0$$

$$= 0 \qquad \text{when } r \neq r_0$$

subject to the integrability condition (in three dimensions)

$$\iiint_{-\infty}^{\infty} \delta(r - r_0)\, dv = 1$$

This three-dimensional impulse function has the following sifting property in three dimensions (see Appendix D):

$$f(r_0) = \iiint_{-\infty}^{\infty} f(r)\delta(r - r_0)\, dv$$

Now, the delta function term on the RHS of (1.24) will give rise to fields that vary as a function of radius only, so this is the type of solution we'll look for. And since we know that the delta function is zero everywhere except at the source point, that is, at $r = r_0$, we can first find the radially symmetric solution to the homogeneous wave equation (whose forcing function is identically zero) to get a solution that is at least valid everywhere except at the source point. We'll solve (1.22) for the three rectangular coordinates of magnetic field separately. So, we'll now solve the scalar wave equation

$$\nabla^2 H_u + k^2 H_u = -\delta(r - r_0) \tag{1.25}$$

where

$$u = x, y, z$$

Equation (1.25) was obtained thanks to the very simple form of the vector Laplacian in rectangular coordinates, that is,

$$\nabla^2 H = \nabla^2 H_x \hat{x} + \nabla^2 H_y \hat{y} + \nabla^2 H_z \hat{z}$$

We'll take r_0 at the origin, so that

$$|r - r_0| = r$$

where r is the radial distance from the origin. The form of the scalar Laplacian in spherical coordinates may be used (see Appendix A) to obtain an explicit expression for the radially symmetric scalar Laplacian. Hence, the radially symmetric form of (1.25) is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial H_u}{\partial r}\right) + k^2 H_u = 0$$

With this, the radially symmetric solution to (1.25) for $r \neq 0$ may be verified to be

$$H_u(r) = A\frac{e^{-jkr}}{r} \tag{1.26}$$

where A is some unknown constant. The reader should verify this solution by substituting (1.26) into (1.25).

Equation (1.26) verifies that Huygens was indeed correct in taking a spherical wave as the field due to a point source radiator. (The phase fronts of this field are surfaces of constant radius, r; thus, the wave is referred to as a *spherical wave.*) This equation also shows, however, that this wave has not only spherical phase, but also an amplitude that varies inversely with distance from the source, due to the spreading of the spherical wave. This "one-over-r" dependence is the same as that of the static potential arising from a charged particle. In the static case, however, the potential varies inversely with radius, and the *field* varies inversely as the *square* of the radius. In this (dynamic) case, however, the actual field varies only inversely with distance, not inversely as the square of the distance. In the time-varying case, the fields extend outward much farther than in the static field case (i.e., they separate from their sources and are able to radiate to distant receivers).

It's worth taking a look at the solution to (1.25) at the origin. This is actually very easy. All we have to do is note that the scalar Laplacian in (1.25) is defined as the divergence of the gradient. So, if we integrate both sides of (1.25) over a small spherical ball (of radius, ϵ) about the origin, taking the incremental volume element in spherical coordinates as

$$dv = r^2 \sin\theta \, dr \, d\theta \, d\phi$$

we see that the second term on the LHS of (1.25) integrates to zero as ϵ tends to zero. The definition of the three-dimensional delta function above shows that the RHS integrates to -1. The first term on the LHS of (1.25) is integrated using the divergence theorem, to convert the integral over the spherical volume to an integral over the spherical surface. Then, taking the gradient of the Green's function in spherical coordinates (using the formula in Appendix A), along with the formula for the element of surface area,

$$ds = r^2 \sin\theta \, d\theta \, d\phi$$

it's readily shown that the integral of the first term is equal to -4π. So, the constant, A, in (1.27) is equal to $1/4\pi$.

The solution to (1.25),

$$\psi(r) = \frac{e^{-jkr}}{4\pi r} \tag{1.27}$$

is called the free-space *scalar Green's function.* To get the Green's function for the actual *field* (called the free-space *tensor Green's function*), we simply use the superposition principle (applied to all three Cartesian components of magnetic field) to get

$$H(r) = \iiint_{v'} \nabla' \times J(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} \, dv' \tag{1.28}$$

We obtained this equation by convolving the Green's function (1.27) with the forcing function from the RHS of (1.19). This is the solution to the magnetic field wave equation (1.19). It is also the tensor Green's function for the magnetic field. Using an entirely analogous procedure, we can show that the tensor Green's function for the electric field is

$$\begin{aligned} E(r) = & -jk\eta \iiint_{-\infty}^{\infty} J(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} \, dv' \\ & -\frac{1}{\epsilon} \iiint_{-\infty}^{\infty} \nabla' \rho(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} \, dv' \end{aligned} \tag{1.29}$$

where

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

is the characteristic impedance of the medium.

One small feature of these two equations may be unfamiliar to many readers. This is the concept of *source coordinates* and *field coordinates.* In Maxwell's equations, there is never any confusion as to the meaning of the vector differential operators. You just

pick a point in space and—say, in the case of Ampère's law—you calculate the curl of the magnetic field at that point and relate it to the current and electric field at that same point. Things are a little more complicated in (1.28) and (1.29), however. Because of the convolution (superposition) nature of the solution, there are currents and charges distributed over the "source coordinates," denoted with primes in (1.28) and (1.29), and there are the resulting fields that appear at the "field coordinates," denoted by unprimed coordinates. When we derived the solution to (1.25), we arbitrarily placed the origin at r' and calculated the scalar Green's function with respect to that origin. Now we have the situation where both r and r' are variable and the origin is located at some fixed point elsewhere.

When we take differential operators now, they can be with respect either to the source coordinates or field coordinates. (This is much different than was the case with Maxwell's equations.) When the operator is taken with respect to the source coordinates, the field point is assumed to be fixed—it is the temporary origin. When the operator is taken with respect to the field coordinates, all the source currents are assumed to be fixed in space and only the field point is variable.

In Eqs. (1.28) and (1.29), the differential operators are taken with respect to the source (primed) coordinates. However, the tensor Green's functions are traditionally expressed in terms of field region (unprimed) differential operators. Appendix C shows how the differential operators in (1.28) and (1.29) can be shifted onto the field coordinates. The resulting equations for the fields are

$$H(r) = \nabla \times \iiint_{v'} J(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} dv' \tag{1.30}$$

$$E(r) = -jk\eta \iiint_{-\infty}^{\infty} J(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} dv' - \frac{1}{\epsilon} \nabla \iiint_{-\infty}^{\infty} \rho(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} dv' \tag{1.31}$$

These two equations may be rewritten in terms of the scalar and vector potentials (mentioned at the outset of this chapter) as

$$H(r) = \nabla \times A(r) \tag{1.32}$$

$$E(r) = -jk\eta A(r) - \frac{1}{\epsilon} \Phi(r) \tag{1.33}$$

where

$$A(r) = \iiint_{v'} J(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} dv' \tag{1.34}$$

and

$$\Phi(r) = \iiint_{v'} \rho(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} dv' \tag{1.35}$$

By the earlier discussion on superposition, it's evident from the convolution form of A, Φ that they themselves satisfy wave equations of the form

$$\nabla^2 A + k^2 A = -J \tag{1.36}$$

$$\nabla^2 \Phi + k^2 \Phi = -\rho \tag{1.37}$$

The continuity equation (1.12) may be used to rewrite Eq. (1.33) entirely in terms of the vector potential, A as

$$E(r) = -jk\eta A(r) - j\frac{\eta}{k}\nabla[\nabla \cdot A(r)] \tag{1.38}$$

The two equations (1.32) and (1.38) for the electric and magnetic fields represent the solutions to Maxwell's equations. (We've extracted E, H from under the vector differential operators and expressed them directly in terms of the currents.) A considerable amount of time was spent developing these equations, even though the principles behind their derivations weren't all that complicated. This development time reflects directly on the importance of these equations. They form the foundation of physical optics diffraction theory, one of the three main techniques we'll study in this book for analyzing optical phenomena. (In the next chapter, we'll derive the equations for the plane wave spectrum theory, and in Chapter 3 we'll derive the equations for geometrical optics theory.) The time spent on these equations was worthwhile for forming a good understanding of the origin of these equations and their meaning.

