VIII

Elements of the theory of diffraction

8.1 Introduction

In carrying out the transition from the general electromagnetic field to the optical field, which is characterized by very high frequencies (short wavelengths), we found that in certain regions the simple geometrical model of energy propagation was inadequate. In particular, we saw that deviations from this model must be expected in the immediate neighbourhood of the boundaries of shadows and in regions where a large number of rays meet. These deviations are manifested by the appearance of dark and bright bands, the diffraction fringes. Diffraction theory is mainly concerned with the field in these special regions; such regions are of great practical interest as they include the part of the image space in which the optical image is situated (region of focus).

The first reference to diffraction phenomena appears in the work of Leonardo da Vinci (1452–1519). Such phenomena were, however, first accurately described by Grimaldi in a book, published in 1665, two years after his death. The corpuscular theory, which, at the time, was widely believed to describe correctly the propagation of light, could not explain diffraction. Huygens, the first proponent of the wave theory, seems to have been unaware of Grimaldi's discoveries; otherwise he would have undoubtedly quoted them in support of his views. The possibility of explaining diffraction effects on the basis of a wave theory was not noticed until about 1818. In that year there appeared the celebrated memoir of Fresnel (see Historical introduction) in which he showed that diffraction can be explained by the application of Huygens' construction (see §3.3.3) together with the principle of interference. Fresnel's analysis was later put on a sound mathematical basis by Kirchhoff (1882), and the subject has since then been extensively discussed by many writers.*

Diffraction problems are amongst the most difficult ones encountered in optics. Solutions which, in some sense, can be regarded as rigorous are very rare in diffraction theory. The first such solution was given as late as 1896 by A. Sommerfeld when, in an important paper, he discussed the diffraction of a plane wave by a perfectly conducting semi-infinite plane screen. Since then rigorous solutions of a small number of other diffraction problems (mainly two-dimensional) have also been found (see Chapter XI), but, because of mathematical difficulties, approximate methods must be used in most cases of practical interest. Of these the theory of Huygens and Fresnel is by far the

* For a fuller historical account of the development of the subject see C. F. Meyer, *The Diffraction of Light, X-rays, and Material Particles* (Chicago, The University Press, 1934).

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

most powerful and is adequate for the treatment of the majority of problems encountered in instrumental optics. This theory and some of its applications form the main subject matter of this chapter.

8.2 The Huygens–Fresnel principle

According to Huygens' construction (§3.3.3), every point of a wave-front may be considered as a centre of a secondary disturbance which gives rise to spherical wavelets, and the wave-front at any later instant may be regarded as the envelope of these wavelets. Fresnel was able to account for diffraction by supplementing Huygens' construction with the postulate that the secondary wavelets mutually interfere. This combination of Huygens' construction with the principle of interference is called the *Huygens–Fresnel principle*. Before applying it to the study of diffraction effects we shall verify that (with certain simple additional assumptions) the principle correctly describes the propagation of light in free space.

Let S (Fig. 8.1) be the instantaneous position of a spherical monochromatic wave-front of radius r_0 which proceeds from a point source P_0, and let P be a point at which the light disturbance is to be determined. The time periodic factor $\mathrm{e}^{-\mathrm{i}\omega t}$ being omitted, the disturbance at a point Q on the wave-front may be represented by $A\mathrm{e}^{\mathrm{i}kr_0}/r_0$, where A is the amplitude at unit distance from the source. In accordance with the Huygens–Fresnel principle we regard each element of the wave-front as the centre of a secondary disturbance which is propagated in the form of spherical wavelets, and obtain for the contribution $\mathrm{d}U(P)$ due to the element $\mathrm{d}S$ at Q the expression

$$\mathrm{d}U(P) = K(\chi)\frac{A\mathrm{e}^{\mathrm{i}kr_0}}{r_0}\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\,\mathrm{d}S,$$

where $s = QP$ and $K(\chi)$ is an *inclination factor* which describes the variation with

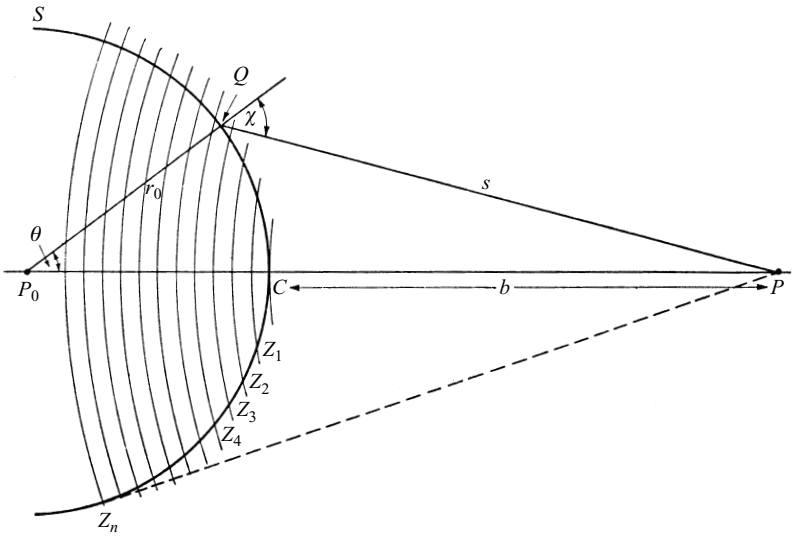


Fig. 8.1 Fresnel's zone construction.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

direction of the amplitude of the secondary waves, χ being the angle (often called the *angle of diffraction*) between the normal at Q and the direction QP. Following Fresnel we assume that K is maximum in the original direction of propagation, i.e. for $\chi = 0$, and that it rapidly decreases with increasing χ, being zero when QP is tangential to the wave-front, i.e. when $\chi = \pi/2$; and finally, that only that part S' of the primary wave contributes to the effect at P, which is not obstructed by obstacles which may be situated between P_0 and P. Hence the total disturbance at P is given by

$$U(P) = \frac{A\mathrm{e}^{\mathrm{i}kr_0}}{r_0} \iint_S \frac{\mathrm{e}^{\mathrm{i}ks}}{s} K(\chi)\mathrm{d}S. \tag{1}$$

To evaluate (1) we shall use the so-called *zone construction* of Fresnel. With centre at P, we construct spheres of radii

$$b, \qquad b + \frac{\lambda}{2}, \qquad b + \frac{2\lambda}{2}, \qquad b + \frac{3\lambda}{2}, \qquad \ldots, \qquad b + \frac{j\lambda}{2}, \qquad \ldots,$$

where $b = CP$, C being the point of intersection of P_0P with the wave-front S (see Fig. 8.1). The spheres divide S into a number of zones $Z_1, Z_2, Z_3, \ldots, Z_j, \ldots$.

We assume that both r_0 and b are large compared to the wavelength; then K may be assumed to have the same value, K_j, for points on one and the same zone. From the figure

$$s^2 = r_0^2 + (r_0 + b)^2 - 2r_0(r_0 + b)\cos\theta,$$

so that

$$s\,\mathrm{d}s = r_0(r_0 + b)\sin\theta\,\mathrm{d}\theta, \tag{2}$$

and therefore

$$\mathrm{d}S = r_0^2 \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi = \frac{r_0}{r_0 + b} s\,\mathrm{d}s\,\mathrm{d}\phi,$$

ϕ being the azimuthal angle. Hence the contribution of the jth zone to $U(P)$ is

$$\begin{aligned} U_j(P) &= 2\pi \frac{A\mathrm{e}^{\mathrm{i}kr_0}}{r_0 + b} K_j \int_{b+(j-1)\lambda/2}^{b+j\lambda/2} \mathrm{e}^{\mathrm{i}ks}\,\mathrm{d}s \\ &= -\frac{2\pi i}{k} K_j \frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0 + b} \mathrm{e}^{\mathrm{i}kj\lambda/2}(1 - \mathrm{e}^{-\mathrm{i}k\lambda/2}). \end{aligned}$$

Since $k\lambda = 2\pi$, the last two factors reduce to

$$\mathrm{e}^{\mathrm{i}kj\lambda/2}(1 - \mathrm{e}^{-\mathrm{i}k\lambda/2}) = \mathrm{e}^{\mathrm{i}\pi j}(1 - \mathrm{e}^{-\mathrm{i}\pi}) = (-1)^j 2,$$

so that

$$U_j(P) = 2i\lambda(-1)^{j+1} K_j \frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0 + b}. \tag{3}$$

We note that the contributions of the successive zones are alternately positive and negative. The total effect at P is obtained by summing all the contributions:

$$U(P) = 2\mathrm{i}\lambda \frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0 + b} \sum_{j=1}^{n} (-1)^{j+1} K_j. \tag{4}$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

