CHAPTER

# The Optics of Microscope Image Formation

2

**David E. Wolf**

*Sensor Technologies, LLC, Shrewsbury, Massachusetts, USA*

## CHAPTER OUTLINE

**Methods in Cell Biology, Volume 114**

**ISSN 0091-679X**

http://dx.doi.org/10.1016/B978-0-12-407761-4.00002-6

## Abstract

Although geometric optics gives a good understanding of how the microscope works, it fails in one critical area, which is explaining the origin of microscope resolution. To accomplish this, one must consider the microscope from the viewpoint of physical optics. This chapter describes the theory of the microscope-relating resolution to the highest spatial frequency that a microscope can collect. The chapter illustrates how Huygens' principle or construction can be used to explain the propagation of a plane wave. It is shown that this limit increases with increasing numerical aperture (NA). As a corollary to this, resolution increases with decreasing wavelength because of how NA depends on wavelength. The resolution is higher for blue light than red light. Resolution is dependent on contrast, and the higher the contrast, the higher the resolution. This last point relates to issues of signal-to-noise and dynamic range. The use of video and new digital cameras has necessitated redefining classical limits such as those of Rayleigh's criterion.

*The nature of light is a subject of no material importance to the concerns of life or to the practice of the arts, but it is in many other respects extremely interesting.*

**(Thomas Young[1], 1773–1829)**
**(M. Shamos, 1987)**

## INTRODUCTION

Although geometric optics gives us a good understanding of how the microscope works, it fails in one critical area: explaining the origin of microscope resolution. Why is it that one objective will resolve a structure whereas another will not? This is the question we will examine in this chapter. To accomplish this, we must consider the microscope from the viewpoint of physical optics. Useful further references are

[1] There is a coincidental connection between the lives of Jean Baptiste Fourier and Thomas Young that transcends optics. In 1798, Fourier joined Napoleon's army during the invasion of Egypt. Word of the French Expeditionary Force had reached Admiral Nelson, who unsuccessfully chased the French fleet across the Mediterranean. On August 1, his luck changed. The French fleet was anchored in a wide crescent at the bay at Aboukir, 37-km east of Alexandria. Secure on their shoreward side, the French had moved all of their cannons to the seaward side. At 5:40 PM, Nelson began the assault. British men-of-war Goliath and Zealous managed to wedge their way between the French and the shore. It was a bold move that led to the destruction of the French fleet and stranded the French Expeditionary Force in Egypt, among them Fourier (www.discovery.ca/napoleon/battle.cfm). The French Expeditionary Force is perhaps most famous for its discovery in 1799 by Dhautpol and Pierre-Francois Bouchard of the Rosetta stone near the town of Rosetta in the Nile delta. Translation of the Rosetta stone represented one of the great intellectual challenges of time and, similar to the work of Young and Fourier in physics and mathematics, respectively, it expanded humanity's worldview. Thomas Young was a brilliant linguist and was the first to recognize that some of the hieroglyphics on the Rosetta stone were alphabetic characters. Young's work was the critical starting point that ultimately enabled Jean Francois Champollion to complete the translation of the Rosetta stone (www.rosetta.com/RosettaStone.html).

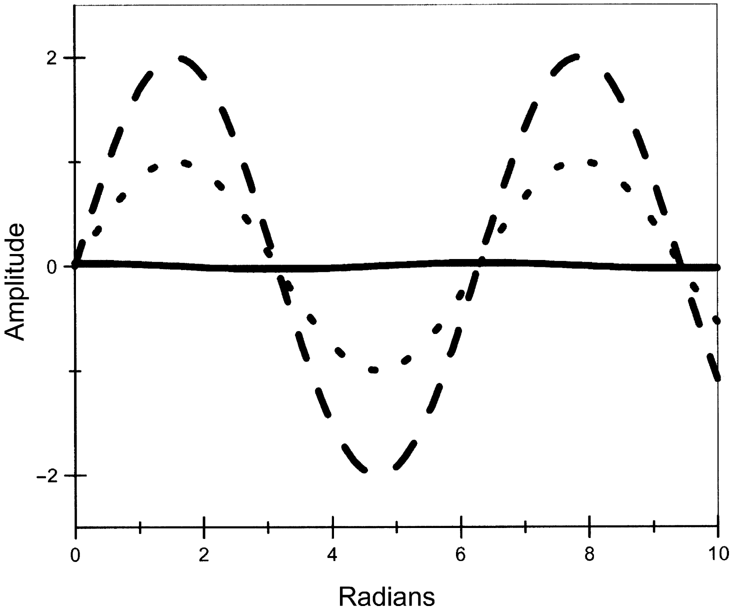

FIGURE 2.1

Superposition of sine waves: *dotted line*, sin($x$); *dashed line*, the condition of constructive interference sin($x$) + sin($2\pi + \phi$); *solid line*, the condition of destructive interference sin($x$) + sin($x + \pi - \phi$) where $\phi \ll 1$.

Inoué & Spring, 1997; Jenkins & White, 1957; Sommerfeld, 1949a; Born & Wolf, 1980 for the optics of microscope image formation.

## 2.1 PHYSICAL OPTICS: THE SUPERPOSITION OF WAVES

Let us consider a simple type of wave, namely, a sine or cosine wave, such as that illustrated in Fig. 2.1. Mathematically, the equation of the wave shown in Fig. 2.1 (dotted line) is

$$y(t) = y_0 \sin(\omega t). \tag{2.1}$$

Here, we are considering the abstract concept of a wave in general. However, a clear example of a wave is the light wave. Light is a wave in both space and time. The equation for its electric field vector, $E$, takes the form

$$E(x,t) = E_0 \sin(kx - \omega t), \tag{2.2}$$

where $\omega$ (the frequency in radians per second) is related to the frequency of the light, $\nu$, by the relation $\omega = 2\pi\nu$; $k$ is the wave number, which is related to the wavelength, $\lambda$, by the relation $k = 2\pi/\lambda = 2\pi\nu/c$, where $c$ is the speed of light; and $E_0$ is the amplitude. In

Appendix A, we show that Eq. (2.2) represents the unique solution of the wave equation, which governs optics. Two additional points are worth mentioning here: First, that the intensity of light is the square of the electric field vector; second, that the spatial and temporal components of the light wave can be separated. As a result, you can view light as being a spatial sine or cosine wave moving through space with time.

Now let us suppose that we have two simultaneous waves with the same frequency,

$$y(t) = y_0 \sin(\omega t) + y_1 \sin(\omega t + \phi), \tag{2.3}$$

but that are phase-shifted with respect to one another with a phase shift $\phi$. The composite wave is determined by pointwise addition of the individual waves, a principle known as the superposition theorem. When the two waves are completely in phase (i.e., $\phi = 0$), we have the condition of constructive interference shown in Fig. 2.1 (dashed line). When the two waves are 180° out of phase, we have the condition of destructive interference shown in Fig. 2.1 (solid line).

This can be most readily shown by adopting the exponential form of the wave. In general, a complex number can be expressed in one of two ways, the right- and left-hand sides of Eq. (2.4):

$$\mathrm{e}^{i\theta} = \cos\theta + i\sin\theta. \tag{2.4}$$

Thus, our sine wave may be expressed as

$$\begin{aligned} E(x,t) &= E_0\,\mathbf{Re}\{\mathrm{e}^{-i(kx-\omega t)}\}, \\ E(x,t) &= E_0\,\mathbf{Re}\{\mathrm{e}^{-ikx}\mathrm{e}^{+i\omega t}\}, \end{aligned} \tag{2.5}$$

where **Re** stands for "the real part of."

## 2.2 HUYGENS' PRINCIPLE

In 1678, the Dutch physicist Christiaan Huygens (1629–1695) evolved a theory of wave propagation that remains useful in understanding such phenomena as reflection, refraction, interference, and diffraction. Huygens' principle is that:

> *All points on a wave front act as point sources for the production of spherical wavelets. At a later time the composite wave front is the surface which is tangential to all of these wavelets.*
>
> **(Halliday & Resnick, 1970)**

In Fig. 2.2, we illustrate how Huygens' principle or construction can be used to explain the propagation of a plane wave. We have chosen this example because it seems otherwise counterintuitive that one can construct a plane out of a set of finite radius spheres.

Expressed in this way, Huygens' principle is an empirical construct that will not explain all aspects and phenomena of light. Clearly, a complete description requires the application of James Clerk Maxwell's (1831–1879) wave equation and the

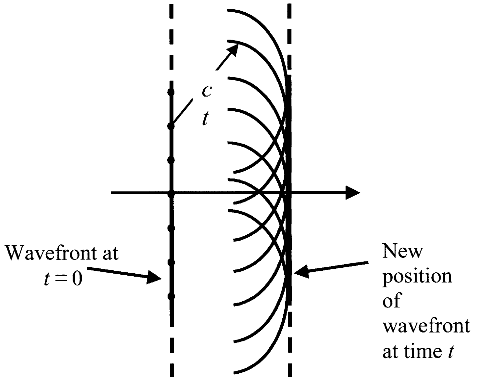

**FIGURE 2.2**

Huygens' construct illustrating how a plane wave front at time $t = 0$ propagates as a plane wave front at time $t$.

boundary conditions of his electromagnetic theory. In Appendix A, we demonstrate that Eq. (2.2) is, indeed, a solution to Maxwell's wave equation. Gustav Kirchoff (1824–1887) has developed a more robust form of Huygens' principle that incorporates electromagnetic theory. In Appendix B, we develop the rudiments of Kirchoff's approach. Subsequently, in Appendix C, we use Kirchoff's solution to develop a mathematical treatment of the Airy disk. The reader is referred to Sommerfeld (1868–1951) (Sommerfeld et al., 1949) for an excellent description of diffraction theory.

## 2.3 YOUNG'S EXPERIMENT: TWO-SLIT INTERFERENCE

One usually sees Huygens' principle described using the quotation above. In practice, however, it is more often applied by considering a wave surface and then asking what the field will be on at some point, $P$, away from the surface. The composite field will be given by the sum of spherical wavelets reaching this point at a given time. Because the distances between points on the surface and point $P$ vary, it must be the case that for them to arrive simultaneously, they must have been generated at different points in the past. In this context, Huygens' principle is really an expression of the superposition theorem.