As it turns out, Eqs. (1.32) and (1.38) aren't quite complete. They only allow us to calculate the fields due to a current distribution when the currents radiate in an unbounded homogeneous dielectric medium having infinite extent. Such an assumption won't be general enough for later needs in this book, where we'll have lots of boundaries—between lenses and air. Therefore, we must add extra terms to (1.32) and (1.38), which take material boundaries (such as those that exist between a glass lens and the air) into account.

To see how the field equations are modified for material boundaries, consider Fig. 1.1, which shows a two-region problem. The reader may be familiar with the fact that a sheet of current causes a discontinuity in magnetic field (known as a *jump discontinuity*). This is readily seen from Ampère's law, for if

$$\nabla \times H = j\omega\epsilon E + K \tag{1.39}$$

where K is a sheet current (measured in Amps/meter, and illustrated in Fig. 1.2), then integrating this equation over the surface shown (as ϵ tends to zero), using Stokes' theorem yields

$$H_{\tan}(z = 0^+) - H_{\tan}(z = 0^-) = K \times \hat{z} \tag{1.40}$$

Thus, the electric surface current, K, supports a jump discontinuity in the tangential magnetic field, H. By analogy, what would happen if we were to modify Faraday's law (1.8) to read

$$\nabla \times E = -j\omega\mu H - M \tag{1.41}$$

where M is a "sheet magnetic current," measured in volts/meter? Well, if M occupies the $z = 0$ plane, we may use the exact same logic as above to show that

$$E_{\tan}(z = 0^+) - E_{\tan}(z = 0^-) = -M \times \hat{z} \tag{1.42}$$

In other words, this sheet magnetic current causes a jump discontinuity in the tangential electric field. Of course, all we're doing here is playing games with Maxwell's equations, just to see how they might be modified in order to yield step discontinuities in the tangential electric and magnetic fields. These games have important practical uses, however. For example, say we were to take the situation in Fig. 1.1 and change it to the one

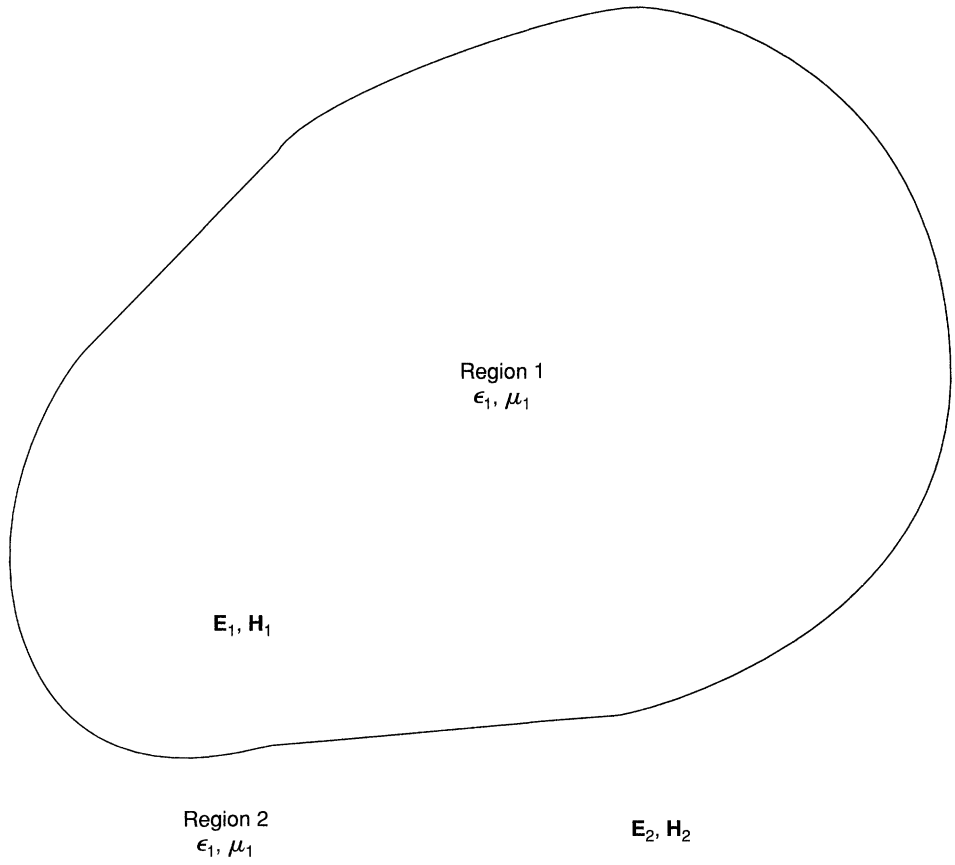


Figure 1.1 Field calculations in bounded dielectric media.

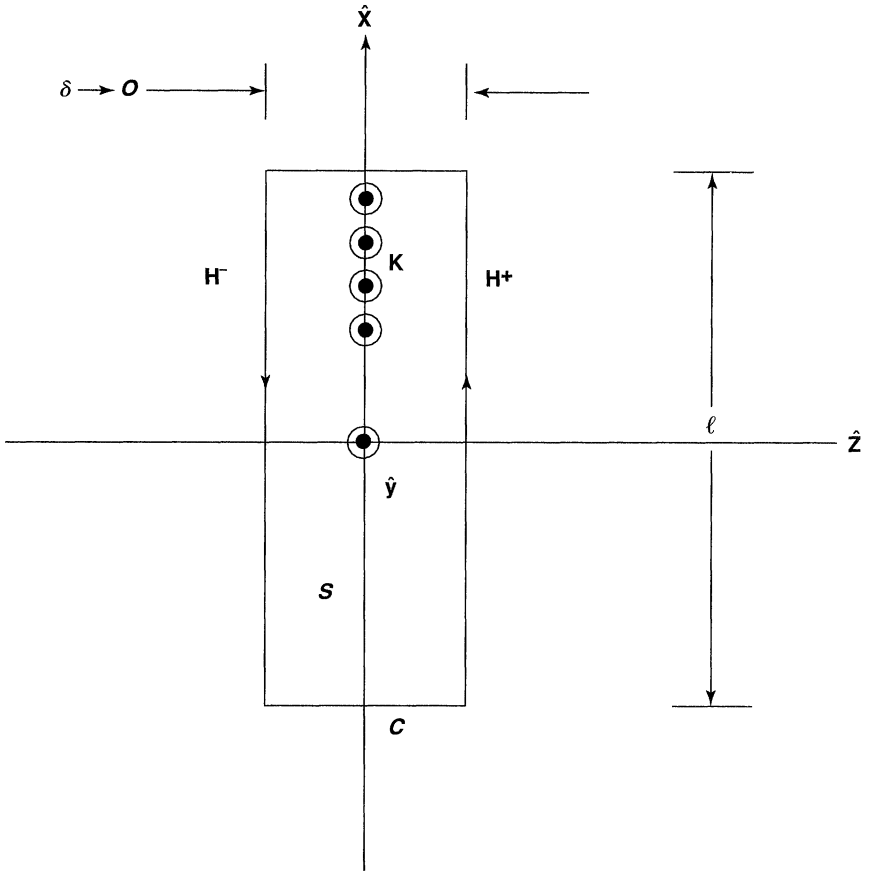
shown in Fig. 1.3. In Fig. 1.3, we have produced an infinite homogeneous medium [to which we can apply (1.30, 1.32)] from the mixed-medium situation shown in Fig. 1.1. The only difference now is that we've placed tangential electric and magnetic currents on the boundary surface between the two media, to support the discontinuities in the fields (from the original field values just inside the boundary to the null field values just outside the boundary). Both configurations will be equivalent (within medium 1, at least) if the following conditions hold at the boundary. These conditions are taken directly from the jump conditions above as

$$E_1 = \hat{n} \times M \tag{1.43a}$$

$$H_1 = K \times \hat{n} \tag{1.43b}$$

where the normal vector, $\hat{n}$ points into region 1.