The series

$$\Sigma = \sum_{j=1}^{n} (-1)^{j+1} K_j = K_1 - K_2 + K_3 - \cdots + (-1)^{n+1} K_n \tag{5}$$

can now be approximately summed by a method due to Schuster.*

First we write (5) in the form

$$\Sigma = \frac{K_1}{2} + \left(\frac{K_1}{2} - K_2 + \frac{K_3}{2}\right) + \left(\frac{K_3}{2} - K_4 + \frac{K_5}{2}\right) + \cdots, \tag{6}$$

the last term being $\frac{1}{2}K_n$ or $\frac{1}{2}K_{n-1} - K_n$ according to n being odd or even. Let us assume for the moment that the law which specifies the directional variation is such that K_j is *greater* than the arithmetic mean of its two neighbours K_{j-1} and K_{j+1}. Then each of the bracketed terms in (6) is negative and it follows that

$$\left.\begin{aligned} &\Sigma < \frac{K_1}{2} + \frac{K_n}{2} && \text{when } n \text{ is odd} \\ \text{and} \qquad &\\ &\Sigma < \frac{K_1}{2} + \frac{K_{n-1}}{2} - K_n && \text{when } n \text{ is even.} \end{aligned}\right\} \tag{7}$$

We can also write (5) in the form

$$\Sigma = K_1 - \frac{K_2}{2} - \left(\frac{K_2}{2} - K_3 + \frac{K_4}{2}\right) - \left(\frac{K_4}{2} - K_5 + \frac{K_6}{2}\right) - \cdots, \tag{8}$$

the last term now being $-\frac{1}{2}K_{n-1} + K_n$ when n is odd and $-\frac{1}{2}K_n$ when n is even. Hence

$$\left.\begin{aligned} &\Sigma > K_1 - \frac{K_2}{2} - \frac{K_{n-1}}{2} + K_n && (n \text{ odd}) \\ \text{and} \qquad &\\ &\Sigma > K_1 - \frac{K_2}{2} - \frac{K_n}{2} && (n \text{ even}). \end{aligned}\right\} \tag{9}$$

Now each K_j differs only slightly from its neighbouring values K_{j-1} and K_{j+1} so that the right-hand sides of the corresponding relations in (7) and (9) are practically equal, and, therefore, approximately,

$$\left.\begin{aligned} &\Sigma = \frac{K_1}{2} + \frac{K_n}{2} && (n \text{ odd}) \\ \text{and} \qquad &\\ &\Sigma = \frac{K_1}{2} - \frac{K_n}{2} && (n \text{ even}). \end{aligned}\right\} \tag{10}$$

It may easily be verified that (10) remains valid when each K_j is *smaller* than the arithmetic mean of its two neighbours, each of the bracketed terms in (6) and (8) then being positive. Moreover, (10) may be expected to remain valid even when only some of the bracketed terms are negative whilst the others are positive, for the series may then be divided into two parts according to the signs of the bracketed terms and a

* A. Schuster, *Phil. Mag.* (5), **31** (1891), p. 77.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

similar argument may be applied to each part. We may, therefore, conclude that the sum of the series is given by (10) unless the bracketed terms in (6) and (8) change sign so frequently that the error terms add up to an appreciable amount. If we exclude the later case it follows from (10) and (4) that

$$U(P) = \mathrm{i}\lambda(K_1 \pm K_n)\frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0+b}, \tag{11}$$

the upper or lower sign being taken according as n is odd or even. Using (3), (11) may be also written in the form

$$U(P) = \tfrac{1}{2}[U_1(P) + U_n(P)]. \tag{12}$$

For the last zone (Z_n) that can be seen from P, QP is a tangent to the wave, i.e. $\chi = \pi/2$, and for this value of χ, as already mentioned, K was assumed to be zero. Hence $K_n = 0$ and (11) reduces to

$$U(P) = \mathrm{i}\lambda K_1 \frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0+b} = \tfrac{1}{2}U_1(P), \tag{13}$$

showing that *the total disturbance at P is equal to half of the disturbance due to the first zone*.

Eq. (13) is in agreement with the expression for the effect of the spherical wave if

$$\mathrm{i}\lambda K_1 = 1,$$

i.e. if

$$K_1 = -\frac{\mathrm{i}}{\lambda} = \frac{\mathrm{e}^{-\mathrm{i}\pi/2}}{\lambda}. \tag{14}$$

The factor $\mathrm{e}^{-\mathrm{i}\pi/2}$ may be accounted for by assuming that the secondary waves oscillate a quarter of a period out of phase with the primary wave; the other factor can be explained by assuming that the amplitudes of the secondary vibrations are in the ratio of $1{:}\lambda$ to those of the primary vibrations. We can therefore conclude that, with these assumptions about the amplitude and phase of the secondary waves, the Huygens–Fresnel principle leads to the correct expression for the propagation of a spherical wave in free space. The additional assumptions must, however, be regarded as purely a convenient way of interpreting the mathematical expressions and as being devoid of any physical significance; the real justification of the factor (14) will become evident later (§8.3).

Still following Fresnel, let us consider the effect at P when some of the zones are obstructed by a plane screen with a circular opening, perpendicular to P_0P and with its centre on this line. The total disturbance at P must now be regarded as due to wavelets from only those zones that are not obstructed by the screen. When the screen covers all but half of the first zone, (3) gives, on setting $j = 1$, and multiplying by $\frac{1}{2}$,

$$U(P) = \mathrm{i}\lambda K_1 \frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0+b} = \frac{A\mathrm{e}^{\mathrm{i}k(r_0+b)}}{r_0+b}; \tag{15}$$

hence the disturbance at P is now the same as would be obtained if no screen were present. When all the zones are covered except the first, (3) gives

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$U(P) = 2i\lambda K_1 \frac{Ae^{ik(r_0+b)}}{r_0+b} = 2\frac{Ae^{ik(r_0+b)}}{r_0+b}, \tag{16}$$

so that the intensity $I(P) = |U(P)|^2$ is four times larger than if the screen were absent. When the opening is increased still further the intensity will decrease, since the first two terms in (4) have different signs. Moreover, since K_1 and K_2 are nearly equal, it follows that there will be almost complete darkness at P when the opening is approximately equal to the first two zones. Thus, when the size of the opening is varied, there is a periodic fluctuation in the intensity at P. Similar results are obtained when the size of the opening and the source are fixed, but the position of the point P of observation is varied along the axis; for then, as P gradually approaches the screen, an increasingly larger number of zones is required to fill the opening completely.

The total number of zones has a simple form when the radius of the opening is much smaller than the perpendicular distance from the screen to the observation point P. Let us consider an opening of radius a, centered on the point C, perpendicular to P_0P (see Fig. 8.1). The outermost (Nth say) Fresnel zone contained within the aperture will have a radius roughly equal to the distance from the point P to the edge of the aperture, i.e.

$$b + \frac{N\lambda}{2} = \sqrt{b^2 + a^2}. \tag{17}$$

When $a \ll b$ one has $\sqrt{b^2+a^2} \approx b + (b/2)(a/b)^2$. On substituting from this expression into (17) it follows that

$$N = \frac{a^2}{b\lambda}. \tag{18}$$

The number N, which represents the number of Fresnel zones in the opening, is known as the *Fresnel number* of the system. This parameter also plays an important role in the theory of laser resonators.*

All these results were found to be in good agreement with experiment. One prediction of Fresnel's theory made a strong impression on his contemporaries, and was, in fact, one of the decisive factors which temporarily ended the long battle between the corpuscular and the wave theory of light in favour of the latter. It concerns the effect which arises when the first zone is obstructed by a small circular disc placed at right angles to P_0P. According to (5) the complex amplitude at P is then given by

$$U(P) = 2i\lambda \frac{Ae^{ik(r_0+b)}}{r_0+b}[-K_2 + K_3 - K_4 + \cdots], \tag{19}$$

and, by a similar argument as before, the sum of the series in the brackets is $-K_2/2$. Since K_2 is assumed to differ only slightly from $K_1 = 1/i\lambda$, it follows that there is light in the geometrical shadow of the disc, and, moreover, that the intensity there is the same as if no disc were present.†

* See, for example, A. E. Siegman, *Lasers* (Mill Valley, University Science Books, 1986). pp. 769–70.

† That a bright spot should appear at the centre of the shadow of a small disc was deduced from Fresnel's theory by S. D. Poisson in 1818. Poisson, who was a member of the committee of the French Academy which reviewed Fresnel's prize memoir, appears to have considered this conclusion contrary to experiment and so refuting Fresnel's theory. However, Arago, another member of the committee, performed the experiment and found that the surprising prediction was correct. A similar observation had been made a century earlier by Maraldi but had been forgotten.

8.3 Kirchhoff's diffraction theory

8.3.1 The integral theorem of Kirchhoff

The basic idea of the Huygens–Fresnel theory is that the light disturbance at a point P arises from the superposition of secondary waves that proceed from a surface situated between this point and the light source. This idea was put on a sounder mathematical basis by Kirchhoff*, who showed that the Huygens–Fresnel principle may be regarded as an approximate form of a certain integral theorem† which expresses the solution of the homogeneous wave equation, at an arbitrary point in the field, in terms of the values of the solution and its first derivatives at all points on an arbitrary closed surface surrounding P.

We consider first a strictly monochromatic scalar wave

$$V(x, y, z, t) = U(x, y, z)\mathrm{e}^{-\mathrm{i}\omega t}. \tag{1}$$

In vacuum the space-dependent part then satisfies the time-independent wave equation

$$(\nabla^2 + k^2)U = 0, \tag{2}$$

where $k = \omega/c$. Eq. (2) is also known as the Helmholtz equation.