This was the approach taken by Thomas Young (1773–1829) in 1801 in explaining interference phenomena (Young, 1801). Young's now classic experiments demonstrated the fundamental wave nature of light and brought Young into conflict with Sir Isaac Newton (1643–1727) and his Corpuscular Theory of Light. Young's experiment is illustrated in Fig. 2.3. Young used a slit at A to convert sunlight to a coherent spherical wave (Huygens' wavelet). Two slits are symmetrically positioned on B relative to the slit at A. Huygens' wavelets propagate from the two slits and will,

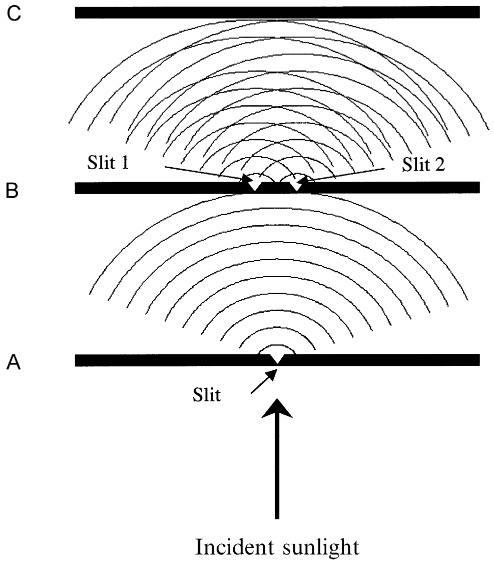

**FIGURE 2.3**

Young's double-slit experiment in terms of Huygens' construct.

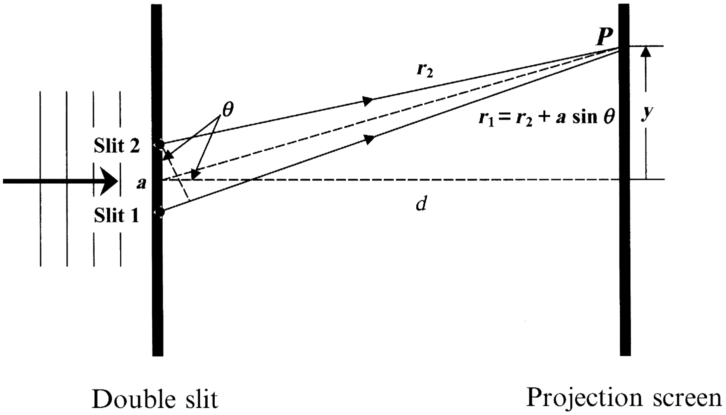

**FIGURE 2.4**

Young's double-slit experiment showing the source of the phase shift geometrically.

at various points, constructively and destructively interfere with one another. Figure 2.4 considers some arbitrary point $P$, a distance $y$ from the center of the surface $C$, which is a distance $D$ from $B$. If the distance between the slits is $d$, the path length difference between the two wavelets at $P$ will be $d \sin(\theta)$ and intensity maxima caused by constructive interference will occur when

$$d \sin(\theta) = m\lambda, \tag{2.6}$$

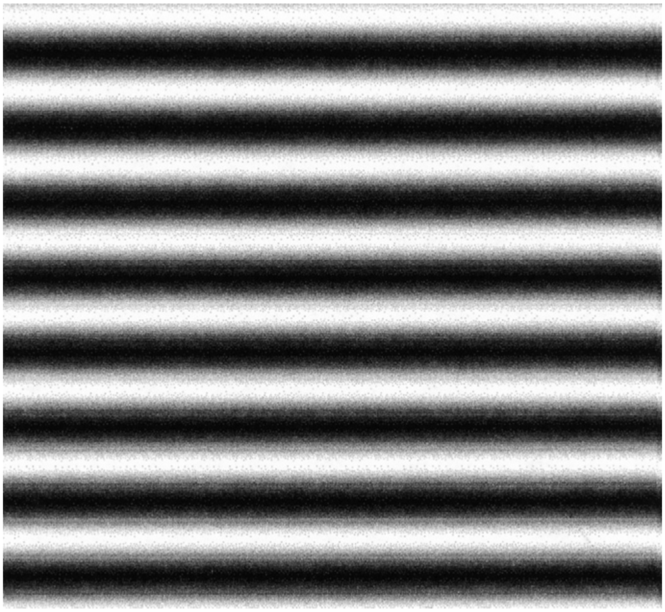

FIGURE 2.5

Young's double-slit interference pattern.

where $m = 0, 1, 2, 3$, and so forth. The pattern that Young saw is shown in Fig. 2.5 and is referred to as the double-slit interference pattern. One clearly observes the alternating intensity maxima and minima. The maxima correspond to the angles $\theta$ given by Eq. (2.4).

It is worthwhile to examine this problem more closely and to determine the actual intensity profile of the pattern in Fig. 2.5. Huygens' principle promises us that it derives from the superposition theorem. If the time-dependent wave from slit 2 is $E_2 = E_0 \sin \omega t$ and the wave from the slit 1 is $E_1 = E_0 \sin(\omega t + \phi)$, then the time-dependent wave at point $P$ is:

$$E = E_1 + E_2 = E_0(\sin \omega t + \sin[\omega t + \phi]), \tag{2.7}$$

where again $\phi = a \sin \theta$. Eq. (2.5) may be algebraically manipulated as follows:

$$\begin{aligned} E &= E_0(\sin \omega t + \sin \omega t \cos \phi + \cos \omega t \sin \phi) \\ &= E_0(\sin \omega t\,(1 + \cos \phi) + \cos \omega t \sin \phi). \end{aligned} \tag{2.8}$$

The intensity is given by $E^2$, therefore,

$$I = E_0^2(\sin^2 \omega t(1 + \cos \phi)^2 + \cos^2 \omega t \sin^2 \phi + 2 \sin \omega t \cos \omega t \sin \phi[1 + \cos \phi]). \tag{2.9}$$

What we observe physically is the time-averaged intensity, $I_{\mathrm{AV}}$. Recalling that the time average of $\sin^2 \omega t$ and $\cos^2 \omega t$ is 1/2 whereas that of $2\sin \omega t \cos \omega t = \sin 2\omega t = 0$, we obtain

$$I_{\mathrm{AV}} = \frac{E_0^2}{2}([1 + \cos\phi]^2 + \sin^2\phi) \tag{2.10}$$

$$\begin{aligned} I_{\mathrm{AV}} &= \frac{E_0^2}{2}(1 + 2\cos\phi + \cos^2\phi + \sin^2\phi) \\ &= E_0^2(1 + \cos\phi) \\ &= 2E_0^2 \cos^2\frac{\phi}{2}. \end{aligned} \tag{2.11}$$

Considering Fig. 2.4, because the angle $\theta$ is small,

$$\phi = a\sin\theta \simeq a\tan\theta = \frac{ay}{d}, \tag{2.12}$$

and therefore,

$$I_{\mathrm{AV}} = 2E_0^2 \cos^2\left(\frac{ay}{2d}\right). \tag{2.13}$$

In deriving Eq. (2.11), we have ignored the $1/r^2$ falloff of intensity. To allow for this, one would have to replace $E$ in Eq. (2.5) by $E/r$. That is, we assume a spherical rather than a plane wave. This, of course, causes the intensity of the bands to fall off with increasing $y$.

In this derivation, one sees the fundamental application of the superposition theorem to derive the composite wave. In this case, the application is relatively straightforward as the wave at $P$ is the superposition of only two waves. In other cases, such as the single-slit diffraction example that follows, the composite is a superposition of an infinite number of waves, and the summation becomes an integration. In Appendix B, we consider Kirchoff's generalization of this problem to a scalar theory of diffraction.

## 2.4 DIFFRACTION FROM A SINGLE SLIT

The double-slit interference experiment is an example of Huygens' principle of superposition where we have only two generating sites. A related interference phenomenon is that of single-slit diffraction, which is illustrated in Fig. 2.6. Here, we envision a plane wave impinging on a narrow slit of width $a$. We imagine the slit divided into infinitesimally narrow slits separated by distance d$x$, each of which acts as a site that generates a Huygens' wavelet. Two neighboring wavelets generate parallel wavelets. A lens collects these wavelets and brings them to a focus at the focal plane. We consider two wavelets from neighboring regions of the slit, which ultimately converge on point $P$ of the focal plane. The path difference will be d$x$ sin ($\theta$). Calculation of the resulting interference pattern referred to as single-slit

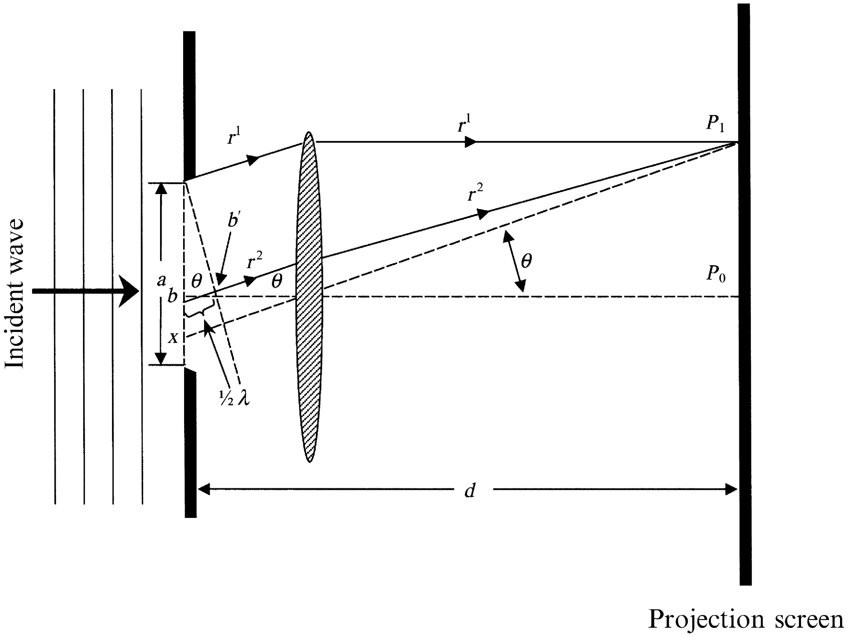

**FIGURE 2.6**

The single-slit experiment showing the source of the phase shifts geometrically.

Fraunhoffer (we define in the appendixes what we mean by Fraunhoffer diffraction) diffraction requires summing or integrating over the entire surface of the slit and is illustrated in Fig. 2.7. Effectively, the single slit acts as an infinite set of slits. Indeed, you are probably more familiar with diffraction produced by a grating, which is an infinite set of equally spaced slits separated by some fixed distance.