All we have to do now is include the effects of K, M into (1.32) and (1.38). This isn't hard, because we've already done all the work. Noting that the magnetic current, M (in Faraday's law), has the same relationship to H as J has to E (in Ampère's law) and that M has the negative relationship to E that J has to H (with the roles of ϵ and μ



$$\nabla \times \mathbf{H} = j\omega\epsilon\,\mathbf{E} + \mathbf{K}$$

So,

$$\iint_s (\nabla \times \mathbf{H}) \cdot \hat{y}\, ds = \iint_s \mathbf{K} \cdot \hat{y}\, ds = K_y l$$

By Stokes' theorem however,

$$\iint_s (\nabla \times \mathbf{H}) \cdot \hat{y}\, ds = \oint_c \mathbf{H} \cdot d\mathbf{l} = (H_x^+ - H_x^-)l$$

So,

$$\mathbf{H}^+ - \mathbf{H}^- = \mathbf{K} \times \hat{z}$$

Figure 1.2 A jump discontinuity in the tangential magnetic field due to a sheet of current.

interchanged), we may immediately write the solutions to Maxwell's equations for bounded regions directly from (1.32) and (1.38). Thus,

$$E(r) = -j\omega\mu A(r) + \frac{1}{j\omega\epsilon}\nabla[\nabla \cdot A(r)] - \nabla \times F(r) \tag{1.44}$$

$$H(r) = -j\omega\epsilon F(r) + \frac{1}{j\omega\mu}\nabla[\nabla \cdot F(r)] + \nabla \times A(r) \tag{1.45}$$

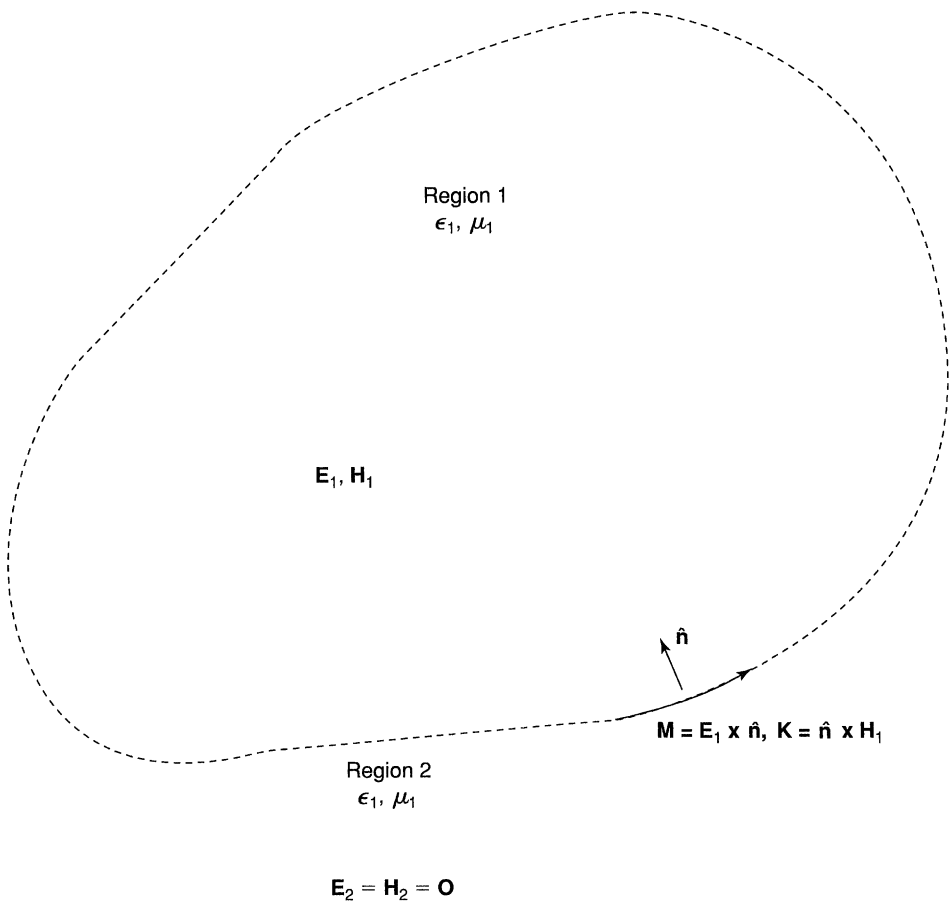


Figure 1.3 An electromagnetic problem that is equivalent (in region 1) to the two-region problem shown in Fig. 1.1.

where

$$F(r) = \iiint_{v'} M(r') \frac{e^{-jk|r-r'|}}{4\pi|r - r'|} dv' \tag{1.46}$$

These equations are written in terms of ω, ϵ, μ, rather than in terms of k and η, in order to show the dual nature of the equations. These are the equations we'll use in the diffraction integral analysis of optical systems, and they represent a mathematical statement of Huygens' principle. (Note: since M is a surface current, the integral in (1.46) is actually taken over a surface, not a volume.)

In the old days of optics, scalar equations similar to these were derived in various ad hoc ways. The names of numerous well-known researchers in classical optics have traditionally been associated with those earlier equations (usually hyphenations of names such as Huygens, Kirchhoff, Fresnel, Rayleigh, Sommerfeld, and Helmholtz). Coincidentally, the equations we've just derived also go by the hyphenation of two names, but they are the names of two twentieth-century researchers in electromagnetics, not optics. These are: J. A. Stratton and L. J. Chu, formerly of MIT, and their formulation of this solution to Maxwell's equations is known as the Stratton-Chu formulation [1].

1.4 ELECTROMAGNETIC POWER

The Poynting vector theorem relates stored and dissipated power within a reference volume to the integral of a "power flux" vector over the bounding surface of the volume. A very well-worn derivation of the Poynting vector theorem is generally used in electromagnetics [2], which, though rigorous from a mathematical standpoint, yields little physical understanding of the concept of electromagnetic power. Fortunately, it is a relatively simple matter to obtain a sound understanding of electromagnetic power from a simple intuitive approach. That is the approach we'll follow here.

Any concept of power must ultimately incorporate the idea of some type of mechanical motion. After all, power is a mechanical construct, being equal to work performed per unit time. When we think of ordinary electrical power, for example, what we're really thinking of is: how much mechanical power can this "electrical power" deliver? That is, can it drive a motor or cause a speaker to vibrate? The power available in an electromagnetic field can also be thought of in mechanical terms. For example, suppose we were to place a planar sheet of charged particles (charge/area) in the path of an electromagnetic wave. How much mechanical energy is available in the wave to expend on moving the charges in this sheet?