Let v be a volume bounded by a closed surface S, and let P be any point within it; we assume that U possesses continuous first- and second-order partial derivatives within and on this surface. If U' is any other function which satisfies the same continuity requirements as U, we have by Green's theorem

$$\iiint_v (U\nabla^2 U' - U'\nabla^2 U)\mathrm{d}v = -\iint_S \left(U\frac{\partial U'}{\partial n} - U'\frac{\partial U}{\partial n}\right)\mathrm{d}S, \tag{3}$$

where $\partial/\partial n$ denotes differentiation along the *inward*‡ normal to S. In particular, if U' also satisfies the time-independent wave equation, i.e. if

$$(\nabla^2 + k^2)U' = 0, \tag{4}$$

then it follows at once from (2) and (4) that the integrand on the left of (3) vanishes at every point of v, and consequently

$$\iint_S \left(U\frac{\partial U'}{\partial n} - U'\frac{\partial U}{\partial n}\right)\mathrm{d}S = 0. \tag{5}$$

Suppose we take $U'(x, y, z) = \mathrm{e}^{\mathrm{i}ks}/s$, where s denotes the distance from P to the point (x, y, z). This function has a singularity for $s = 0$, and since U' was assumed to be continuous and differentiable, P must be excluded from the domain of integration. We shall therefore surround P by a small sphere of radius ε and extend the integration

* G. Kirchhoff, *Berl. Ber.* (1882), 641; *Ann. d. Physik.* (2), **18** (1883), 663; *Ges. Abh. Nachtr.*, 22.
Kirchhoff's theory applies to the diffraction of scalar waves. As will be shown in §8.4 a scalar theory is usually quite adequate for the treatment of the majority of problems of instrumental optics.
Vectorial generalizations of the Huygens–Fresnel principle have been proposed by many authors. The first satisfactory generalization is due to F. Kottler, *Ann. d. Physik*, **71** (1923), 457; **72** (1923), 320. (See B. B. Baker and E. T. Copson, *The Mathematical Theory of Huygens' Principle* (Oxford, Clarendon Press, 2nd edition, 1950), p. 114.)

† For monochromatic waves this theorem was derived earlier in acoustics by H. von Helmholtz, *J. f. Math.*, **57** (1859), 7.

‡ Green's theorem is usually expressed in terms of the outward normal, but the inward normal is more convenient in the present application.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

throughout the volume between S and the surface S' of this sphere (Fig. 8.2). In place of (5), we then have

$$\left\{\iint_S + \iint_{S'}\right\}\left[U\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial n}\right]\mathrm{d}S = 0,$$

so that

$$\iint_S\left[U\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial n}\right]\mathrm{d}S = -\iint_{S'}\left[U\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\left(\mathrm{i}k - \frac{1}{s}\right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial n}\right]\mathrm{d}S'$$

$$= -\iint_\Omega\left[U\frac{\mathrm{e}^{\mathrm{i}k\varepsilon}}{\varepsilon}\left(\mathrm{i}k - \frac{1}{\varepsilon}\right) - \frac{\mathrm{e}^{\mathrm{i}k\varepsilon}}{\varepsilon}\frac{\partial U}{\partial s}\right]\varepsilon^2\,\mathrm{d}\Omega, \quad (6)$$

where $\mathrm{d}\Omega$ denotes an element of the solid angle. Since the integral over S is independent of ε, we may replace the integral on the right-hand side by its limiting value as $\varepsilon \to 0$; the first and third terms in this integral give no contribution in the limit, and the total contribution of the second term is $4\pi U(P)$. Hence

$$U(P) = \frac{1}{4\pi}\iint_S\left[U\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial n}\right]\mathrm{d}S. \quad (7)$$

This is one form of the *integral theorem of Helmholtz and Kirchhoff*.*

We note, that as $k \to 0$, the time-independent wave equation (2) reduces to Laplace's equation $\nabla^2 U = 0$, and (7) then goes over into the well-known formula of potential theory

$$U(P) = \frac{1}{4\pi}\iint_S\left[U\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{s}\frac{\partial U}{\partial n}\right]\mathrm{d}S. \quad (8)$$

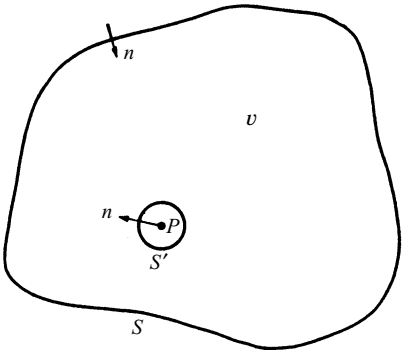


Fig. 8.2 Derivation of the Helmholtz–Kirchhoff integral theorem: region of integration.

* This theorem expresses $U(P)$ in terms of the values of both U and $\partial U/\partial n$ on S. It may, however, be shown from the theory of Green's functions that the values of either U or $\partial U/\partial n$ on S are sufficient to specify U at every point P within S. (See, for example, F. Pockels: *Über die Partielle Differentialgleichung* $(\nabla^2 + k^2)U = 0$ (Leipzig, Teubner, 1891).) However, only in the simplest cases, e.g. when S is a plane, is it possible to determine the appropriate Green's function (see A. Sommerfeld, *Optics* (New York, Academic Press, 1954), p. 199). The resulting expression for $U(P)$ is then known as the Rayleigh diffraction integral and will be derived in §8.10.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

If P lies outside the surface S, but U is still assumed to be continuous and differentiable up to the second order within S, and if as before we take $U' = \mathrm{e}^{\mathrm{i}ks}/s$, (3) remains valid throughout the whole volume within S. According to (5) the surface integral then has the value zero.

There is a complementary form of the Helmholtz–Kirchhoff theorem for the case when U is continuous and differentiable up to the second order *outside* and on a closed surface S (sources inside). In this case, however, as in other problems of propagation in an infinite medium, the boundary values on S are no longer sufficient to specify the solution uniquely and additional assumptions must be made about the behaviour of the solution as $s \to \infty$. For a discussion of this case we must, however, refer elsewhere.*

So far we have considered strictly monochromatic waves. We now derive the general form of Kirchhoff's theorem which applies to waves that are not necessarily monochromatic.

Let $V(x, y, z, t)$ be a solution of the wave equation

$$\nabla^2 V = \frac{1}{c^2}\frac{\partial^2 V}{\partial t^2}, \tag{9}$$

and assume that V can be represented in the form of a Fourier integral

$$V(x, y, z, t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega(x, y, z)\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega. \tag{10}$$

Then, by the Fourier inversion formula

$$U_\omega(x, y, z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} V(x, y, z, t)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}t. \tag{11}$$

Since $V(x, y, z, t)$ is assumed to satisfy the wave equation (9), $U_\omega(x, y, z)$ will satisfy the time-independent wave equation (2). If, moreover, V obeys the appropriate regularity conditions within and on a closed surface S, we may apply the Kirchhoff formula separately to each Fourier component $U_\omega(x, y, z) = U_\omega(P)$:

$$U_\omega(P) = \frac{1}{4\pi}\iint_S \left\{U_\omega \frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\frac{\partial U_\omega}{\partial n}\right\}\mathrm{d}S. \tag{12}$$

When we change the order of integration and set $k = \omega/c$, (10) becomes,

$$\begin{aligned} V(P, t) &= \frac{1}{4\pi}\iint_S \mathrm{d}S \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left\{U_\omega \frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{-\mathrm{i}\omega(t-s/c)}}{s}\right) - \frac{\mathrm{e}^{-\mathrm{i}\omega(t-s/c)}}{s}\frac{\partial U_\omega}{\partial n}\right\}\mathrm{d}\omega \\ &= \frac{1}{4\pi}\iint_S \mathrm{d}S \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left\{U_\omega\left\{\frac{\partial}{\partial n}\left(\frac{1}{s}\right) + \frac{\mathrm{i}\omega}{sc}\frac{\partial s}{\partial n}\right\}\mathrm{e}^{-\mathrm{i}\omega(t-s/c)}\right. \\ &\qquad \left. - \frac{\mathrm{e}^{-\mathrm{i}\omega(t-s/c)}}{s}\frac{\partial U_\omega}{\partial n}\right\}\mathrm{d}\omega \end{aligned}$$

or using (10),

$$V(P, t) = \frac{1}{4\pi}\iint_S \left\{[V]\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\frac{\partial s}{\partial n}\left[\frac{\partial V}{\partial t}\right] - \frac{1}{s}\left[\frac{\partial V}{\partial n}\right]\right\}\mathrm{d}S. \tag{13}$$

* See for example B. B. Baker and E. T. Copson, *loc. cit.*, p. 26.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

The square brackets denote 'retarded values', i.e. values of the functions taken at the time $t - s/c$. The formula (13) is the general form of *Kirchhoff's theorem*.

It can also be seen by analogy with the previous case, that the value of the integral in (13) is zero when P is outside S.

The last term in (13) represents the contribution of a distribution of sources of strength $-(1/4\pi)(\partial V/\partial n)$ per unit area, whilst the first two terms may be shown to represent a contribution of doublets of strength $V/4\pi$ per unit area, directed normally to the surface. Naturally these sources and doublets are fictitious, there being no deep physical significance behind such an interpretation.