## 2.5 THE AIRY DISK AND THE ISSUE OF MICROSCOPE RESOLUTION

We are now in a position to turn our attention to how interference affects microscope images: How does a microscope treat a point source of light? You might ask, why do we care? We care because ultimately all objects can be represented as the sum of an infinite set of point sources. If we know how the microscope treats (distorts, if you will) a point object, we know how it treats any object. This is the concept of the point spread function because a point of light is confused or spread out by both the optical and the electronic systems of the digital or video microscope system. We will encounter this concept progressively throughout this volume. This spreading is an example of what mathematicians refer to as convolution. In Chapter 9 by Salmon et al., this volume, we will be introduced to the concept of convolution.

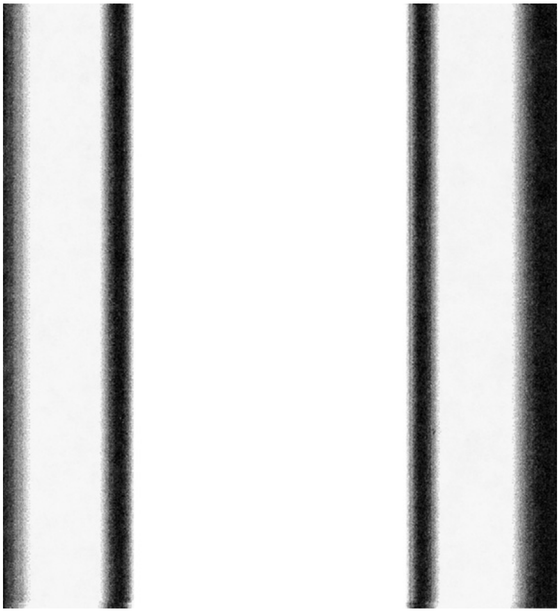

FIGURE 2.7

The single-slit diffraction pattern.

Unfortunately, the reverse process, which we might refer to as deconvolution, is not mathematically trivial. Later in this chapter, we will introduce the concept of Fourier series and transforms, both as an introduction to the idea of spatial frequencies and to provide the mathematical tools necessary to understand deconvolution. The concept of deconvolution using Fourier transforms will finally be considered in Chapter 16 by Salmon and Tran, this volume; and Chapter 14 by Wolf et al., this volume.

In Fig. 2.8, we illustrate how a microscope views a point source of light. One expects from consideration of geometric optics that a point of light at A will be focused by the objective to a point on the image plane. However, Huygens teaches us that the two points on the wave front act as generation sites for Huygens' wavelets. Although most of the light goes to the center point on the image plane, where it constructively interferes to produce a bright spot, some of it also goes to other points. At the second point shown in the figure, the two waves will be phase-shifted. The result is an interference pattern on the image plane. Rather than focusing the point source to a spot, the actual image shown in Fig. 2.9 is known as the Airy disk, after George Biddell Airy (1801–1892). The ability of the microscope to resolve two spots as two ultimately relates to how sharply the Airy disks are defined (Fig. 2.10). For self-luminous point sources, such as the ones obtained in fluorescence microscopy, the Airy disks are independent and do not interfere with one another.

An empirical criterion for resolution for such self-luminous objects that is related to the separation where the primary maximum of the second Airy disk coincides with the primary minimum of the first is

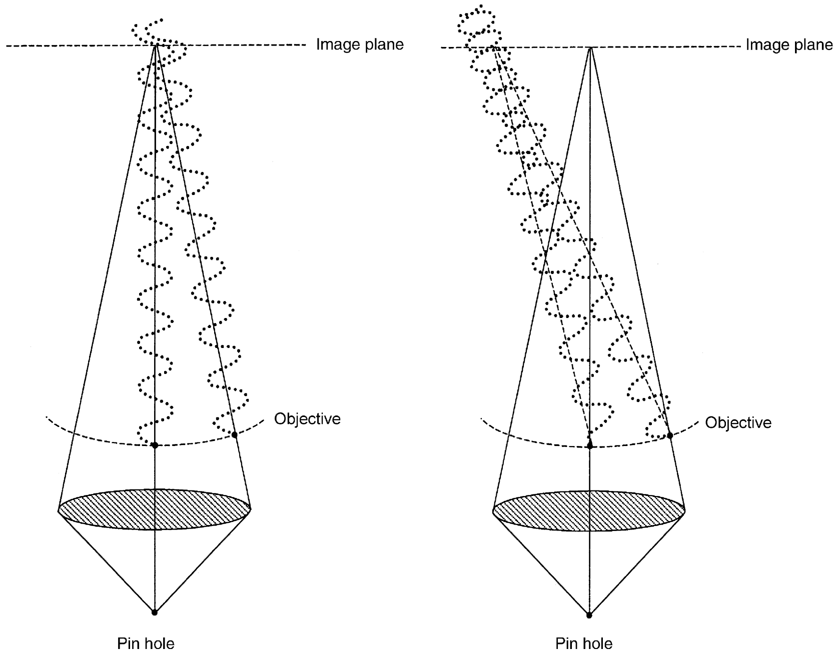

**FIGURE 2.8**

A point of light viewed in the microscope showing how interference leads to the Airy disk pattern.

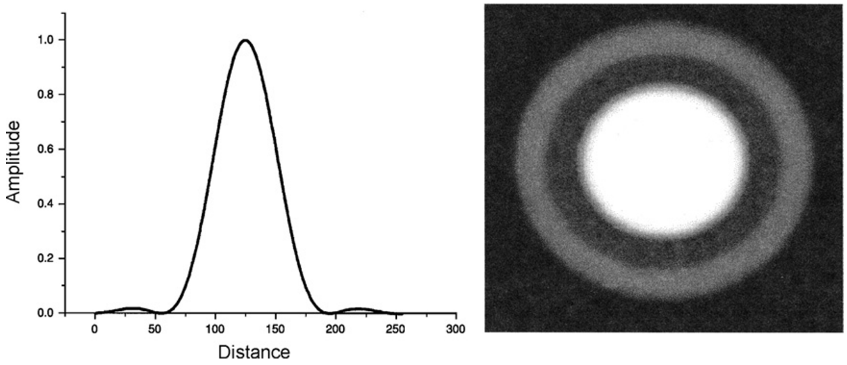

**FIGURE 2.9**

The Airy disk: Notice how the intensity of the secondary ring is only about 2% of the intensity of the primary maximum. The ability to see this ring in the microscope offers an empirical test of noise rejection in the microscope.

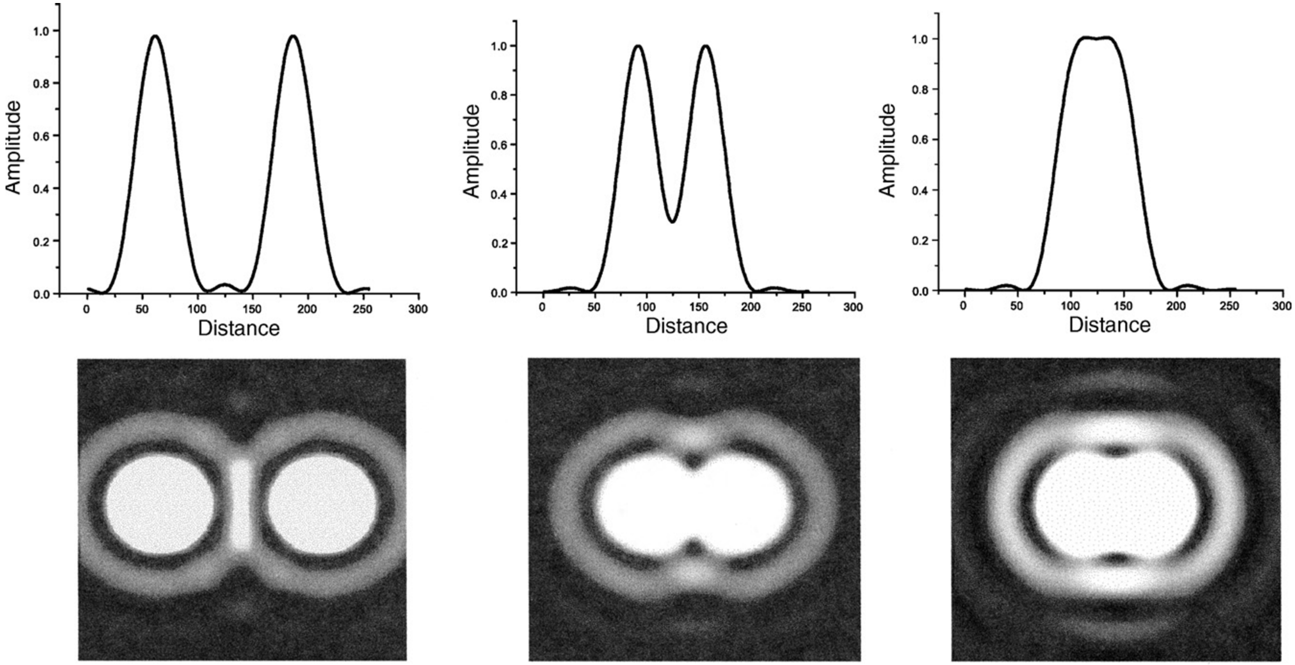

**FIGURE 2.10**

Defining resolutions as the ability to distinguish two Airy disks: as the disks come closer, the intensity dip between them becomes smaller. Resolution definitions such as Rayleigh's criteria set are based on the ability of the human eye to detect this dip. Other criteria may be more appropriate for digital systems.

$$r = \frac{1.22\lambda_0}{2\mathrm{NA}_{\mathrm{obj}}} \tag{2.14}$$

and is known as Rayleigh's criterion. The NA (numerical aperture) of an objective is related to the collection angle $\theta$ and is given by

$$\mathrm{NA} = n\sin(\theta), \tag{2.15}$$

where $n$ is the index of refraction. For bright field images, the point sources will be partially coherent, and the resolution will also be dependent on the NA of the condenser. Thus, for bright field point sources, resolution is given by

$$r = \frac{1.22\lambda_0}{\mathrm{NA}_{\mathrm{obj}} + \mathrm{NA}_{\mathrm{cond}}}. \tag{2.16}$$

In a modern context, it is important to recognize that Rayleigh's criterion is empirical and ultimately based on the human eye's ability to resolve intensity differences. It asks the question, or more accurately, states the answer to the question, what is the minimum intensity dip between the two maxima required for the human eye to resolve them as two Airy disks? With modern image detectors and image processing, it is possible to push these limits. We will explore this issue further in subsequent chapters.