This question is readily answered. We know from elementary physics that the force exerted on a charge, q, by an electromagnetic field is given by

$$F = q(E + v \times B) \tag{1.47}$$

We also know from elementary physics that power is given as the dot product of force times velocity. (In the phasor domain, this is force times the *conjugate* of velocity; see Appendix B.) Thus, for an elemental square patch in the sheet,

$$P = F \cdot v^* = dq(E \cdot v^*) = E \cdot dK^* \tag{1.48}$$

where K is the surface current density, defined in the previous section. We know from the jump condition in the previous section that

$$H(z = 0^+) - H(z = 0^-) = K \times \hat{z} \tag{1.49}$$

So, say that all of the energy from the electromagnetic wave is absorbed by the sheet. That is, no energy gets past the sheet; hence,

$$H(z = 0^+) = 0$$

and

$$H(z = 0^-) = -K \times \hat{z}$$

or

$$K = H(z = 0^-) \times \hat{z}$$

Therefore,

$$P = E \cdot (H \times \hat{z})^* = \hat{z} \cdot (E \times H^*) \tag{1.50}$$

This leads naturally to the concept of the power vector (the so-called *Poynting vector*) as

$$P = E \times H^* \tag{1.51}$$

With this, the power dissipated (per unit area) in the sheet of charge is given by (1.50) as

$$P_z = \hat{z} \cdot P = \hat{z} \cdot (E \times H^*) \tag{1.52}$$

This equation directly relates the mechanical power dissipated on a sheet of charge to the component of the Poynting vector *normal* to the sheet. So, the Poynting vector is a "flux" type of vector; it points in the direction of power flow, and its integral over a surface determines the amount of power crossing that surface.

1.5 IMAGE THEORY

In this section, we'll show how to make the diffraction integral method practical. If this method is applied hastily to the analysis of optical systems, it can result in many unnecessary calculations. *Image theory* can dramatically reduce the number of calculations involved whenever planar (or nearly planar) boundaries are involved.

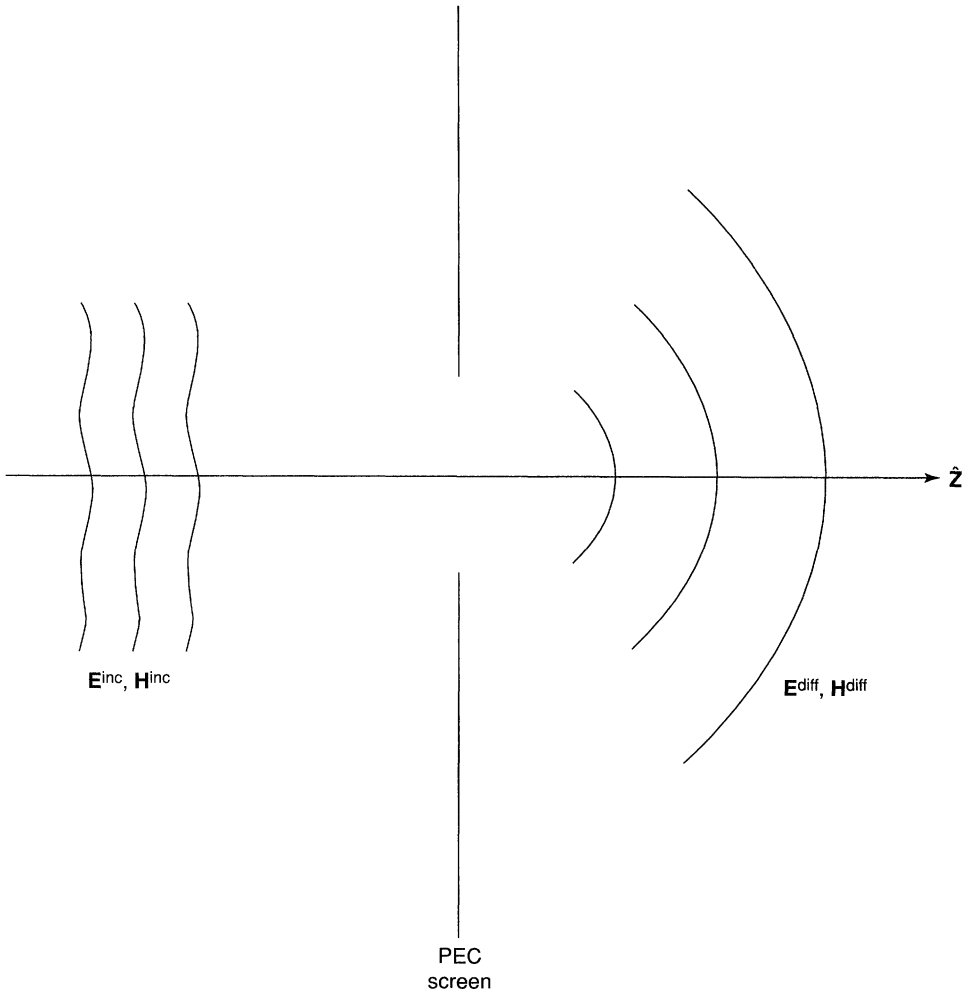


Figure 1.4 Diffraction by an aperture.

For example, consider the aperture in the planar wall shown in Fig. 1.4. We assume the wall to be perfectly electrically conducting (PEC); that is, the wall is so highly conductive that the tangential electric field can never take on a finite value. (We'll discuss perfect conductors in more detail below.) If we were to use the Stratton-Chu equations to calculate the electric field transmitted through the aperture to the $z > 0$ region, we'd have to integrate over the electric currents on the entire $z = 0$ plane, as well as integrate over the "equivalent electric and magnetic currents" on the aperture (which are expressed in terms of the tangential aperture fields by [1.43]). This is a lengthy calculation that is greatly simplified through the use of images.

We may verify from (1.44) and (1.45) that a planar sheet of "magnetic current" will radiate symmetric magnetic fields and antisymmetric electric fields. This is shown in Fig. 1.5. Since the tangential radiated electric field is antisymmetric, it must therefore be zero on those portions of the $z = 0$ plane where there are no magnetic currents. (Magnetic currents in the $z = 0$ plane will produce a step discontinuity, from a nonzero positive value on one side to the opposing negative value on the other.) Thus, magnetic currents in the $z = 0$ plane radiate fields that exactly satisfy the aperture and wall boundary conditions of the original aperture diffraction problem. That is, the tangential electric field is zero at $z = 0$, outside the aperture portions of the plane, and equal to the original aperture electric field in the aperture.

We may use these unique radiative properties of planar magnetic current sheets to advantage in analyzing the aperture diffraction problem. As shown in Fig. 1.6, the original aperture diffraction problem may be modified by setting the electric field on the left side

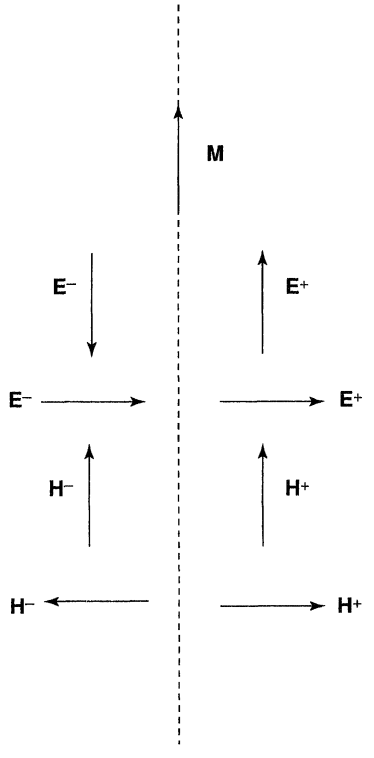


Figure 1.5 Fields produced by a planar magnetic current distribution.