8.3.2 Kirchhoff's diffraction theory

Whilst the integral theorem of Kirchhoff embodies the basic idea of the Huygens–Fresnel principle, the laws governing the contributions from different elements of the surface are more complicated than Fresnel assumed. Kirchhoff showed, however, that in many cases the theorem may be reduced to an approximate but much simpler form, which is essentially equivalent to the formulation of Fresnel, but which in addition gives an explicit formula for the inclination factor that remained undetermined in Fresnel's theory.

Consider a monochromatic wave, from a point source P_0, propagated through an opening in a plane opaque screen, and let P as before be the point at which the light disturbance is to be determined. We assume that the linear dimensions of the opening, although large compared to the wavelength, are small compared to the distances of both P_0 and P from the screen.

To find the disturbance at P we take Kirchhoff's integral over a surface S formed by (see Fig. 8.3(a)): (1) the opening $\mathcal{A}$, (2) a portion $\mathcal{B}$ of the nonilluminated side of the screen, and (3) a portion $\mathcal{C}$ of a large sphere of radius R, centred at P which, together with $\mathcal{A}$ and $\mathcal{B}$, forms a closed surface.

Kirchhoff's theorem, expressed by (7), then gives

$$U(P) = \frac{1}{4\pi}\left[\iint_{\mathcal{A}} + \iint_{\mathcal{B}} + \iint_{\mathcal{C}}\right]\left\{U\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right) - \left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right)\frac{\partial U}{\partial n}\right\}\mathrm{d}S, \qquad (14)$$

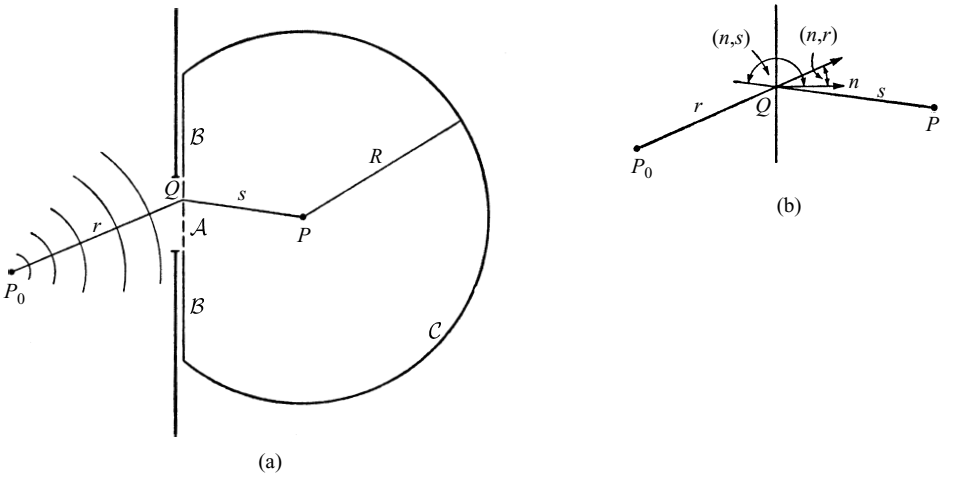


Fig. 8.3 Illustrating the derivation of the Fresnel–Kirchhoff diffraction formula.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

where, as before, s is the distance of the element $\mathrm{d}S$ from P and $\partial/\partial n$ denotes differentiation along the inward normal to the surface of integration.

The difficulty is encountered that the values of U and $\partial U/\partial n$ on $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ which should be substituted into (14) are never known exactly. However, it is reasonable to suppose that everywhere on $\mathcal{A}$, except in the immediate neighbourhood of the rim of the opening, U and $\partial U/\partial n$ will not appreciably differ from the values obtained in the absence of the screen, and that on $\mathcal{B}$ these quantities will be approximately zero. Kirchhoff accordingly set

$$\left.\begin{array}{lll} \text{on } \mathcal{A}: & U = U^{(i)}, & \dfrac{\partial U}{\partial n} = \dfrac{\partial U^{(i)}}{\partial n}, \\ \text{on } \mathcal{B}: & U = 0 & \dfrac{\partial U}{\partial n} = 0, \end{array}\right\} \tag{15}$$

where

$$U^{(i)} = \frac{A\mathrm{e}^{\mathrm{i}kr}}{r}, \qquad \frac{\partial U^{(i)}}{\partial n} = \frac{A\mathrm{e}^{\mathrm{i}kr}}{r}\left[\mathrm{i}k - \frac{1}{r}\right]\cos(n,\, r) \tag{16}$$

are the values relating to the incident field (see Fig. 8.3(b)) and A is a constant. The approximations (15) are called *Kirchhoff's boundary conditions* and are the basis of *Kirchhoff's diffraction theory.*

It remains to consider the contribution from the spherical portion $\mathcal{C}$. Now it is evident that by taking the radius R sufficiently large, the values of U and $\partial U/\partial n$ on $\mathcal{C}$ may be made arbitrarily small, which suggests that the contribution from $\mathcal{C}$ may be neglected. However, by letting R increase indefinitely, the area of $\mathcal{C}$ also increases beyond all limits, so that the condition $U \to 0$ and $\partial U/\partial n \to 0$ as $R \to \infty$ is not sufficient to make the integral vanish. A more precise assumption about the behaviour of the wave function at a large distance from the screen must therefore be made, a point which we have already touched upon on p. 419 in connection with the uniqueness of solutions in problems involving an infinite medium. For our purposes it is sufficient to make the physically obvious assumption that the radiation field does not exist at all times but that it is produced by a source that begins to radiate at some particular instant of time $t = t_0$.* (This, of course, implies, that we now depart from strict monochromaticity, since a perfectly monochromatic field would exist for all times.) Then at any time $t > t_0$, the field fills a region of space the outer boundary of which is at distance not greater than $c(t - t_0)$ from P_0, c being the velocity of light. Hence if the radius R is chosen so large that at the time when the disturbance at P is considered no contributions from $\mathcal{C}$ could have reached P because at the appropriate earlier time the field has not reached these distant regions, the integral over $\mathcal{C}$ will vanish. Thus finally, on substituting into (14), and neglecting in the normal derivatives the terms $1/r$ and $1/s$ in comparison to k, we obtain

$$U(P) = -\frac{\mathrm{i}A}{2\lambda}\iint_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{rs}[\cos(n,\, r) - \cos(n,\, s)]\mathrm{d}S, \tag{17}$$

which is known as the *Fresnel–Kirchhoff diffraction formula*.

* This assumption is not essential but shortens the discussion. For a more formal argument see M. Born, *Optik* (Berlin, Springer, 1933, reprinted 1965), p. 149 or B. B. Baker and E. T. Copson, *loc. cit.*, p. 25.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

It is evident that in place of $\mathcal{A}$ any other open surface, the rim of which coincides with the edge of the aperture, could have been chosen. In particular we may choose instead of $\mathcal{A}$ a portion W of an incident wave front which approximately fills the aperture, together with a portion $\mathcal{C}$ of a cone with vertex at P_0 and with generators through the rim of the aperture (Fig. 8.4). If the radius of curvature of the wave is sufficiently large, the contribution from $\mathcal{C}$ may obviously be neglected. Also, on W, $\cos(n, r_0) = 1$. If further we set $\chi = \pi - (r_0, s)$ we obtain, in place of (17),

$$U(P) = -\frac{\mathrm{i}}{2\lambda}\frac{A\mathrm{e}^{\mathrm{i}kr_0}}{r_0}\iint_W \frac{\mathrm{e}^{\mathrm{i}ks}}{s}(1+\cos\chi)\mathrm{d}S, \tag{18}$$

where r_0 is the radius of the wave-front W. This result is in agreement with Fresnel's formulation of Huygens' principle if, as the contribution from the element $\mathrm{d}W$ of the wave-front we take

$$-\frac{\mathrm{i}}{2\lambda}\frac{A\mathrm{e}^{\mathrm{i}kr_0}}{r_0}\frac{\mathrm{e}^{\mathrm{i}ks}}{s}(1+\cos\chi)\mathrm{d}S. \tag{19}$$

Comparison of (18) with §8.2 (1) gives for the inclination factor of Fresnel's theory the expression*

$$K(\chi) = -\frac{\mathrm{i}}{2\lambda}(1+\cos\chi). \tag{20}$$

For the central zone $\chi = 0$, and (20) gives $K_1 = K(0) = -\mathrm{i}/\lambda$ in agreement with §8.2. (14). It is, however, not true, as Fresnel assumed, that $K(\pi/2) = 0$.

Returning now to the Fresnel–Kirchhoff diffraction formula (17), we note that it is symmetrical with respect to the source and the point of observation. This implies that *a point source at P_0 will produce at P the same effect as a point source of equal intensity placed at P will produce at P_0*. This result is sometimes referred to as the *reciprocity theorem* (or the *reversion theorem*) *of Helmholtz*.

So far we have assumed that the light on its passage from the source to P does not encounter any other surface than the diffracting screen; the incident waves are then spherical. The analysis can be easily extended to cover more complicated cases, where

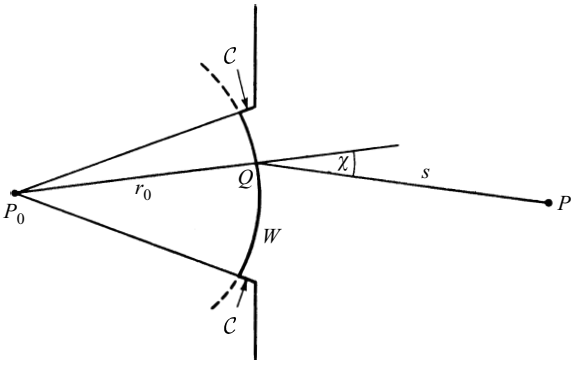


Fig. 8.4 Illustrating the diffraction formula (18).