On deeper consideration of Fig. 2.8, the reader may be tempted to cry foul. There is a point at which the spherical wavelets emitted from the two points destructively interfere, but what about wavelets from all of the other generating centers between these points? In reality, we must consider and superimpose all of the wavelets to generate the composite field at any point $P$. In Appendix B, we will discuss Kirchoff's approach to this problem, and in Appendix C, we will apply the approach to the problem of the Airy disk.

## 2.6 FOURIER OR RECIPROCAL SPACE: THE CONCEPT OF SPATIAL FREQUENCIES

The issue of resolution also relates to the amount of contrast (the difference between light and dark) and the sharpness of edges. To understand these questions, we must take a detour into what may at first appear to be a strictly mathematical domain. This is the concept of Fourier, or reciprocal, space. This concept is fundamental to understanding the optics of microscope image formation.

We must begin with the concept of a spatial frequency. Consider, for instance, the picture in Fig. 2.11A. When we look at this (or any) scene, our first inclination is to say that the sizes of the objects are random. On closer examination, we realize that this is not really the case. Every scene has certain characteristic spacings in it. We have drawn some of these characteristic frequencies as sine$^2$ waves over the figure. Mathematically, these spacings are referred to as spatial frequencies (more accurately, we might refer to spatial wavelengths, $L$, and define the spatial frequencies as $2\pi/L$). In every scene, certain spatial frequencies dominate over others. This

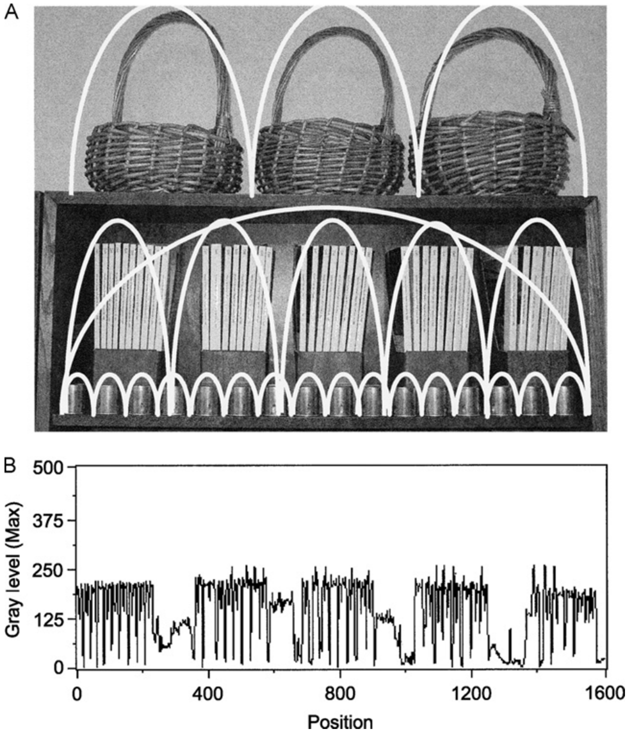

**FIGURE 2.11**

(A) Every object or image has characteristic spatial frequencies. (B) A line scan across the books in Fig. 2.11.

concept may seem very abstract at first. When you think about it, one point that you will probably recognize is that fine detail is described by high spatial frequencies. So in some sense, the highest spatial frequency, which we can detect, must be related to the resolution limit of our optical system. If you take a magnifying glass and examine Fig. 2.11A closely, you will see dots. What this tells us is that there is a cutoff, a highest spatial frequency, beyond which the printer did not need to go to accurately portray the photograph. This is very clear in digital photography. When you look at a high-resolution digital photograph, it is nice and sharp; as you zoom in, however, you will eventually see the pixels. The maximum spatial frequency used in the photograph is $2\pi$ divided by the interpixel distance.

One might think that although light waves can be represented by sine waves, or more accurately a sum of sine waves, this might not be the case for all waves. Let us phrase this differently: Can we put this abstract concept of spatial frequencies into a coherent mathematical context? To approach this problem, let us simplify it a bit. Suppose that rather than dealing with the two-dimensional image of Fig. 2.11A, we scan one horizontal line across this image and look at the intensity as a function of position. Such a line is shown in Fig. 2.11B. Again, the information is not random. We can clearly see characteristic spatial frequencies.

Jean-Baptiste Fourier (1768–1830), in the early nineteenth century, demonstrated the remarkable fact that any wave or function could be described as a sum of sine waves (actually, sine + cosine waves). Rather than work with the scan of Fig. 2.11B, let us consider a simpler but more challenging pattern, namely, that of a step function, as illustrated below in Fig. 2.12A—a step function that is an

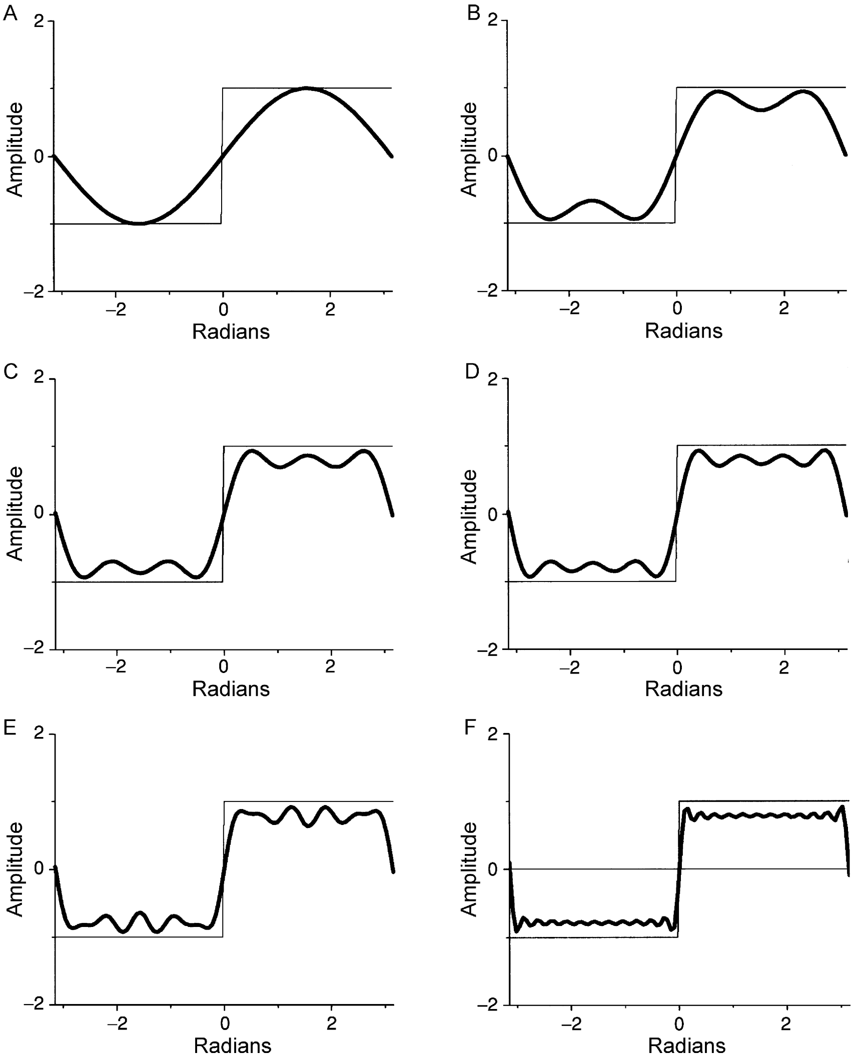

**FIGURE 2.12**

Fourier's construction of a step from a composite of sine waves: (A) sin *x*, (B) sin *x* + 1/3 sin 3*x*, (C) sin *x* + 1/3 sin 3*x* + 1/5 sin 5*x*, (D) sin *x* + 1/3 sin 3*x* + 1/5 sin 5*x* + 1/7 sin 7*x*, (E) sin *x* + 1/3 sin 3*x* + 1/5 sin 5*x* + 1/7 sin 7*x* + 1/9 sin 9*x*, and (F) up to the term 1/23 sin 23*x*. Notice that even in (F) there is ringing of the intensity pattern.

alternating pattern of white and black bands. One might think that nothing is farther from a sine wave.

However, think about it a little more closely. If we draw a sine wave with the same spacing, it approximates the step function (Fig. 2.12A). This is the lowest spatial frequency, $2\pi/L$. We next try to fill in the gaps by adding some amount of a sine wave with spatial frequency $3(2\pi/L)$ (Fig. 2.12B). As one successively adds sine waves of shorter and shorter wavelength [spatial frequencies $m(2\pi/L)$, where $m$ is odd] (Fig. 2.12C–F) in the appropriate proportions, the resulting sum looks more and more like a step function. To achieve perfect match, one has to use an infinite number of sine waves. However, as the amplitudes of the waves decrease rapidly and monotonically, any given degree of approximation can be achieved with a finite number of sine functions. It is significant to note that when one cuts off the summation at a finite number of sine waves, the resultant wave shows oscillations or "ringing," which is most apparent at the edge of the step (Fig. 2.12F). As we will see, the optical ringing at the edges of objects in the light microscope is caused by incomplete reconstruction of Fourier series. Indeed, microscopes always act as low-pass filters with some maximum spatial frequency making it through, and as we shall see, is the basis of resolution in the microscope.

In Fig. 2.13, we show the amplitudes of the spatial frequencies of the step function. Such a plot is referred to as the Fourier transform of the original function. Note that in the case of the step function, the function is not continuous but rather discrete.