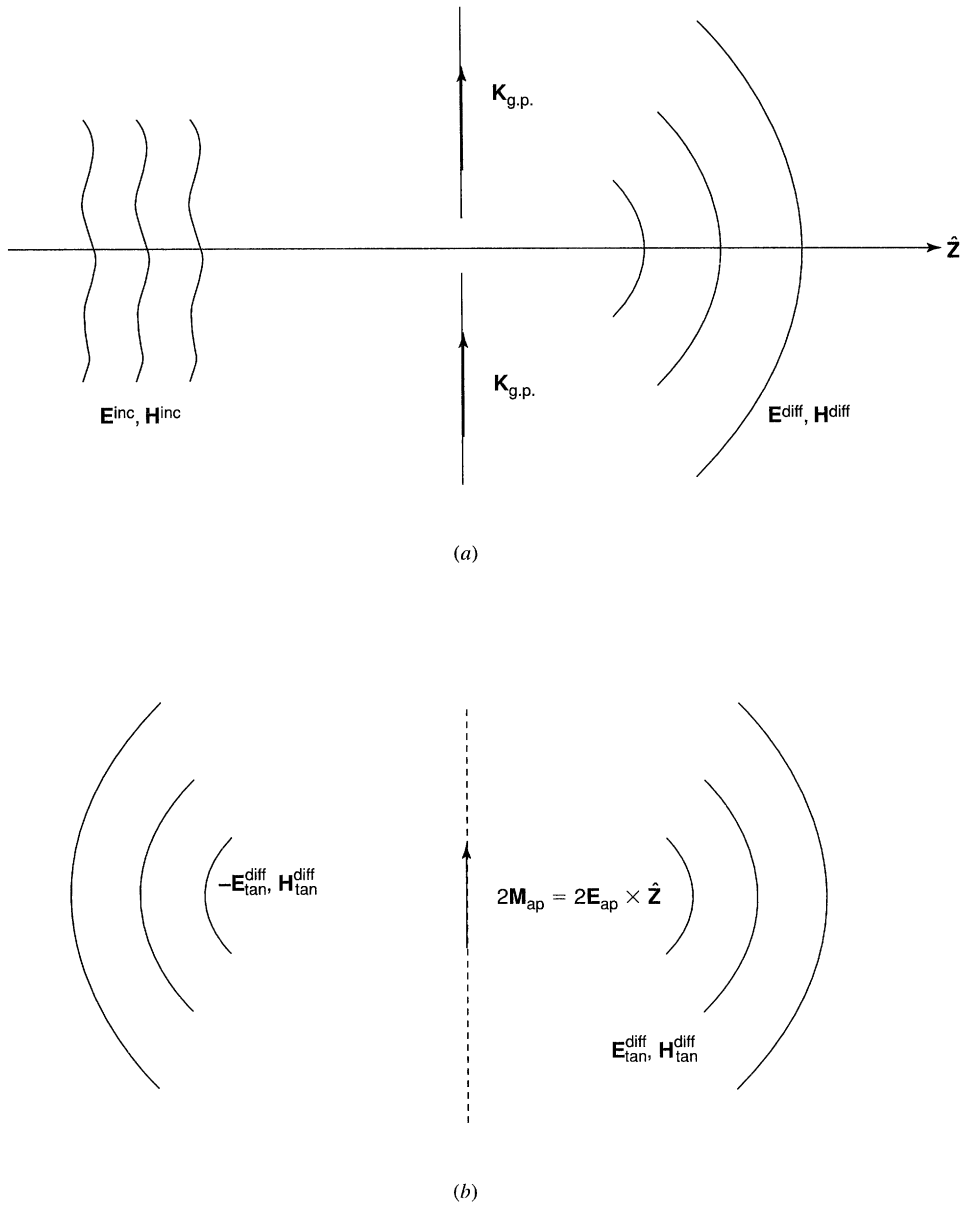


Figure 1.6 Simplification of the aperture diffraction problem: (a) Original aperture diffraction problem and (b) Transformed aperture diffraction problem equivalent to the original in the right-hand half space.

exactly equal to the negative of the diffracted electric field on the right side, and the magnetic field on the left side equal to the diffracted magnetic field on the right side. Thus, a step discontinuity has been created in the tangential electric field at $z = 0$ (equal to twice the aperture field), and the tangential magnetic field has been made continuous at $z = 0$. So, since the tangential magnetic field is now continuous everywhere at $z =$

0, we may remove all the electric currents from the $z = 0$ plane of our transformed problem. And since we've created a step discontinuity in the electric field (in the aperture), we must now add aperture magnetic currents to support this newly created discontinuity.

The original aperture diffraction problem shown in Fig. 1.6*a* now becomes the transformed aperture diffraction problem shown in Fig. 1.6*b*. The fields on the right-hand side of the aperture are exactly the same as those in the case of the original aperture diffraction problem—only the fields in the left half space have been altered. So, the price we've paid for obtaining a simplified mathematical problem is that we now have a problem that is equivalent to the original problem in the $z > 0$ region only.

Calculating the diffracted fields for the problem shown in Fig. 1.6*c* is relatively easy. We know the aperture electric field distribution. (In the optical limit, it's just equal to the incident field in the aperture.) And we can easily calculate the electric field radiated by the magnetic currents using (1.44). So, we've succeeded in making the initially intractable aperture diffraction problem tractable.

In electromagnetics, the construct of a PEC material is often used. This material is so highly conductive that an infinitesimal tangential electric field on its surface will result in a finite electrical current on the surface of the material. Materials that simulate PEC surfaces include metals such as silver, copper, and gold. Today's high-temperature superconducting materials are (for all practical purposes) PEC materials, at least up to moderate microwave frequencies. The optical analogue of a PEC material is a reflective mirror. In many types of optical systems (cameras, telescopes, etc.), however, nonreflective (absorptive) black-colored materials are preferable to highly reflective PEC materials. These opaque materials absorb stray light and prevent it from propagating further through the system. However, the PEC construct is still used for aperture penetration problems. (It is the only possible way to "image away" the electric currents in the aperture and on the groundplane, and confine the integration to the magnetic currents in the aperture.) It is remarkable that this assumption produces accurate results for aperture-diffracted fields.

1.6 APPLICATIONS: FRAUNHOFER REGION FIELDS OF AN ILLUMINATED APERTURE

In the majority of calculations in optics, the diffraction integral formulation is applied in only two main regimes. One of these is the *Fraunhofer,* or far-field region, which is used primarily for analyzing the interaction between an optical instrument and its outside environment. For example, this situation would describe the field incident on a camera lens from some sort of environmental scene, or the field radiated by a laser beam. The second regime is the *Fresnel* regime, which is used primarily to describe the movement, or propagation, of optical energy from one transverse plane to another *within* an optical system. The diffraction integral formulation can be used to describe either regimes, and in this section we'll consider the Fraunhofer regime.