* Expression (20) for the inclination factor was first derived by G. G. Stokes, *Trans. Camb. Phil. Soc.*, **9** (1849), 1; reprinted in his *Math. and Phys. Papers*, Vol. 2 (Cambridge, Cambridge University Press, 1883), p. 243.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

the waves are no longer of such simple form. It is again found that, provided the radii of curvature at each point of the wave-front are large compared to the wavelength, and provided that the angles involved are sufficiently small, the results of Kirchhoff's theory are substantially equivalent to predictions based on the Huygens–Fresnel principle.

From the preceding discussion we can also immediately draw a conclusion which concerns the distribution of light diffracted by complementary screens, i.e. by screens which are such that the openings in one correspond exactly to the opaque portions of the others and vice versa. Let $U_1(P)$ and $U_2(P)$ denote respectively the values of the complex displacement when the first or the second screen alone is placed between the source and the point P of observation, and let $U(P)$ be the value when no screen is present. Then, since U_1 and U_2 can be expressed as integrals over the openings, and since the openings in the two screens just add up to fill the whole plane,

$$U_1 + U_2 = U. \tag{21}$$

This result is known as *Babinet's principle*.*

From Babinet's principle two conclusions follow at once: If $U_1 = 0$, then $U_2 = U$; hence at points at which the intensity is zero in the presence of one of the screens, the intensity in the presence of the other is the same as if no screen was present. Further if $U = 0$, then $U_1 = -U_2$; this implies that, at points where U is zero, the phases of U_1 and U_2 differ by π and the intensities $I_1 = |U_1|^2$, $I_2 = |U_2|^2$ are equal. If, for example, a point source is imaged by an error-free lens, the light distribution U in the image plane will be zero except in the immediate neighbourhood of the image O of the source. If then complementary screens are placed between the object and the image one has $I_1 = I_2$ except in the neighbourhood of O.

The consequences of the basic approximation (15) of Kirchhoff's theory have been subject to many critical discussions, which showed, for example, that Kirchhoff's solution does not reproduce the assumed values in the plane of the aperture†. However, more recently it was shown by Wolf and Marchand‡ that Kirchhoff's theory may be interpreted in a mathematically consistent way, as providing an exact solution to a somewhat different boundary value problem than that specified by (15) and (16). It turns out that Kirchhoff's theory is entirely adequate for the treatment of the majority of problems encountered in instrumental optics. This is mainly due to the smallness of the optical wavelengths in comparison with the dimensions of the diffracting obstacles.§ In other problems, such as those relating to the behaviour of the field in the

* A. Babinet, *Compt. Rend.*, **4** (1837), 638. An analogous theorem of this type, which involves the electromagnetic field vectors rather than the single scalar U and which may be considered as rigorous formulation of Babinet's principle, is given in §11.3.

† H. Poincaré, *Théorie mathématique de la lumière* (Paris, George Carré, II (1892)), pp. 187–8. See also B. B. Baker and E. T. Copson, *loc. cit.*, pp. 71–72 and G. Toraldo di Francia, *Atti Fond. Giorgio Ronchi*, **XI** (1956), §6.

‡ E. Wolf and E. W. Marchand, *J. Opt. Soc. Amer.*, **56** (1966), 1712. Also, it has been shown by F. Kottler, *Ann. d. Physik*, **70** (1923), 405 that Kirchhoff's theory may be regarded as providing a rigorous solution to a certain saltus problem (a problem with prescribed discontinuities rather than prescribed boundary values). This interpretation is of particular interest in connection with the problem of diffraction at a black (completely absorbing) screen. (See also F. Kottler, *Progress in Optics*, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1964), p. 281 and B. B. Baker and E. T. Copson, *loc. cit.*, p. 98.)

An article by C. J. Bouwkamp, *Rep. Progr. Phys.* (London, Physical Society), **17** (1954), 35, contains references to numerous papers concerned with various modifications of Kirchhoff's theory.

§ See S. Silver, *J. Opt. Soc. Amer.*, **52** (1962), 131.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

immediate neighbourhood of screens and obstacles, more refined methods have to be used; they must then be considered as boundary-value problems of electromagnetic theory, with the sources as appropriate singularities of the wave functions. Only in a very limited number of cases have such solutions been found; some of them will be discussed in Chapter XI.

8.3.3 Fraunhofer and Fresnel diffraction

We now examine more closely the Fresnel–Kirchhoff diffraction integral (17),

$$U(P) = -\frac{A\mathrm{i}}{2\lambda}\iint_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{rs}[\cos(n, r) - \cos(n, s)]\mathrm{d}S. \tag{22}$$

As the element $\mathrm{d}S$ explores the domain of integration, $r + s$ will in general change by very many wavelengths, so that the factor $\mathrm{e}^{\mathrm{i}k(r+s)}$ will oscillate rapidly. On the other hand, if the distances of the points P_0 and P from the screen are large compared to the linear dimensions of the aperture, the factor $[\cos(n, r) - \cos(n, s)]$ will not vary appreciably over the aperture. Further, we assume that if O is any point in the aperture, the angles which the lines P_0O and OP make with P_0P are not too large. We may then replace this factor by $2\cos\delta$, where δ is the angle between the line P_0P and the normal to the screen. Finally the factor $1/rs$ may be replaced by $1/r's'$, where r' and s' are the distance of P_0 and P from the origin and (22) then reduces to

$$U(P) \sim -\frac{A\mathrm{i}}{\lambda}\frac{\cos\delta}{r's'}\iint_{\mathcal{A}} \mathrm{e}^{\mathrm{i}k(r+s)}\,\mathrm{d}S. \tag{23}$$

We take a Cartesian reference system with origin in the aperture and with the x- and y-axes in the plane of the aperture and choose the positive z direction to point into the half-space that contains the point P of observation (Fig. 8.5).

If (x_0, y_0, z_0) and (x, y, z) are the coordinates of P_0 and of P respectively, and (ξ, η) the coordinates of a point Q in the aperture, we have

$$\left.\begin{aligned} r^2 &= (x_0 - \xi)^2 + (y_0 - \eta)^2 + {z_0}^2, \\ s^2 &= (x - \xi)^2 + (y - \eta)^2 + z^2, \end{aligned}\right\} \tag{24}$$

$$\left.\begin{aligned} r'^2 &= x_0^2 + y_0^2 + z_0^2, \\ s'^2 &= x^2 + y^2 + z^2. \end{aligned}\right\} \tag{25}$$

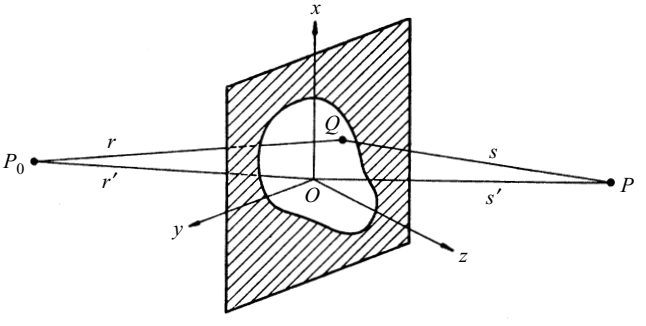


Fig. 8.5 Diffraction at an aperture in a plane screen.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Hence

$$\left.\begin{aligned} r^2 &= r'^2 - 2(x_0\xi + y_0\eta) + \xi^2 + \eta^2, \\ s^2 &= s'^2 - 2(x\xi + y\eta) + \xi^2 + \eta^2. \end{aligned}\right\} \tag{26}$$

Since we assumed that the linear dimensions of the aperture are small compared to both r' and s' we may expand r and s as power series in ξ/r', η/r', ξ/s' and η/s'. We then obtain

$$\left.\begin{aligned} r &\sim r' - \frac{x_0\xi + y_0\eta}{r'} + \frac{\xi^2 + \eta^2}{2r'} - \frac{(x_0\xi + y_0\eta)^2}{2r'^3} - \cdots, \\ s &\sim s' - \frac{x\xi + y\eta}{s'} + \frac{\xi^2 + \eta^2}{2s'} - \frac{(x\xi + y\eta)^2}{2s'^3} - \cdots. \end{aligned}\right\} \tag{27}$$

Substitution from (27) into (23) gives

$$U(P) = -\frac{\mathrm{i}\cos\delta}{\lambda}\frac{A\mathrm{e}^{\mathrm{i}k(r'+s')}}{r's'}\iint_{\mathcal{A}} \mathrm{e}^{\mathrm{i}kf(\xi,\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta, \tag{28}$$

where

$$\begin{aligned} f(\xi, \eta) = &-\frac{x_0\xi + y_0\eta}{r'} - \frac{x\xi + y\eta}{s'} + \frac{\xi^2 + \eta^2}{2r'} + \frac{\xi^2 + \eta^2}{2s'} \\ &- \frac{(x_0\xi + y_0\eta)^2}{2r'^3} - \frac{(x\xi + y\eta)^2}{2s'^3} \cdots. \end{aligned} \tag{29}$$

If we denote by $(l_0,\ m_0)$ and $(l,\ m)$ the first two direction cosines

$$\left.\begin{aligned} l_0 &= -\frac{x_0}{r'}, \qquad l = \frac{x}{s'}, \\ m_0 &= -\frac{y_0}{r'}, \qquad m = \frac{y}{s'}, \end{aligned}\right\} \tag{30}$$

(29) may be written in the form

$$\begin{aligned} f(\xi, \eta) = &(l_0 - l)\xi + (m_0 - m)\eta \\ &+ \tfrac{1}{2}\left[\left(\frac{1}{r'} + \frac{1}{s'}\right)(\xi^2 + \eta^2) - \frac{(l_0\xi + m_0\eta)^2}{r'} - \frac{(l\xi + m\eta)^2}{s'}\right] \cdots. \end{aligned} \tag{31}$$

We have reduced the problem of determining the light disturbance at P to the evaluation of the integral (28). Naturally the evaluation is simpler to carry out when the quadratic and higher-order terms in ξ and η may be neglected in f. In this case one speaks of *Fraunhofer diffraction*; when the quadratic terms cannot be neglected, one speaks of *Fresnel diffraction*. Fortunately the simpler case of Fraunhofer diffraction is of much greater importance in optics.