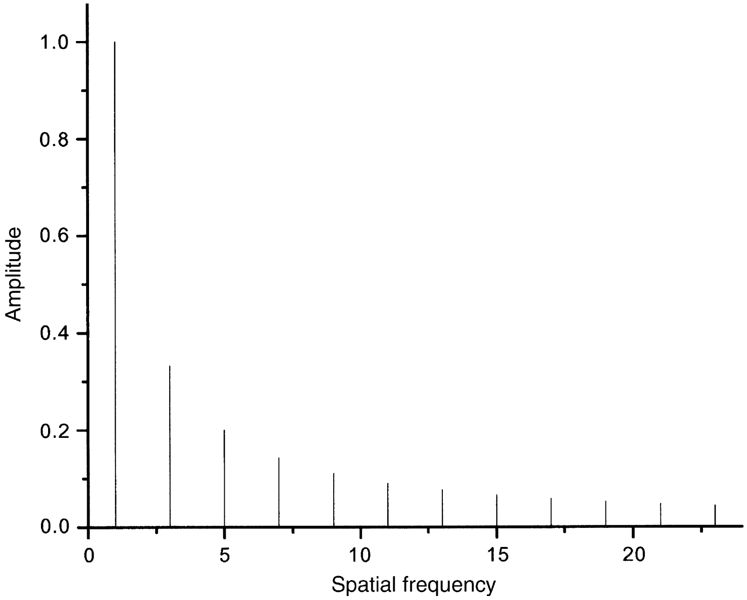

**FIGURE 2.13**

The amplitudes of the frequency terms: this is the so-called Fourier transform of the step function.

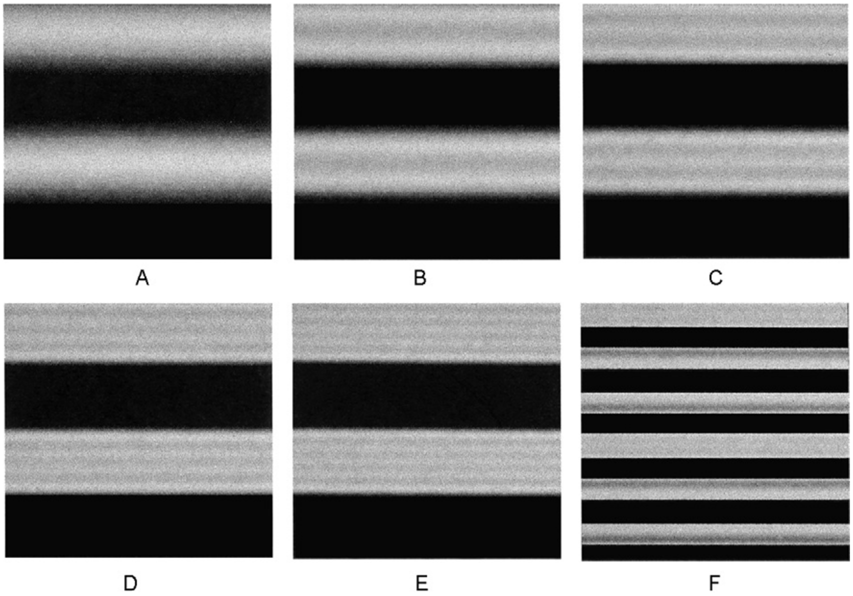

**FIGURE 2.14**

The Fourier series of Fig. 2.13 shown as an image; (A–E) correspond to Fig. 2.13. In (F), we have removed the sin *x* term, which causes a doubling of the frequency pattern.

That is, only those spatial frequencies that are odd numbers have nonzero amplitude. In fact, for a spatial frequency *m*, the amplitude is 1/*m*.

In Fig. 2.14, we use the Fourier series to recreate the step function as an image. In Fig. 2.14A, we have generated an image with a sinusoidally varying intensity. We can get a sense of the bars of the step function, but the edges are very fuzzy. In Fig. 2.14B–E, we progressively add the sin 3×, sin 5×, sin 7×, and sin 9× terms. With the addition of each higher-order spatial frequency, the edges become more and more distinct. However, in each case we can see ringing at the edges because we lack higher order terms.

Next we ask the question, what will happen if we eliminate the first term in the series? This is shown in Fig. 2.14F. What happens is that we now have a perfectly reasonable looking step function, only its bars are now twice as close as before.

In this discussion, we have considered a one-dimensional Fourier series. One can also perform Fourier transforms on two-dimensional patterns (i.e., images) or even three-dimensional patterns such as a *z*-stack of images. In Appendix A, we develop the mathematics of Fourier series and transforms. Reducing a wave to a Fourier series is derived, following Carl Friedrich Gauss (1777–1855), as being essentially equivalent to a problem in nonlinear least squares curve fitting: the solution of the step function is derived explicitly, the integral form is developed, and finally, the application of Fourier transforms to solving the wave equation of electrodynamics and optics is discussed.

## 2.7 RESOLUTION OF THE MICROSCOPE

Physical systems such as microscopes tend to pass only a certain set of spatial frequencies. In a sense, the highest frequency passed becomes a measure of resolution. An object such as diffraction grating, which is essentially the step wave shown in Fig. 2.14, requires all spatial frequencies to see it correctly. How a microscope treats such an object is illustrated in Fig. 2.15. When the grating is illuminated with a bundle of planar light, say at the center, the zeroth-order undiffracted light propagates straight through as one bundle, whereas higher orders are diffracted and propagate as parallel bundles at appropriate angles. Because all of these bundles are parallel, they will be focused by the objective at the back focal plane, with higher orders appropriately displaced from the center. A Fourier transform of the image forms at the

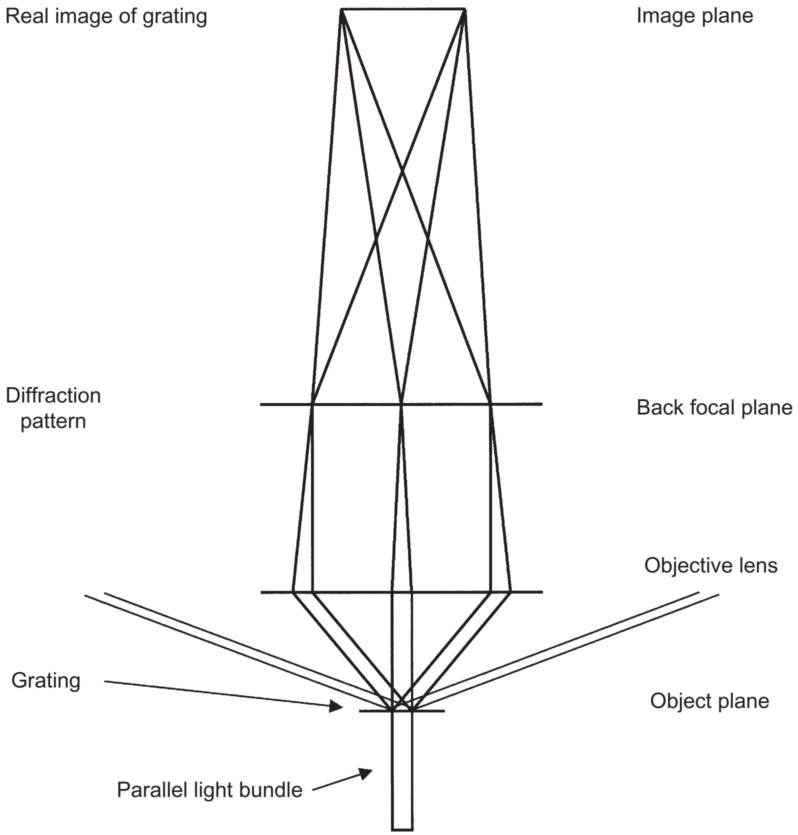

**FIGURE 2.15**

In the microscope, the resolution limit results from the microscope's inability to collect all spatial frequencies. Objects act as diffraction gratings and Fourier transform of the object forms at the back focal plane.

back focal plane. These bundles recombine at the image plane to form a real image of the diffraction grating. You can see the problem immediately: Not all of the orders will make it through the objective. There is a cutoff and higher orders fail to enter the objective. The larger the collection angle (the NA) of the objective, the higher the cutoff and the higher the resolution. So far, we have discussed what is going on at the center of the object. The situation becomes more complex as one moves away from the center. The cutoff frequency on the two sides becomes asymmetric. The resolution limit is effectively defined by the highest spatial frequency that enters the objective.

In Fig. 2.16, we image the back focal plane for Fourier transform of biological specimens. We have chosen specimens with some well-defined structure.

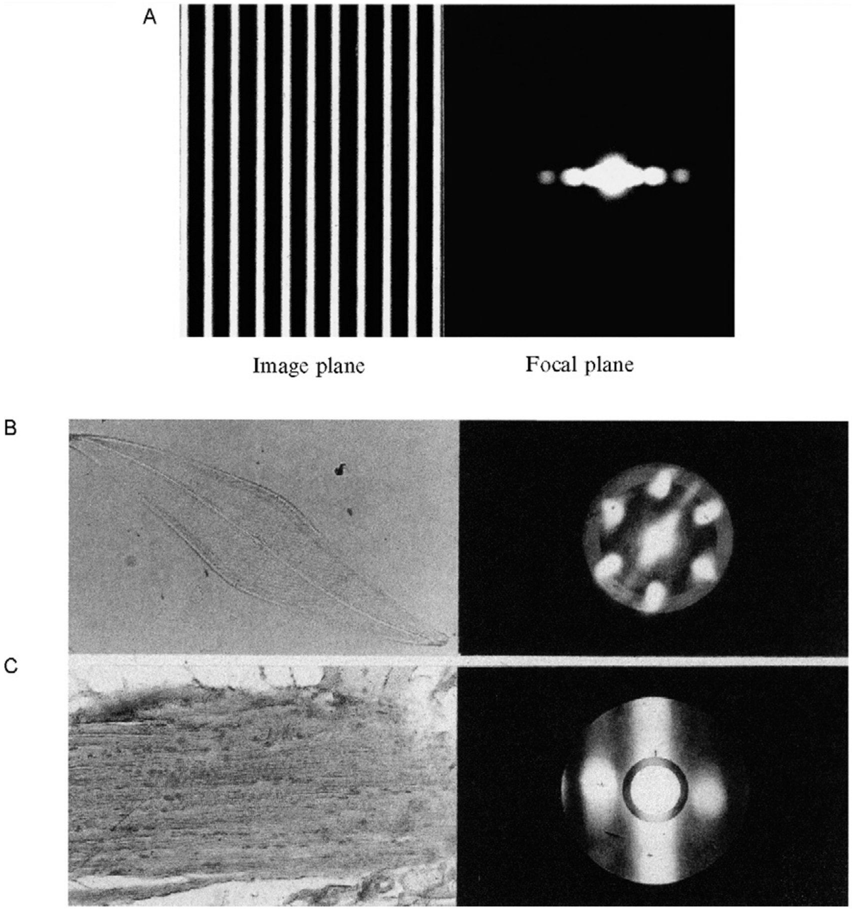

**FIGURE 2.16**

Fourier patterns at the back focal plane of the objective for: (A) a Ronchi ruling, (B) a diatom, and (C) a muscle fiber. Courtesy of Greenfield Sluder.