With reference to Fig. 1.7, we can obtain the far-zone expressions for the electric and magnetic fields by invoking the *parallel rays* approximation. In this approximation, the free-space scalar Green's function is approximated as

$$\frac{e^{-jk|r-r'|}}{4\pi|r - r'|} \cong \frac{e^{-jkr}}{4\pi r} e^{jk(\hat{r}\cdot r')} \tag{1.53}$$

Using this approximation, it is shown [3] that

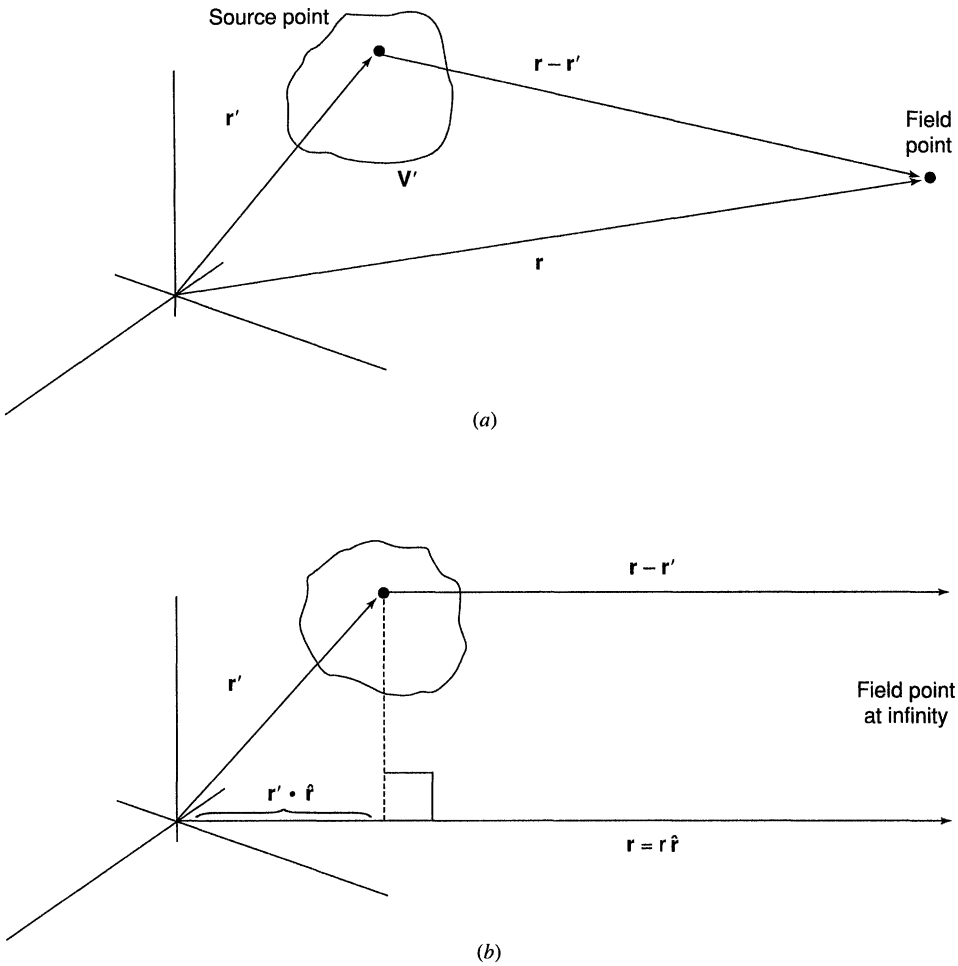


Figure 1.7 The parallel rays approximation for Fraunhofer region fields: (a) field point at finite distance from source region and (b) field point infinitely far away from sources.

$$E(\hat{r}) = -j\omega\mu[A - \hat{r}(\hat{r} \cdot A)] - \hat{r} \times F \tag{1.54}$$

$$H(\hat{r}) = -j\omega\epsilon[F - \hat{r}(\hat{r} \cdot F)] + \hat{r} \times A \tag{1.55}$$

These expressions show that in the far-zone region, both electric and magnetic fields are transverse to each other and to the radial vector, r. They also show that in the far-zone region, the electric and magnetic fields of a planar distribution are related to the Fourier transform of the source fields and/or currents since

$$J(r') = J(x', y')$$

and

$$\hat{r} \cdot r' = (\alpha x + \beta y)$$

This is a very important result in the study of optical systems.

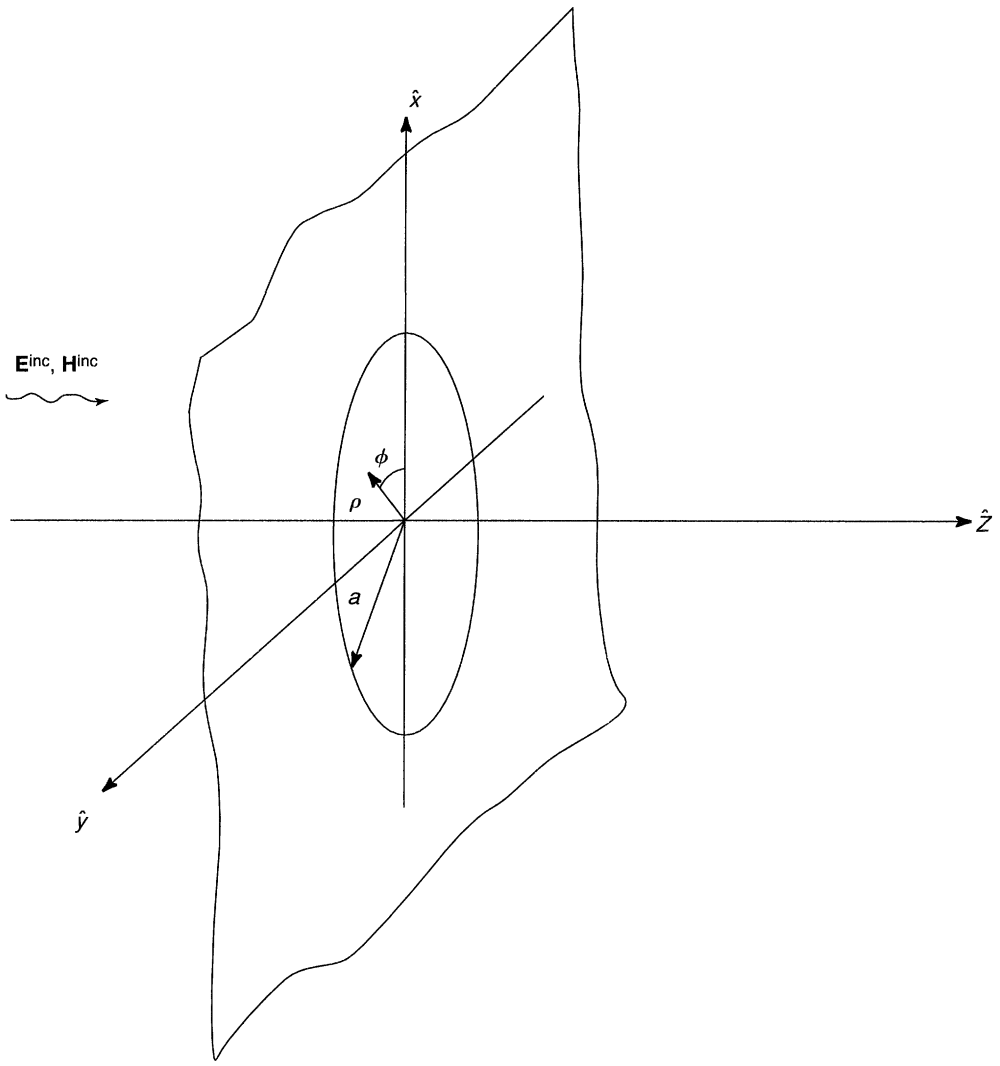


Figure 1.8 Plane wave diffraction by a circular aperture in a PEC screen.

EXAMPLE 1.1 THE AIRY PATTERN FOR CIRCULAR APERTURES

A uniformly illuminated circular aperture radiates a pattern that is often referred to as the *Airy pattern.* We may easily derive this Airy function pattern using (1.54) (with $A = 0$), in conjunction with an image transformation of the type described in Section 1.6. If the electric field incident on the aperture is a plane wave field, as shown in Fig. 1.8, then (from the results of Section 1.6),

$$M(x, y) = 2E^{\text{inc}}(x, y, z = 0) \times \hat{z}$$

$$\text{so, if} \quad E^{\text{inc}}(x, y, z = 0) = \frac{1}{2}\hat{y}$$

$$\text{then,} \quad M(x, y) = \hat{x}$$

(where we have approximated the true field in the aperture with the incident field in the aperture via the Physical Optics approximation).