Strictly speaking, the second and higher-order terms disappear only in the limiting case $r' \to \infty$, $s' \to \infty$, i.e. when both the source and the point of observation are at infinity (the factor A outside the integral must then be assumed to tend to infinity like $r's'$). It is, however, evident that the second-order terms do not appreciably contribute to the integral if

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$\tfrac{1}{2}k\left|\left(\frac{1}{r'}+\frac{1}{s'}\right)(\xi^2+\eta^2)-\frac{(l_0\xi+m_0\eta)^2}{r'}-\frac{(l\xi+m\eta)^2}{s'}\right|\ll 2\pi. \tag{32}$$

We can immediately recognize certain conditions under which (32) will be satisfied. If we make use of inequalities of the form $(l_0\xi+m_0\eta)^2\leqslant(l_0{}^2+m_0{}^2)(\xi^2+\eta^2)$ and remember that $l_0{}^2$, $m_0{}^2$, l^2 and m^2 cannot exceed unity, we find that (32) will be satisfied if

$$|r'|\gg\frac{(\xi^2+\eta^2)_{\max}}{\lambda}\quad\text{and}\quad|s'|\gg\frac{(\xi^2+\eta^2)_{\max}}{\lambda}, \tag{33}$$

or if

$$\frac{1}{r'}+\frac{1}{s'}=0\quad\text{and}\quad l_0^2,\ m_0^2,\ l^2,\ m^2\ll\frac{|r'|\lambda}{(\xi^2+\eta^2)_{\max}}. \tag{34}$$

Conditions (33) give an estimate of the distances r' and s' for which the Fraunhofer representation may be used. Conditions (34) imply that Fraunhofer diffraction also occurs when the point of observation is situated in a plane parallel to that of the aperture, provided that both the point of observation and the source are sufficiently close to the z-axis. Here two cases may be distinguished: When r' is negative, the wave-fronts incident upon the aperture are concave to the direction of the propagation, i.e. P_0 is a centre of convergence and not of divergence of the incident wave. This case is of great practical importance, as it arises in the image space of a well-corrected centred system that images a point source which is not far from the axis. A Fraunhofer pattern is then formed in the Gaussian image plane and may be considered as arising from the diffraction of the image-forming wave on the exit pupil. When r' is positive, the wave-fronts are convex to the direction of propagation. The diffraction phenomena are virtual, being apparently formed on a screen through the source P_0. This case arises, for example, when an aperture is held in front of the eye, or the object glass of a telescope adjusted for distant vision of the light source.

To understand in more physical terms why Fraunhofer phenomena are observed in the focal plane of a well-corrected lens, let us compare first the two situations illustrated in Fig. 8.6. In Fig. 8.6(a) a pencil of rays from an infinitely distant point is incident upon the aperture in the direction specified by the direction cosines l_0, m_0, n_0. The effect observed at a very distant point P in the direction l, m, n may be regarded as arising from the superposition of plane waves originating at each point of the aperture and propagated in this direction. These waves (which have no existence in the domain of geometrical optics) may be called *diffracted waves* and the corresponding wave-normals the *diffracted rays*.

If now a well-corrected lens is placed behind the screen [Fig. 8.6(b)] all the light diffracted in the (l, m, n) direction will be brought to a focus P' in the focal plane of the lens. Since the optical path from a wave-front of the diffracted pencil to P' is the same for all the rays, one obtains substantially the same interference effects as in the first case; it being assumed, of course, that the lens is so large that it introduces no additional diffraction. More generally still, the restriction that the wave incident upon the aperture is plane may also be removed, provided that the path lengths from the source to P' are substantially the same for all the rays.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

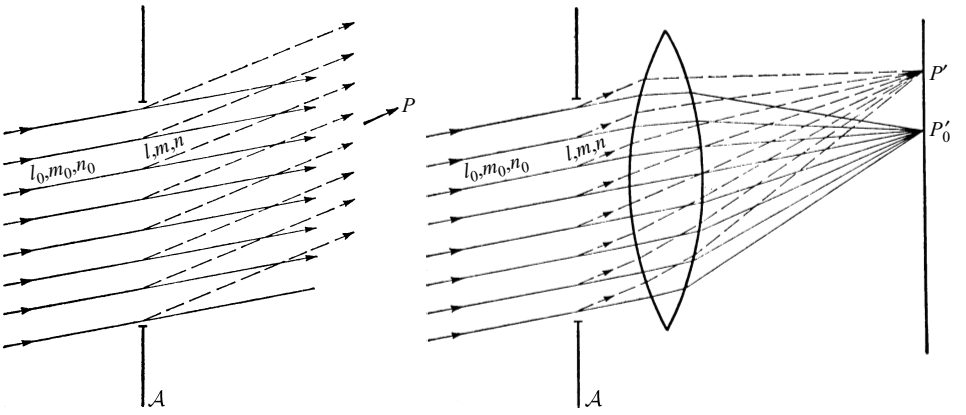


Fig. 8.6 Comparison of two cases of Fraunhofer diffraction.

In the case of Fraunhofer diffraction, the four quantities l_0, m_0, l, m enter (31) only in the combinations

$$p = l - l_0, \qquad q = m - m_0. \tag{35}$$

Hence, within the range of validity of the above approximation, the effect is unchanged when the aperture is displaced in its own plane.

We shall write the integral governing Fraunhofer diffraction in the form

$$U(P) = C\iint_{\mathcal{A}} \mathrm{e}^{-\mathrm{i}k(p\xi+q\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta, \tag{36}$$

C being the constant appearing in front of the integral (28). C is defined in terms of quantities depending on the position of the source and of the point of observation, but in practice it is often more convenient to express it in terms of other quantities. Let $\mathcal{P}$ be the total power incident upon the aperture. By the law of conservation of energy the total power that reaches the plane of observation must also be equal to $\mathcal{P}$, so that we have the normalizing condition

$$R^2\iint |U(p,\,q)|^2\,\mathrm{d}p\,\mathrm{d}q = \mathcal{P}, \tag{37}$$

where R is the distance from O (see Fig. 8.5) to the point at which the line P_0O intersects that plane. In (37) we used the fact that, for small angles of diffraction, $R^2\mathrm{d}p\mathrm{d}q$ is the area element of the observation plane formed by the diffracted rays, the integration extending over all effective p and q values. Eq. (36) may be re-written as a Fourier integral

$$U(p,\,q) = \iint G(\xi,\,\eta)\mathrm{e}^{-\frac{2\pi\mathrm{i}}{\lambda}(p\xi+q\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta, \tag{38}$$

where the *pupil function** G is given by

* More general pupil functions will be considered in §8.6 and §9.5.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$G(\xi, \eta) = \text{constant } (C) \text{ at points in the opening}$$
$$= 0 \qquad \text{at points outside opening} \tag{39}$$

and the integral extends over the whole ξ, η plane.

By Parseval's theorem for Fourier transforms*

$$\iint |G(\xi, \eta)|^2 \, \mathrm{d}\xi \, \mathrm{d}\eta = \left(\frac{1}{\lambda}\right)^2 \iint |U(p, q)|^2 \, \mathrm{d}p \, \mathrm{d}q, \tag{40}$$

or, substituting from (37) and (39) and denoting by D the area of the opening,

$$\frac{\mathcal{P}}{\lambda^2 R^2} = |C|^2 D \tag{41}$$

and consequently†

$$C = \frac{1}{\lambda R} \sqrt{\frac{\mathcal{P}}{D}}. \tag{42}$$

The basic integral for Fraunhofer diffraction then takes the form

$$U(p, q) = \frac{1}{\lambda R} \sqrt{\frac{\mathcal{P}}{D}} \iint_{\mathcal{A}} \mathrm{e}^{-\mathrm{i}k(p\xi + q\eta)} \, \mathrm{d}\xi \, \mathrm{d}\eta. \tag{43}$$

We note that the intensity $I_0 = |U(0, 0)|^2$ at the centre of the pattern $p = q = 0$ is given by

$$I_0 = \left(\frac{1}{\lambda R}\right)^2 \frac{\mathcal{P}}{D} \left(\iint_{\mathcal{A}} \mathrm{d}\xi \, \mathrm{d}\eta \right)^2 = \frac{\mathcal{P} D}{\lambda^2 R^2} = C^2 D^2. \tag{44}$$

In deriving (43) we have disregarded the fact that (36) was obtained subject to certain restrictions on the range of p and q. The errors introduced by extending the integration in (40) over all p and q values is, however, negligible, since $U(p, q)$ is very small except in the neighbourhood of $p = q = 0$.