Figure 2.16A shows a diatom and Fig. 2.16B a muscle section. For comparison, we also show the specimens viewed in the image plane.

## 2.8 RESOLUTION AND CONTRAST

We have previously seen how resolution can be defined in terms of our ability to distinguish between two Airy disks. As the disks come closer and closer, which corresponds to trying a higher and higher spatial frequency, the contrast between the two disks decreases. The resolution limit is reached when the contrast between the two points becomes sufficiently small such that they can no longer be distinguished as two separate points. We see that image contrast and resolution become highly intertwined factors. This is illustrated in Fig. 2.17, where sine waves of increasing spatial frequency (left to right) and increasing contrast (bottom to top) are displayed as an image; the higher the contrast, the higher the detectable spatial frequency. In Fig. 2.18, we show this same fact for hypothetical microscope

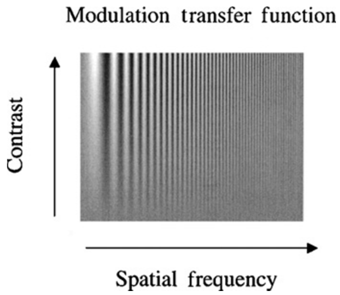

**FIGURE 2.17**

Increasing contrast increases resolution (the highest detectable spatial resolution).

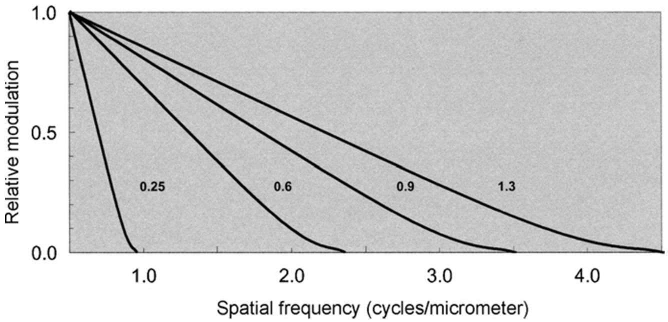

**FIGURE 2.18**

The relationship between contrast (relative modulation) and resolution (highest discernable spatial frequency) for objectives with NA = 0.25, 0.6, 0.9, and 1.3.

objectives. Resolution (highest collected spatial frequency) clearly increases with increasing NA. However, for a given objective, the greater the contrast, the greater the resolution becomes.

## CONCLUSIONS

In this chapter, we have described Abbe's theory of the microscope-relating resolution to the highest spatial frequency that a microscope can collect. We have shown that this limit increases with increasing NA. As a corollary to this, resolution increases with decreasing wavelength because of how NA depends on wavelength. The resolution is higher for blue light than for red light. Finally, we see that resolution is dependent on contrast and that the higher the contrast, the higher the resolution.

Ultimately, this last point relates to issues of signal-to-noise and dynamic range. The use of video and new digital cameras has necessitated redefining classical limits such as those of Rayleigh's criterion. In subsequent chapters, we will further explore these critical issues.

## 2.9 APPENDIX A

### 2.9.1 Fourier series

In this chapter, we indicated that a light wave could be represented by the equation

$$E(x_1 t) = E_0 \mathrm{e}^{-i(kx-\omega t)}. \tag{2.A1}$$

One might ask, where this comes from? We know that the mathematical form of the wave must satisfy Maxwell's wave equation, and we can, indeed, show that it is a solution of this equation. Yet we may still ask to what extent is it the only solution? It was considerations such as these that led to the development of Fourier series and transforms. At the end of this appendix we will see why. We begin by exploring briefly the concept of the Fourier series and their importance in optics. For a more extensive discussion, the reader is referred to the elegant treatment by Arnold Sommerfeld in his book *Partial Differential Equations in Physics* (Sommerfeld, Sommerfeld, 1949b).

Suppose that one has an arbitrary function, a wave form, $f(x)$, which is defined over the interval of $-\pi$ to $\pi$ (by suitable rescaling, any interval $-L/2$ to $L/2$ can be redefined to meet this criterion). We next suppose that this function can be approximated by a function $S_n(x)$, which is the sum of a series of $n$ sine waves and $n$ cosine waves with coefficients $A_m$ and $B_m$ plus a constant $A_0$. That is,

$$S_n(x) = A_0 + \sum_{m=1}^{n} \{A_m \sin(mx) + B_m \cos(mx)\}. \tag{2.A2}$$

The problem, then, is to determine the $2n+1$ coefficients, $A_m$ and $B_m$, that define the $S_n(x)$. This problem is equivalent to the problem of fitting experimental data to a curve or mathematical function. As in the curve-fitting problem, we recognize that for every value of $x$, there is a deviation or difference $\delta(x)$ between the approximation $S_n(x)$ and the function $f(x)$. That is,

$$f(x) = S_n(x) + \delta(x). \tag{2.A3}$$

The goal is to minimize the sum of the squares of the deviations over the entire interval $-\pi$ to $\pi$. That is, we seek to minimize $\chi^2$, where

$$\chi^2 = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \delta_n^2(x)\mathrm{d}x. \tag{2.A4}$$

Minimization requires that all of the partial derivatives of $\chi^2$ with respect to the $A_m$'s and the $B_m$'s be simultaneously equal to 0. For an arbitrary value of $m$, we have that,

$$\begin{aligned} -\frac{\partial\chi^2}{\partial A_m^2} &= \frac{1}{\pi}\int_{-\pi}^{+\pi}\{f(x) - S_n(x)\}\cos(mx)\mathrm{d}x = 0 \text{ and} \\ -\frac{\partial\chi^2}{\partial B_m^2} &= \frac{1}{\pi}\int_{-\pi}^{+\pi}\{f(x) - S_n(x)\}\sin(mx)\mathrm{d}x = 0. \end{aligned} \tag{2.A5}$$

One can then solve Eq. (2.A5) for the $2n+1$ coefficients by putting Eq. (2.A2) into Eq. (2.A5) and making use of the orthogonal properties of the sine and cosine functions, namely, that:

$$\frac{1}{\pi}\int_{-\pi}^{+\pi}\cos(px)\sin(qx)\mathrm{d}x = 0 \quad \text{for (all values of } p \text{ and } q), \tag{2.A6a}$$

$$\begin{aligned} \frac{1}{\pi}\int_{-\pi}^{+\pi}\cos(px)\cos(qx)\mathrm{d}x &= 0 \quad \text{for} \quad p \neq q, \\ &= 1 \quad \text{for} \quad p = q, \end{aligned} \tag{2.A6b}$$

$$\begin{aligned} \frac{1}{\pi}\int_{-\pi}^{+\pi}\sin(px)\sin(qx)\mathrm{d}x &= 0 \quad \text{for} \quad p \neq q, \\ &= 1 \quad \text{for} \quad p = q. \end{aligned} \tag{2.A6c}$$

Substituting the conditions of Eq. (2.A6) as well as the trivial condition that

$$\frac{1}{2\pi}\int_{-\pi}^{+\pi}\mathrm{d}x = 1, \tag{2.A7}$$

we find that

$$A_m = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(x)\cos(mx)\mathrm{d}x, \tag{2.A8a}$$

$$B_m = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(x)\sin(mx)\mathrm{d}x, \tag{2.A8b}$$

$$A_0 = \frac{1}{2\pi}\int_{-\pi}^{+\pi} f(x)\mathrm{d}x. \tag{2.A8c}$$

The $2n+1$ coefficients can be determined by means of Eq. (2.A8).

A rigorous mathematical treatment must demonstrate three critical points: that as $n \to \infty$, then $\chi^2$, indeed $\delta(x)$ for all values of $x$, goes to 0; that a solution exists; and that the solution is unique. Here, we will refer the reader to Sommerfeld et al., 1949 on these three points and go on to examine the problem of the step function.

The step function is given by

$$\begin{aligned} f(x) &= -1 \quad \text{for} \quad -\pi < x < 0 \\ f(x) &= +1 \quad \text{for} \quad 0 < x < +\pi. \end{aligned} \tag{2.A9}$$

Before applying Eqs. (2.A7) and (2.A8), we recognize that as we are centering the step function around a DC level of 0 then $A_0 = 0$. Furthermore, we recognize that as the function must have a node at $-\pi$, 0, and $+\pi$, the function is a series in sines with odd coefficients, $m$. This last condition means that all of the $B_m$'s equal 0 and the $A_m$'s equal 0 where $m$ is even. Inserting Eq. (2.A9) into Eq. (2.A8a) gives us the Fourier series for the step function, namely,

$$f(x) = \sum_{m=0}^{\infty}\left(\frac{1}{2m+1}\right)\sin[(2m+1)x]. \tag{2.A10}$$

Related to the Fourier series is the concept of the Fourier transform. The function in real space is described by $f(x)$. We alternatively speak of the function being described in frequency or reciprocal space. This frequency space description is referred to as the Fourier transform, $\mathfrak{J}(v) = \mathfrak{J}\{f(x)\}$, of the function. For the step functions, $\mathfrak{J}(v) = 0$ except for $\nu = 2m+1$, where $\mathfrak{J}(v) = 1/(2m+1)$ and $m$ is an integer. For completeness we also speak of the inverse Fourier transform $\mathfrak{J}^{-1}$. This leads to the trivial identity that

$$\mathfrak{J}^{-1}[\mathfrak{J}\{f(x)\}] = f(x). \tag{2.A11}$$

The requirement that the function have nodes at $-\pi$, 0, and $+\pi$, or more generally, the requirement that the function is defined over a finite interval with specific boundary values, leads to series like that of Eq. (2.A9) wherein the series contains only a subset of all possible sine waves defined by the eigenvalues $m$. In the more general case, all possible sine and cosine waves represent possibilities, and the summation in the equation is replaced by an integral. In such cases, it is more convenient to use the exponential form of the wave, $\mathrm{e}^{ikx}$. The connection between this representation and the sine/cosine representation has already been discussed in the text of the chapter and is contained in the trigonometric identities

$$\sin(kx) = \frac{\mathrm{e}^{ikx} - \mathrm{e}^{-ikx}}{2i}, \tag{2.A12a}$$

$$\cos(kx) = \frac{\mathrm{e}^{ikx} + \mathrm{e}^{-ikx}}{2i}. \tag{2.A12b}$$

We may then define the Fourier transform of the function $f(x)$ as

$$\mathfrak{J}(v) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x)\mathrm{e}^{-ivx}\mathrm{d}x, \tag{2.A13a}$$

and the inverse Fourier transform by

$$\mathfrak{J}^{-1}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathfrak{J}(v)\mathrm{e}^{ivx}\mathrm{d}v. \tag{2.A13b}$$

We have here derived the concept of Fourier series and Fourier transforms as a means of introducing the fact that the diffraction pattern is the Fourier transform of the object and that the diffraction pattern may be observed at the back focal plane of the microscope objective. As the reader will see in subsequent chapters (i.e., Chapter 9 by Salmon et al., this volume; Chapter 16 by Salmon and Tran, this volume; and Chapter 14 by Wolf et al., this volume), the Fourier transform becomes a powerful tool in digital deconvolution techniques to remove out-of-focus fluorescence. Ultimately, this results from the fact that the Fourier transform of the object is the diffraction pattern observed at the back focal plane of the objective. Indeed, some laboratories have created Fourier filters of the image not by digital transformation but by physically masking specific frequencies at the back focal plane (Hui & Parsons, 1975).