The integral for the electric vector potential then becomes

$$F(r,\ \theta,\ \phi) = \frac{e^{-jkr}}{4\pi r}\,\hat{x}\int_0^{2\pi}\int_0^{a} e^{jk_\rho \rho \sin\theta \cos(\phi-\phi_k)}\,\rho d\rho d\phi$$

where

$$k_\rho \rho \sin\theta \cos(\phi - \phi_k) = \alpha x + \beta y$$

with,

$$x = \rho\cos\phi$$
$$y = \rho\sin\phi$$
$$\alpha = k_\rho \cos\phi_k$$
$$\beta = k_\rho \sin\phi_k$$

This integral expression is readily evaluated in closed form using the following two integral identities for Bessel functions (see ref. 16, Chapter 2):

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{\pm jx\cos(\phi-\phi_0)}\,d\phi$$

and

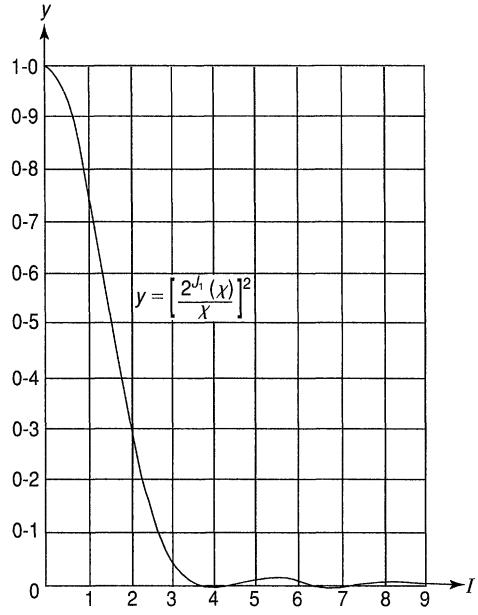


Figure 1.9 The Airy function $J_1(x)/x$ (after, M. Born and E. Wolf, *Principles of Optics*, 6th ed, Pergamon Press).

$$x J_1(x) = \int_0^x J_0(u)\, u\, du$$

to yield the electric vector potential as

$$F(r, \theta, \phi) = \frac{a^2}{2} \frac{e^{-jkr}}{4\pi r} \frac{J_1(ka \sin\theta)}{ka \sin\theta}$$

The function $J_1(x)/x$ is the classic Airy pattern, shown in Fig. 1.9.

EXAMPLE 1.2 THE FAR-FIELD CRITERION

In the text, it is shown how the parallel rays approximation may be used to simplify field calculations when the field point is infinitely distant from the source distribution. In practical calculations, the parallel ray approximation is valid when the *far-field criterion* is met. This criterion determines how far a field point must be away from a radiating aperture (for a given frequency and aperture size) in order for the parallel rays approximation to be valid. We may readily find the far-field criterion with the aid of Fig. 1.10.

Say a field point is located on-axis, a distance R away from the aperture. The aperture has a lateral extent, D, in the $x - z$ plane. Rays from the two extreme edges of the aperture are clearly not parallel, so the parallel rays approximation is not strictly satisfied. We can use the approximation, however, when the optical phase, kR_1 is not too different from the phase kR. So, say we require

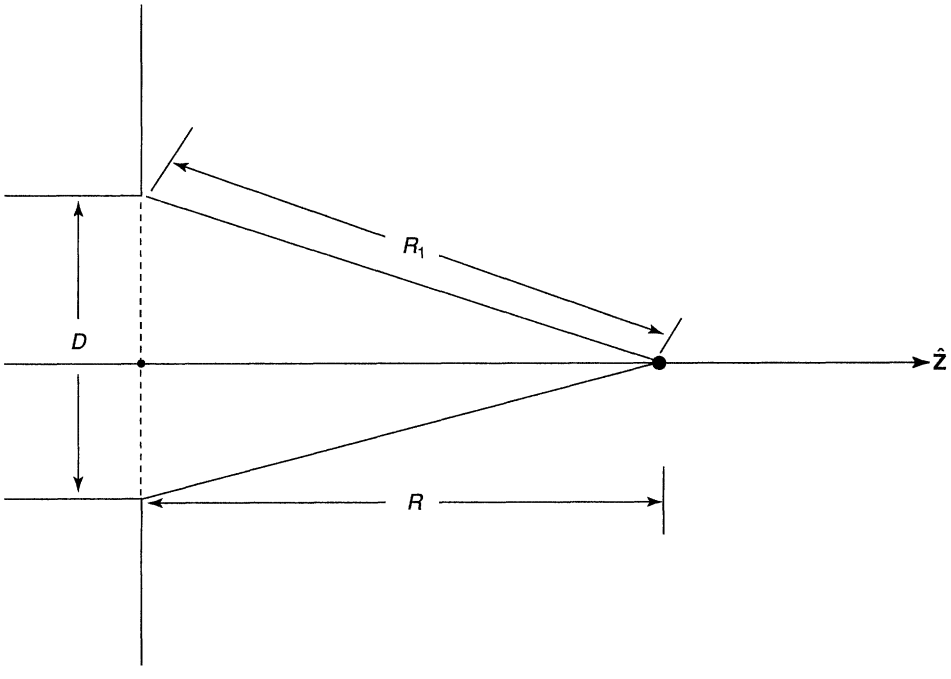


Figure 1.10 On the far-field criterion.

$$k(R_1 - R) < \frac{\pi}{8}$$

Then, since,

$$R_1 = \sqrt{(D/2)^2 + R^2} \cong R\left[1 - \frac{1}{8}(D/R)^2\right]$$

the phase criterion above becomes

$$R > \frac{2D^2}{\lambda}$$

According to this criterion, the far-field range increases as the aperture size increases and as the frequency increases.

EXAMPLE 1.3 FRESNEL ZONES

As we saw in the previous example, for field points located a finite distance from the aperture, the distance function does not necessarily satisfy the far-field criterion. This criterion is satisfied only when the wavefront curvature remains below an acceptable level. (We arbitrarily set that level equal to $\pi/8$ in the example above.) What happens when the wavefront curvature gets larger and perhaps even exceeds 2π? Well, with reference to Fig. 1.11*a*, we may approximately say that when kR_1 is between $-90°$ and $90°$ of kR, the field is basically in phase with the center of the aperture, and when the phase is outside this range, the field is out of phase. In-phase fields tend to add and out-of-phase fields tend to subtract, as shown in Fig. 1.11*c*. The alternate in-phase and out-of-phase zones on an aperture are known as *Fresnel zones*. We may readily find expressions for the Fresnel zones of a circular aperture. (*Note:* Zones similar to Fresnel zones are also created when (1.54) and (1.55) are evaluated for field points that lie off the main axis of the aperture; the computational difficulty involved in evaluating (1.54) and (1.55) in this case is directly proportional to the number of Fresnel zones in the aperture.)