Let us now return to the basic diffraction integral (28). As the point (ξ, η) explores the domain of integration, the function $f(\xi, \eta)$ will change by very many wavelengths, so that both the real and imaginary parts of the integrand will change sign many times. In consequence the contributions from the various elements will in general virtually cancel each other out (destructive interference). The situation is, however, different for an element which surrounds a point (called *critical point* or *pole*) where $f(\xi, \eta)$ is stationary. Here the integrand varies much more slowly and may be expected to give a significant contribution. Hence, when the wavelength is sufficiently small, the value of the integral is determined substantially by the behaviour of f in the neighbourhood of points where f is stationary. This is the principle of the *method of stationary phase* for determining the asymptotic behaviour of a certain class of integrals, and is discussed more fully in Appendix III. Here we only note the bearing of this result on the classification of diffraction phenomena:

* See I. N. Sneddon, *Fourier Transforms* (New York, McGraw-Hill, 1951), pp. 25 and 44.
† We omit here a constant phase factor as it contributes nothing to the intensity $I = |U|^2$.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

On comparing (22) and (28) we see that $r + s = r' + s' + f$, so that (see Fig. 8.5)

$$f = P_0Q + QP + \text{constant}. \tag{45}$$

Obviously f considered as function of Q will be stationary when Q is collinear with P_0 and P. Hence the main contribution to the disturbance at P comes from the immediate neighbourhood of the point $\overline{Q}$ where the line joining the source to the point of observation intersects the plane of the aperture. Now in the special case of Fraunhofer diffraction, P_0 and P are effectively at infinity, so that there is no preferential point $\overline{Q}$. In this case the behaviour of the diffraction integral must, therefore, be expected to be somewhat exceptional.

In §8.5–§8.8 we shall study the most important cases of Fraunhofer and Fresnel diffraction. But first we must justify the use of the single scalar wave function U in calculations of the light intensity.

8.4 Transition to a scalar theory*

The only property of the U function that we used in the derivation of the Kirchhoff integral theorem was that it satisfies the homogeneous scalar wave equation. It therefore follows that this theorem and the conclusion of the preceding section apply to each of the Cartesian components of the field vectors, the vector potential, the Hertz vectors, etc., in regions where there are no currents and charges. To obtain a complete description of the field, the theorem must be applied separately to each of the Cartesian components. Fortunately it turns out that in the majority of problems encountered in optics an approximate description in terms of a single complex scalar wave function is adequate.

A complete description of an electromagnetic field requires the specification of the magnitude of the field vectors as well as their direction (polarization), both as functions of position and time. However, because of the very high frequencies of optical fields (of the order of 10^{14}/s), one cannot measure the instantaneous values of any of these quantities, but only certain time averages over intervals that are large compared to the optical periods. Moreover, one usually deals with natural light, so that there is no preferential polarization direction of the observable (macroscopic) field. The quantity which is then of primary interest is the *intensity* I defined in §1.1.4 as the time average of the energy which crosses a unit area containing the electric and magnetic vector in unit time

$$I = \frac{c}{4\pi}|\langle \mathbf{E} \times \mathbf{H} \rangle|.$$

We shall show that the electromagnetic field which is associated with the passage of natural light through an optical instrument of moderate aperture and of conventional design is such that the intensity may approximately be represented in terms of a single complex scalar wave function by means of the formula†

* We follow here substantially the analysis of O. Theimer, G. D. Wassermann and E. Wolf, *Proc. Roy. Soc.*, A, **212** (1952), 426.

† More generally it was shown by E. Wolf, *Proc. Phys. Soc.*, **74** (1959), 269, that both the (time-averaged) energy density and energy flow in an unpolarized quasi-monochromatic field may always be derived from one complex time-harmonic scalar wave function.

$$I = |U|^2,$$

and that the function U may be calculated from the knowledge of the eikonal function of the system.

8.4.1 The image field due to a monochromatic oscillator

We consider a symmetrical optical system with a point source at P_0 (Fig. 8.7) emitting natural, quasi-monochromatic light of frequency ω_0. We assume that the inclination to the axis of the rays which pass through the system is not large, say not more than 10° or 15°. At P_0 we choose a set of Cartesian axes (x_1, x_2, x_3) with the x_3 direction along the principal ray. The source may be regarded as a dipole of moment $\mathbf{Q}(t)$ which varies both in magnitude and direction with time t. The components of $\mathbf{Q}(t)$ in the three directions will be written in the form of Fourier integrals,

$$Q_j(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} q_j(\omega) \mathrm{e}^{-\mathrm{i}\omega t}\, \mathrm{d}\omega \qquad (j = 1, 2, 3). \tag{1}$$

Since $Q_j(t)$ is real it follows that the complex quantities $q_j(\omega)$ satisfy the relations

$$q_j(-\omega) = q_j{}^{\star}(\omega), \tag{2}$$

where the asterisk denotes the complex conjugate. Consequently (1) may be written as

$$Q_j(t) = \mathcal{R}\left\{ \sqrt{\frac{2}{\pi}} \int_0^{\infty} q_j(\omega) \mathrm{e}^{-\mathrm{i}\omega t}\, \mathrm{d}\omega \right\} \qquad (j = 1, 2, 3), \tag{3}$$

$\mathcal{R}$ denoting the real part. Each Fourier component of (3) represents a monochromatic Hertzian oscillator with its axis along the x_j direction.

Let $|q_j(\omega)|$ and $\delta_j(\omega)$ be the amplitude and the phase of $q_j(\omega)$,

$$q_j(\omega) = |q_j(\omega)| \mathrm{e}^{\mathrm{i}\delta_j(\omega)}. \tag{4}$$

Since the source is assumed to emit quasi-monochromatic light, the modulus $|q_j(\omega)|$ will, for each j, differ appreciably from zero only within a narrow interval $(\omega_0 - \frac{1}{2}\Delta\omega,$

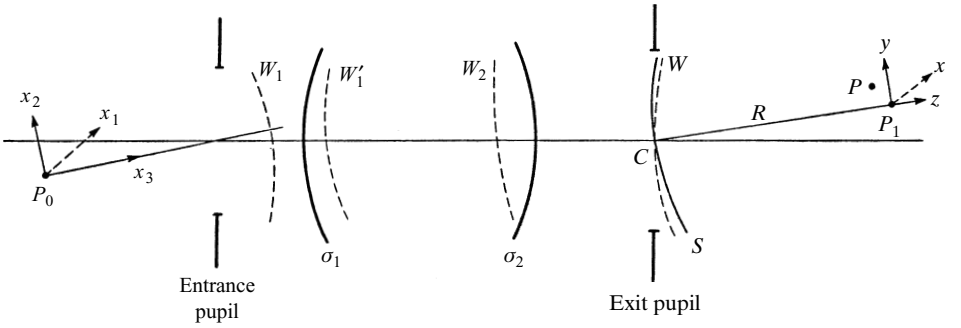


Fig. 8.7 Propagation of an electromagnetic wave through an optical system.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$\omega_0 + \frac{1}{2}\Delta\omega$). The assumption that the light is natural implies that $\delta_j(\omega)$ are rapidly and irregularly varying functions over the frequency range.*

Since the field may be regarded as a superposition of strictly monochromatic fields, it will be convenient to examine first the contributions from a single monochromatic Hertzian oscillator at P_0. As the field of such an oscillator is weak in the neighbourhood of its axis, and as we assume that the angles which the diameters of the entrance pupil subtend at P_0 are small, it follows that only the components $Q_1(t)$ and $Q_2(t)$ of $\mathbf{Q}(t)$ will substantially contribute to the field. We shall therefore take as our typical oscillator one which has its axis in the x_1, x_2-plane.

Let

$$\mathcal{R}\{q(\omega)\boldsymbol{\rho}_0(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\} \tag{5}$$

be the moment of this typical dipole, $\boldsymbol{\rho}_0(\omega)$ being a unit vector in the direction of its axis. Such a dipole will produce at a point T in vacuum, whose distance from P_0 is large compared to the wavelength $\lambda = 2\pi c/\omega$, a field given by (see §2.2 (64)):

$$\left.\begin{aligned}\mathbf{E}_\omega &= \mathcal{R}\left\{\frac{\omega^2}{c^2 r}|q(\omega)|\mathbf{r}_0 \times (\boldsymbol{\rho}_0(\omega) \times \mathbf{r}_0)\mathrm{e}^{\mathrm{i}[\delta(\omega)-\omega(t-r/c)]}\right\},\\ \mathbf{H}_\omega &= \mathcal{R}\left\{\frac{\omega^2}{c^2 r}|q(\omega)|\mathbf{r}_0 \times \boldsymbol{\rho}_0(\omega)\mathrm{e}^{\mathrm{i}[\delta(\omega)-\omega(t-r/c)]}\right\},\end{aligned}\right\} \tag{6}$$

where $\mathbf{r}_0$ denotes the unit radial vector.