Fourier transforms also are very powerful tools for solving partial differential equations such as the wave equations. Let us suppose, as an example, that we have an electromagnetic wave, which is propagating in the $x$-direction and polarized so that it is confined to the $z$-direction. Then we can write the wave equation for the electric field as:

$$\frac{\partial^2 E}{\partial t^2} = c^2 \frac{\partial^2 E}{\partial z^2}. \tag{2.A14}$$

We next Fourier transform both sides of Eq. (2.A14). If we reverse the order of integration and differentiation, we obtain

$$\frac{\mathrm{d}^2 \mathfrak{J}(E)}{\mathrm{d}t^2} = -v^2 c^2 \mathfrak{J}(E). \tag{2.A15}$$

Note that this is a simple rather than a partial differential equation. We can then write,

$$\mathfrak{J}(E) = \mathfrak{J}_0(E)\mathrm{e}^{-ivct}. \tag{2.A16}$$

All that remains to be done to determine $E$ is to inverse Fourier transform Eq. (2.A14). Let us assume that $E(x) = E_0\mathrm{e}^{ikx}$. Fourier has already taught us that this is not a big assumption, as even if it is not true, it will be the case that the wave can be represented as a sum or distribution of terms $E_0\mathrm{e}^{ikx}$. Then the Fourier transform is determined by putting Eq. (2.A11) into Eq. (2.A14):

$$\Im(E) = E_0 \mathrm{e}^{-ivct} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{i(k-v)x} \mathrm{d}x. \tag{2.A17}$$

The integral Eq. (2.A17) is the Dirac Delta Function given by

$$\delta(k - v) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{i(k-v)x} \mathrm{d}x, \tag{2.A18}$$

then

$$\Im(E) = E_0 \mathrm{e}^{-ivct} \delta(k - v). \tag{2.A19}$$

The critical property of the Dirac Delta Function is that it is 0 for all values of $s$ except for $s = k$, where the function has the value of 1.

We next inverse Fourier transform Eq. (2.A19)

$$E = E_0 \int \mathrm{e}^{-i(vct - vx)} \delta(k - v) \mathrm{d}v. \tag{2.A20}$$

The properties of the Dirac Delta Function lead us to

$$E = E_0 \mathrm{e}^{i(kx - kct)}, \tag{2.A21}$$

remembering that

$$k = \frac{2\pi}{\lambda} = \frac{2\pi\upsilon}{c} = \frac{\omega}{c}. \tag{2.A22}$$

We obtain

$$E = E_0 \mathrm{e}^{i(kx - \omega t)}. \tag{2.A23}$$

We have now come full circle. We have shown that $\mathrm{e}^{ikx}$ is a solution to the spatial part of the wave equation. We have further shown that any solution can be represented as a Fourier sum or distribution of such solutions. Finally, we have used the Fourier transform to demonstrate that given the form $\mathrm{e}^{ikx}$ of the spatial dependence, $\mathrm{e}^{-iwt}$ must be the solution of the temporal dependence.

## 2.10 APPENDIX B

### 2.10.1 Kirchoff's scalar theory of diffraction: Recasting Huygens' principle in an electrodynamic context

Our goal in this appendix is to describe Kirchoff's formulation of Huygens' principle in terms of electrodynamic field theory. Here we will follow the derivation of Sommerfeld in his book *Optics, Lectures on Theoretical Physics* (Sommerfeld et al., 1949). We present an outline of the theory, and the reader is referred to the original for more details. Other texts with excellent discussions of this problem and the problem of the diffraction pattern from a circular aperture are *An Introduction to Fourier Optics* (Goodman, 1968) and *Electrodynamics* (Jackson, 1975). The latter is particularly

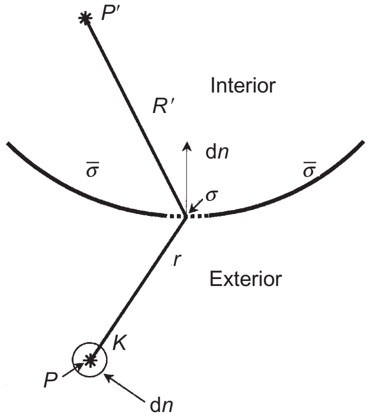

FIGURE 2.A.1

The generalized condition of an opaque surface $\overline{\sigma}$ with an aperture $\sigma$.

valuable as it treats the full vectorial solution rather than the scalar approximation. The reader is further referred to *Principles of Optics* (Born & Wolf, 1980).

### 2.10.2 Generalizing the problem

Figure 2.A.1 represents a generalized view of diffraction problems. We have a point $P$ and a surface. The surface consists of an opaque region, $\overline{\sigma}$, and an aperture region, $\sigma$. The basic question is how the field at $P$ relates to the field in the aperture, $\sigma$.

For mathematical purposes, we define an exterior and an interior to the surface and we create a second infinitesimally small spherical surface $K$ around the point $P$. This serves to isolate the singularity at $P$ and to create a closed surface for purposes of investigation. In the figure, d$n$ represents the normal to the surface at any point $r$.

### 2.10.3 Scalar spherical waves

In this chapter, we have dealt with plane waves of the form $\mathrm{e}^{ikx}$, which depend only on the coordinate $x$. Here we must deal with a scalar spherical wave, which depends only on the radial value $r$. Such waves must, in general, be solutions of the equation

$$\frac{1}{r}\cdot\frac{\partial^2(ru)}{\partial r^2}+k^2u=0. \tag{2.B1}$$

Equation (2.B1) is the wave equation in spherical coordinates assuming spherical symmetry, that is, no angular dependence.

The so-called principal solution of Eq. (2.B1) has the form

$$u=\frac{1}{r}\mathrm{e}^{ikr}. \tag{2.B2}$$

### 2.10.4 Green's theorem

Here we need to integrate over the surface $\sigma$ to find our solution. This is accomplished using Green's function, $G$, to find the solution, which we call $v_p$ for the field at $P$. In general, Green's theorem tells us that

$$\int\left\{G\frac{1}{r}\left(\frac{\mathrm{d}^2 rv}{\mathrm{d}r^2}\right)-v\frac{1}{r}\times\frac{\mathrm{d}^2 rG}{\mathrm{d}r^2}\right\}\mathrm{d}\tau=\int\left(G\frac{\partial v}{\partial n}-v\frac{\partial G}{\partial n}\right)\mathrm{d}\sigma, \tag{2.B3}$$

where d$\tau$ represents the volume element within the surface and d$\sigma$ represents the surface element.

The use of $G = u$ leads to contradictions because of the requirement that $v = 0$ and $\partial v/\partial n = 0$ everywhere on the surface $\overline{\sigma}$. Instead, we define $G$ by the conditions that

$$\frac{1}{r}\times\frac{\partial^2 rG}{\partial r}+k^2G=0 \quad \text{within the volume,} \tag{2.B4a}$$

$$G=0 \text{ on } \sigma, \tag{2.B4b}$$

$$G\rightarrow u \text{ as } r\rightarrow 0, \tag{2.B4c}$$

$$r\left(\frac{\partial G}{\partial n}-ikG\right)\rightarrow 0 \text{ as } r\rightarrow\infty. \tag{2.B4d}$$

It is Eq. (2.B4d) that enables the boundary conditions on $v$ without contradiction. If we define the volume as the exterior exclusive of the volume surrounding $P$ by $K$, then the left-hand side of Eq. (2.B3) vanishes. Integration of the right-hand side over $K$ leads to $-4\pi v_p$. Thus, Eq. (2.B3) reduces to

$$4\pi v_p=\int_\sigma\left\{\frac{\partial v}{\partial n}G-v\frac{\partial G}{\partial n}\right\}\mathrm{d}\sigma. \tag{2.B5}$$

Applying the condition $\partial v/\partial n = 0$ on $\sigma$, Eq. (2.B5) becomes

$$4\pi v_p=\int_\sigma -v\frac{\partial G}{\partial n}\delta\sigma. \tag{2.B6}$$

The boundary conditions on $v$ are

$$v=0 \text{ on } \overline{\sigma}, \tag{2.B7a}$$

$$v=\frac{A\ \exp\ ikr'}{r'} \text{ on } \sigma. \tag{2.B7b}$$

### 2.10.5 Solution for a plane

This problem simplifies where the screen or surface is a plane. In that case, Green's function can be conveniently expressed by the method of images [to satisfy the condition in Eq. (2.B4b) that $G = 0$ on $\sigma$]. Referring to Fig. 2.A.2, if the point $P$

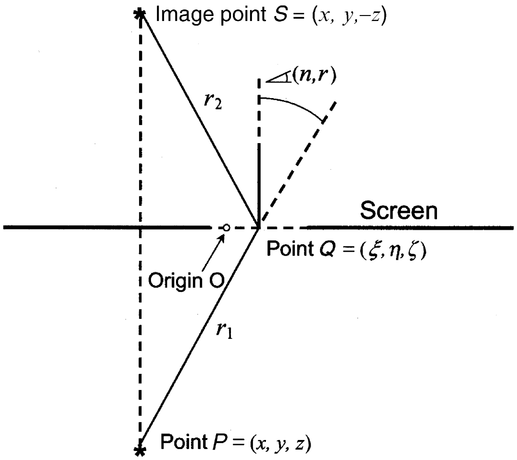

**FIGURE 2.A.2**

Method of images applied to the generalizable diffraction problem.