So, since

$$k(R_1 - R) \cong kR\frac{1}{2}(x/R)^2$$

the in-phase Fresnel zones lie between

$$x = \sqrt{nR\lambda}$$

and

$$x = \sqrt{\left(n + \frac{1}{2}\right)R\lambda}$$

A Fresnel zone lens consists of a mask placed over an unfocused circular aperture, with the out-of-phase Fresnel zones covered. An example of such a masked aperture is shown in Fig. 1.12; the mask produces nearly the same focusing properties as a dielectric lens. It should be noted, however, that since the Fresnel zone mask is a *binary* mask (i.e., the zones are either 0% or 100% transmissive, and these zones coincide with the "in-phase" and "out-of-phase" regions), it may launch both converging and diverging spherical wave fields (the first constituting a real image and the second a virtual image, as we'll read about later). This property of a planar structure to produce multiple transmitted fields is characteristic of many types of devices we'll study in this book,

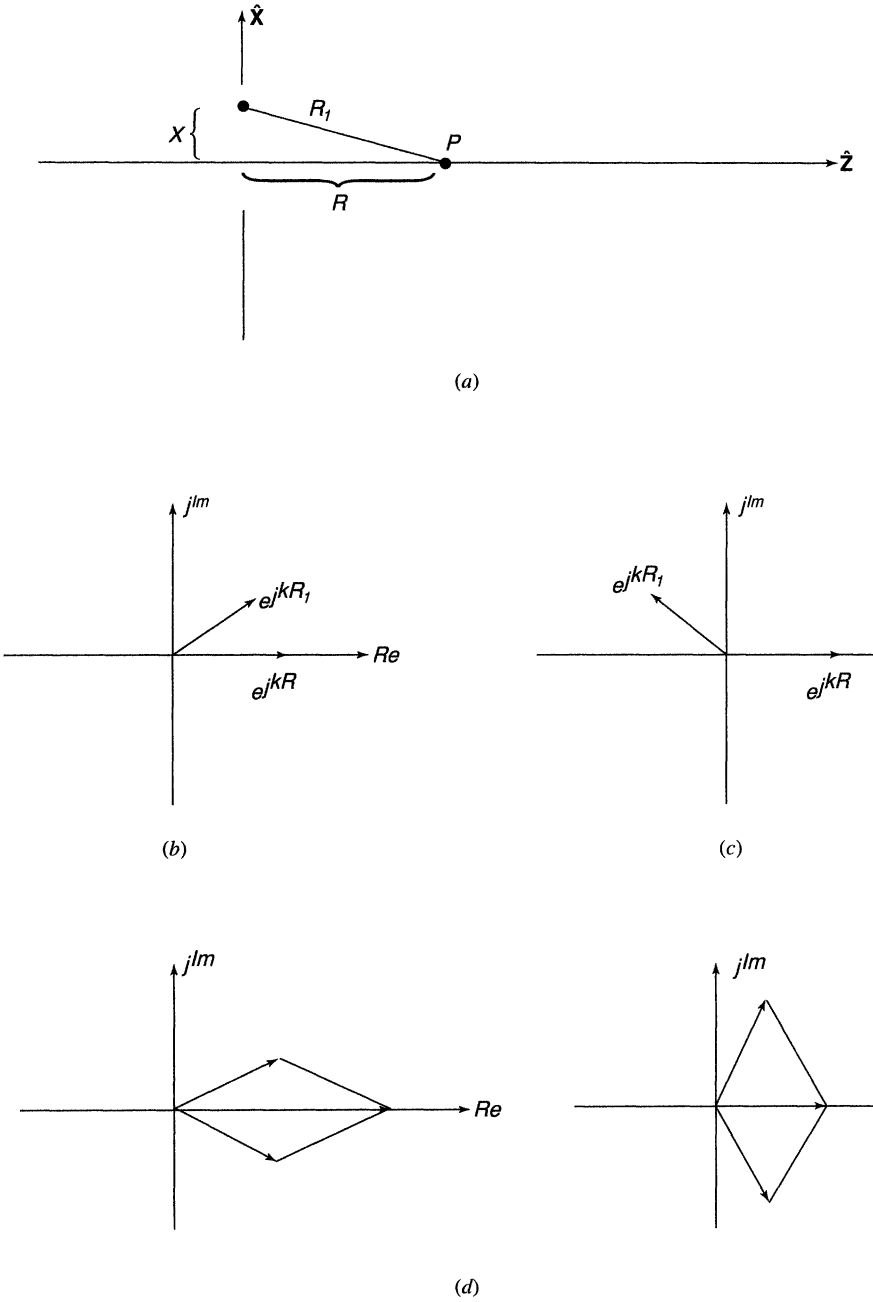


Figure 1.11 On the nature of Fresnel zones: (a) Phase from an aperture point to the axial field point, P; (b) Approximate in-phase condition: $|k\,(R_1 - R)| \leq \pi/2 \pmod{2\pi}$; (c) Approximate out-of-phase condition: $|k\,(R_1 - R)| \geq \pi/2 \pmod{2\pi}$; and (d) Addition of "in-phase" and "out-of-phase" complex numbers.

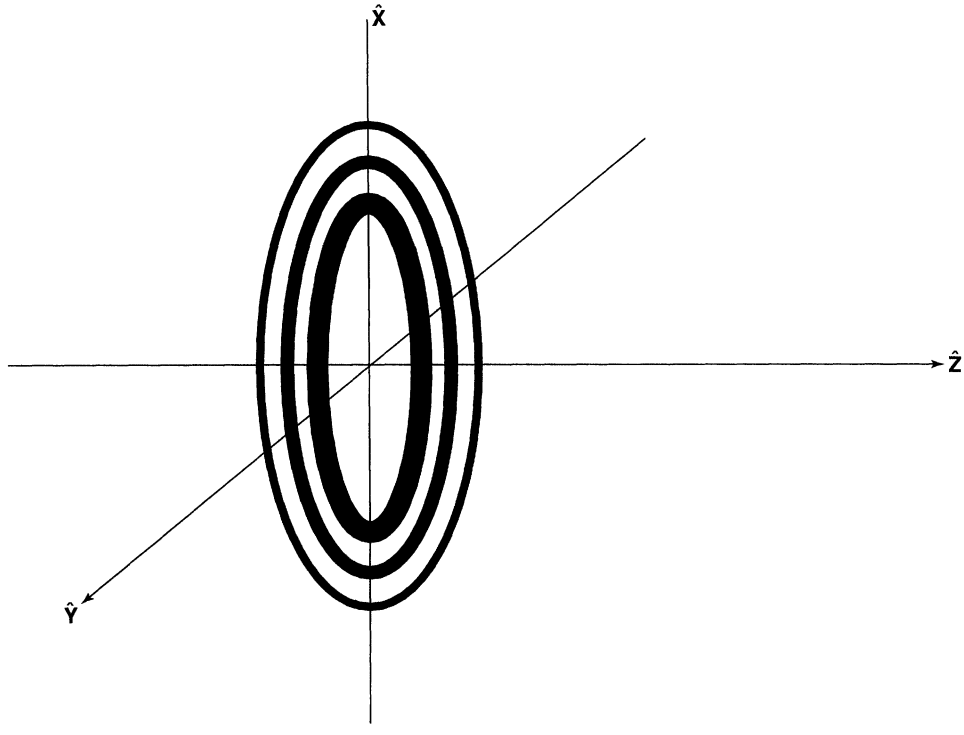


Figure 1.12 Fresnel zone mask.

including diffraction gratings, holograms, holographic lenses, and the like. In fact, the Fresnel zone concept has been used in the past both to understand the operation of holograms on optical fields [4] and to provide insight into the design of holographic devices such as laser beam scanners [5].

REFERENCES

[1] Stratton, J. A., *Electromagnetic Theory,* New York: McGraw-Hill, 1941.

[2] Scott, C. R., *Field Theory of Acousto-Optic Signal Processing Devices,* Norwood, MA: Artech House, 1992.

[3] Scott, C. R., *Modern Methods of Reflector Antenna Analysis and Design,* Norwood, MA: Artech House, 1992.

[4] Siemens-Wapniarski, W. J., and Givens, M. Parker, "The Experimental Production of Synthetic Holograms," *Applied Optics,* vol. 7, no. 3, March 1968.

[5] Lee, Wai-Hon, "Holographic Grating Scaners with Aberration Corrections," *Applied Optics,* vol. 16, no. 5, May 1977.