is at arbitrary coordinate $(x,y,z)$, then the image point is at coordinates $(x,y,-z)$. For an arbitrary point $Q$ with coordinates $(\xi,\eta,\zeta)$, we have

$$G = \frac{e^{ikr_1}}{r_1} - \frac{e^{ikr_2}}{r_2}, \tag{2.B8a}$$

where

$$r_1^2 = (\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2 \tag{2.B8b}$$

and

$$r_2^2 = (\xi - x)^2 + (\eta - y)^2 + (\xi + z)^2. \tag{2.B8c}$$

We next calculate

$$\frac{\partial G}{\partial \zeta} = \frac{d}{dr_1}\left(\frac{e^{ikr_1}}{r_1}\right)\frac{\partial r_1}{\partial \zeta} - \frac{d}{dr_1}\left(\frac{e^{ikr_1}}{r_1}\right)\frac{\partial r_2}{\partial \zeta}, \tag{2.B9}$$

$$\frac{\partial G}{\partial n} = \frac{-\partial G}{\partial \zeta} - 2\frac{\partial}{\partial r}\left(\frac{e^{ikr}}{r}\right)\cos(n,r). \tag{2.B10}$$

We then have that

$$\begin{aligned}\frac{\partial}{\partial r}\left(\frac{e^{ikr}}{r}\right) &= \frac{ike^{ikr}}{r} - \frac{1}{r^2}e^{ikr} \\ &= \frac{ike^{ikr}}{r}\left(1 - \frac{1}{ikr}\right).\end{aligned} \tag{2.B11}$$

In the limit of large $r$ we have that $kr = 2\pi r/\lambda \gg 1$ and

$$\frac{\partial}{\partial}\left(\frac{\mathrm{e}^{ikr}}{r}\right) \sim \frac{2\pi i}{\lambda}\frac{\mathrm{e}^{ikr}}{r}. \tag{2.B12}$$

Putting Eq. (2.B12) into Eq. (2.B10), and that into Eq. (2.B6), we can solve for $v_p$.

$$i\lambda v_p = \int_\sigma \frac{\mathrm{e}^{ikr}}{r}\cos(n,r)v\mathrm{d}\sigma. \tag{2.B13}$$

### 2.10.6 Huygens' principle

Equation (2.B13) is mathematically equivalent to Huygens' principle. A light wave falling on the aperture $\sigma$ propagates as if every element $\delta\sigma$ emitted a spherical wave, the amplitude and phase of which are given by that of the incident wave $v$.

If we assume point source illumination, then

$$v = \frac{A\mathrm{e}^{ikr'}}{r'}, \tag{2.B14}$$

in which case,

$$i\lambda v_p = A\int \mathrm{e}^{ik(r+r')}\frac{\cos(n,r)\delta\sigma}{rr'}. \tag{2.B15}$$

## 2.11 APPENDIX C

### 2.11.1 Diffraction by a circular aperture from which the airy disk comes

Our next goal is to understand the origin of the Airy disk in the context of Kirchoff's formulation of Huygens' principle. We again follow Sommerfeld in this discussion. If we take as our starting point Eq. (2.C15) of Appendix B and further assume that dimensions of the aperture are small compared to the distance $r_1$ and $r_2$, then the term $\cos(n,r)/r_1r_2$ varies little within the opening. Hence, that term may be taken outside of the integral. Replacing $r$ and $r'$ with $R$ and $R'$, their respective values at the center of the aperture, Eq. (2.C15) of Appendix B becomes

$$i\lambda v_p = \frac{A}{RR'}\cos(n,r)\int \mathrm{e}^{ik(r+r')}\delta\eta\delta\xi, \tag{2.C1}$$

where $\eta$ and $\xi$ have the same meaning as in Appendix B. Now,

$$r = \sqrt{(x-\xi)^2 + (y-n)^2 + z^2}. \tag{2.C2}$$

Because $R^2 = x^2 + y^2 + z^2$,

$$\begin{aligned} r &= \sqrt{R^2 - 2(x\xi + y\eta) + \xi^2 + \eta^2} \\ &\cong R - \frac{x}{R}\xi - \frac{y}{R}\eta + \frac{\xi^2 + \eta^2}{2R} - \frac{(x\xi + y\eta)^2}{2R^2} \\ &= R - \alpha\xi - \beta\eta + \frac{\xi^2 + \eta^2 - (\alpha\xi + \beta\eta)^2}{2R}, \end{aligned} \tag{2.C3}$$

to second order in $\xi$ and $\eta$.

Where $\alpha$ and $\beta$ are respectively the directional cosines of the diffracted ray O → P′,

$$r' = R' + \alpha_0\xi + \beta_0\eta + \frac{\xi^2 + \eta^2 - (\alpha_0\xi + \beta_0\eta^2)}{2R'}. \tag{2.C4}$$

It then follows that

$$\mathrm{e}^{ik(r+r')} = \mathrm{e}^{ik(R+R')}\mathrm{e}^{-ik\Phi}, \tag{2.C5}$$

where

$$\Phi = (\alpha - \alpha_0)\xi + (\beta - \beta_0)\eta - \left(\frac{1}{R} + \frac{1}{R'}\right)\frac{\xi^2 + \eta^2}{2} + \frac{(\alpha\xi + \beta\eta)^2}{2R} + \frac{(\alpha_0\xi + \beta_0\eta)^2}{2R'}. \tag{2.C6}$$

With these modifications, Eq. (2.C1) becomes

$$i\lambda v_p = \frac{A}{RR'}\cos(n, R)\mathrm{e}^{ik(R+R')}\int \mathrm{e}^{-ik\Phi}\delta\xi\delta\eta. \tag{2.C7}$$

We have, as they say, reached the proverbial end of the tunnel. For the microscope, both $R$ and $R'$ are large in comparison to the dimensions of the aperture. This is the important condition of Fraunhoffer diffraction. In this limit, $\Phi$ becomes linear in $\xi$ and $\eta$:

$$\Phi \cong (\alpha - \alpha_0)\xi + (\beta - \beta_0)\eta. \tag{2.C8}$$

We define

$$a = \alpha - \alpha_0, \tag{2.C9a}$$

and

$$b = \beta - \beta_0, \tag{2.C9b}$$

in which case,

$$\Phi \cong a\xi + b\eta. \tag{2.C10}$$

Considering the problem of a small circular aperture in the microscope, we recognize that in the Fraunhoffer limit $RR'$ and $\cos(n, r)$ vary very little. Because lexp $ik(R + R')$ξl = 1 and $A$ is a constant, we may rewrite Eq. (2.C7) as

$$v_p = C \int e^{ik\Phi} \delta\xi\delta\eta, \tag{2.C11}$$

where $C$ is a constant.

If we replace the coordinates ($\xi$, $\eta$) and the directional cosines ($a$, $b$) with polar coordinates, that is

$$\xi = r \cos \varphi, \tag{2.C12a}$$

$$\eta = r \sin \varphi, \tag{2.C12b}$$

$$a = s \cos \psi, \tag{2.C12c}$$

$$b = s \sin \psi, \tag{2.C12d}$$

where $r$ is the distance from the center of the opening, and $s$ is the sine of the deflection angle between the diffracted ray and the perpendicular incident ray (the undiffracted ray), thus we may rewrite Eq. (2.C11) as

$$v_p = Ck \int_0^{d/2} r\mathrm{d}r \int_{-\pi}^{\pi} e^{-ikrs \cos(\varphi-\psi)} \mathrm{d}\varphi. \tag{2.C13}$$

We recognize the $\varphi$ integral to be the cylindrical Bessel function:

$$J_0(\rho) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} e^{+i\rho \cos\propto} \mathrm{d}\alpha. \tag{2.C14}$$

Thus,

$$\upsilon_p = 2\pi Ck \int_0^{d/2} J_0(krs) r\mathrm{d}r, \tag{2.C15}$$

$$\upsilon_p = \frac{2\pi Ck}{k^2 s^2} \int_0^{ksd/2} J_0(\rho')\rho' d\rho', \tag{2.C16}$$

where $\rho = ikrs$. Then,

$$v_p = \frac{\pi Cd}{s} J_1\left(ks\frac{d}{2}\right). \tag{2.C17}$$

What we actually observe is not $v_p$ but $|v_p|^2$. If we also normalize this by dividing by the intensity squared at $s = 0$, we obtain

$$\frac{v_p^2}{v_{p_0}^2} = \frac{J_1^2(ks(d/2))}{J_1^2(0)}. \tag{2.C18}$$

This is the equation of the Airy disk observed in the microscope for a point source of light.

## Acknowledgments

I thank Ryan and Kathryn Sears, Leya Bergquist, Dylan Bulseco, and Angela Perno for help in the preparation of this chapter, and Robin G. Jordan, professor of physics at Florida Atlantic University, for the quotation by Thomas Young at the beginning of this chapter.

## References

Born, M., & Wolf, E. (1980). *Principles of optics: Electromagnetic theory of propagation interference and diffraction of light* (6th ed.). New York: Pergamon Press.

Goodman, J. W. (1968). *Introduction to Fourier optics*. New York: McGraw-Hill Book Company.

Halliday, D., & Resnick, R. (1970). *Fundamentals of physics*. New York: John Wiley & Sons.

Hui, S. W., & Parsons, D. (1975). Direct observation of domains in wet lipid bilayers. *Science*, *190*, 383.

Inoué, S., & Spring, K. R. (1997). *Video microscopy* (2nd ed.). New York: Plenum Publishing.

Jackson, J. D. (1975). *Classical electrodynamics* (2nd ed.). New York: John Wiley & Sons.

Jenkins, F. A., & White, H. E. (1957). *Fundamentals of optics* (3rd ed.). New York: McGraw-Hill Book Company.

Shamos, M. (1987). *Great experiments in physics*. New York: Dover Publications.

Sommerfeld, A. (Ed.), (1949a). *Optics: Lectures on theoretical physics* (Vol. 4). New York: Academic Press.

Sommerfeld, A. (Ed.), (1949b). *Physics: Partial differential equations* (Vol. 6). New York: Academic Press.

Young, T. (1801). *Course of lectures on natural philosophy and the mechanical arts*. Lecture 39. London: Press of the Royal Institution.