Let W_1 be a typical geometrical wave-front in the object space at a distance from P_0 which is large compared with the wavelength. Since we assume that the angles which the rays make with the axis of the system are small, it follows immediately from (6) that at any particular instant of time the vectors $\mathbf{E}_\omega$ and $\mathbf{H}_\omega$ do not vary appreciably in magnitude and direction over W_1.

The effect of the first surface† σ_1 on the incident field is twofold. First, the amplitudes of the field vectors are diminished on account of reflection losses; secondly, the directions of vibrations are changed. Fresnel's formulae show that both these effects depend mainly on the magnitudes of the angle of incidence. If this angle is small (say 10° or so), reflection losses are also small (approx. 5 per cent) and the rotations of the planes of vibration do not exceed a few degrees (see §1.5). Moreover, these effects are practically uniform over σ_1. Since the time-independent parts of $\mathbf{E}_\omega$ and $\mathbf{H}_\omega$ do not vary appreciably with position over the wave-front W_1, they will also not vary appreciably over the refracted wave-front W_1' which follows the surface σ_1 (see Fig. 8.7). The same applies to the behaviour of the two fields over any other wave-front in the space between σ_1 and the second surface σ_2. For, as we showed in §3.1.3, in a homogeneous medium the direction of vibration along each ray remains constant, and also, since the wave-fronts are nearly spherical (centred on the Gaussian image of P_0 by the first surface), the amplitudes will be diminished almost in the ratio of their paraxial radii of curvature.

Repeating these arguments we finally arrive at a wave-front W which passes through

* For a detailed discussion of this point see M. Planck, *Ann. d. Physik* (4), **1** (1900), 61.

† We assume here that σ_1 is a refracting surface. If σ_1 is a mirror, no essential modifications of our argument are necessary, as is seen by inspection of Fresnel's formulae.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

the centre C of the exit pupil and find again that the time-independent parts of $\mathbf{E}_\omega$ and $\mathbf{H}_\omega$ do not vary appreciably over this wave-front. This result makes it immediately possible to write down an approximate mathematical representation for the field vectors in the region of the image.

We take rectangular Cartesian axes $(x,\ y,\ z)$ with origin at the Gaussian image P_1 of P_0, with the z direction along CP_1. The field at all points in the region of the aperture except those in the immediate neighbourhood of the edge of the aperture can be approximately expressed in the form (see Chapter III)

$$\left.\begin{aligned}\mathbf{E}_\omega(x,\ y,\ z,\ t) &= \mathcal{R}\left\{\frac{\omega^2}{c^2}\mathbf{e}_\omega(x,\ y,\ z)\mathrm{e}^{\mathrm{i}\{\delta(\omega)-\omega[t-\frac{1}{c}\mathcal{S}_\omega(x,y,z)]\}}\right\},\\ \mathbf{H}_\omega(x,\ y,\ z,\ t) &= \mathcal{R}\left\{\frac{\omega^2}{c^2}\mathbf{h}_\omega(x,\ y,\ z)\mathrm{e}^{\mathrm{i}\{\delta(\omega)-\omega[t-\frac{1}{c}\mathcal{S}_\omega(x,y,z)]\}}\right\},\end{aligned}\right\} \tag{7}$$

which may be regarded as generalization of (6). Hence $\mathcal{S}_\omega(x,\ y,\ z)$ is the optical length from the object point to the point $(x,\ y,\ z)$, and $\mathbf{e}_\omega(x,\ y,\ z)$ and $\mathbf{h}_\omega(x,\ y,\ z)$ are mutually orthogonal real vectors.* In a homogeneous non-magnetic medium of refractive index n, these vectors satisfy the relation [see §3.1 (19) and §3.1 (20)]

$$|\mathbf{h}_\omega| = n|\mathbf{e}_\omega|. \tag{8}$$

We take a reference sphere S, centred on P_1, which passes through the centre C of the exit pupil, and denote by R its radius CP_1. In practice the distance between S and W will nowhere exceed a few dozen wavelengths. Consequently on S just as on W the amplitude vectors $\mathbf{e}_\omega$ and $\mathbf{h}_\omega$ will be practically constant in magnitude and direction.

Let $P(X,\ Y,\ Z)$ be a point in the region of the image where the intensity is to be determined. If the angles which the diameters of the exit pupil subtend at P are small, we may apply Kirchhoff's formula with the same approximation as in the previous section, and we find on integrating the expressions (7) over that part S' of S which approximately fills the exit pupil, if in addition we also neglect the variation of the inclination factor over S', that

$$\left.\begin{aligned}\mathbf{E}_\omega(X,\ Y,\ Z,\ t) &= \mathcal{R}\frac{\omega^3}{2\pi\mathrm{i}c^3}\mathrm{e}^{\mathrm{i}[\delta(\omega)-\omega t]}\iint_{S'}\frac{1}{s}\mathbf{e}_\omega(x',\ y',\ z')\mathrm{e}^{\mathrm{i}\frac{\omega}{c}[\mathcal{S}_\omega(x',y',z')+s]}\,\mathrm{d}S,\\ \mathbf{H}_\omega(X,\ Y,\ Z,\ t) &= \mathcal{R}\frac{\omega^3}{2\pi\mathrm{i}c^3}\mathrm{e}^{\mathrm{i}[\delta(\omega)-\omega t]}\iint_{S'}\frac{1}{s}\mathbf{h}_\omega(x',\ y',\ z')\mathrm{e}^{\mathrm{i}\frac{\omega}{c}[\mathcal{S}_\omega(x',y',z')+s]}\,\mathrm{d}S,\end{aligned}\right\} \tag{9}$$

where s is the distance from a typical point $(x',\ y',\ z')$ on the reference sphere to P.

Since the vectors $\mathbf{e}_\omega(x',\ y',\ z')$ and $\mathbf{h}_\omega(x',\ y',\ z')$ do not vary appreciably over the surface of integration we may replace them by the values $\mathbf{e}_\omega(0,\ 0,\ -R)$ and $\mathbf{h}_\omega(0,\ 0,\ -R)$ which they take at the centre C of the exit pupil. Now these vectors are orthogonal and satisfy (8), so that we may set, if in addition we take $n = 1$,

$$\left.\begin{aligned}\mathbf{e}_\omega(0,\ 0,\ -R) &= a(\omega)\boldsymbol{\alpha}(\omega),\\ \mathbf{h}_\omega(0,\ 0,\ -R) &= a(\omega)\boldsymbol{\beta}(\omega),\end{aligned}\right\} \tag{10}$$

* That $\mathbf{e}_\omega$ and $\mathbf{h}_\omega$ are real follows from the fact that the corresponding vectors in (6) are real (linear polarization) and that the state of polarization remains linear on each refraction (see §1.5.2). Moreover, between any two consecutive surfaces, the state of polarization is constant along each ray, as shown in §3.1.3.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

where $\boldsymbol{\alpha}(\omega)$ and $\boldsymbol{\beta}(\omega)$ are orthogonal unit vectors in the plane perpendicular to the z direction. The relations (9) then become

$$\left.\begin{aligned}\mathbf{E}_\omega(X,\,Y,\,Z,\,t) &= \mathcal{R}\left\{\frac{\omega^2}{c^2}\,U_\omega(X,\,Y,\,Z)a(\omega)\boldsymbol{\alpha}(\omega)\mathrm{e}^{\mathrm{i}[\delta(\omega)-\omega t]}\right\},\\ \mathbf{H}_\omega(X,\,Y,\,Z,\,t) &= \mathcal{R}\left\{\frac{\omega^2}{c^2}\,U_\omega(X,\,Y,\,Z)a(\omega)\boldsymbol{\beta}(\omega)\mathrm{e}^{\mathrm{i}[\delta(\omega)-\omega t]}\right\},\end{aligned}\right\} \tag{11}$$

where U_ω is the scalar wave function

$$U_\omega(X,\,Y,\,Z) = \frac{\omega}{2\pi\mathrm{i}c}\iint_{S'}\frac{1}{s}\,\mathrm{e}^{\frac{\mathrm{i}\omega}{c}[\mathcal{S}_\omega(x',y',z')+s]}\,\mathrm{d}S. \tag{12}$$

From (11) we can immediately deduce by calculating the Poynting vector $\mathbf{S}_\omega = c[\mathbf{E}_\omega \times \mathbf{H}_\omega]/4\pi$ and taking the time average, that the intensity at the point $P(X,\,Y,\,Z)$ due to the single dipole [represented by (5)] at P_0 is proportional to the square of the modulus of the scalar wave function $U_\omega(X,\,Y,\,Z)$. However, to justify the use of a single scalar wave function in calculating the intensity we must carry out the time averaging not for the monochromatic component but for the total field.

8.4.2 The total image field

We saw that the contributions of each frequency component to the total field may be regarded as arising essentially from two dipoles at P_0 with their axes along the x_1 and x_2 directions. Hence it follows from (1) and (11), if we also define contributions from negative frequencies by relations of the form (2), that the total field in the image region may be expressed approximately in the form

$$\left.\begin{aligned}\mathbf{E}(X,\,Y,\,Z,\,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{\omega^2}{c^2}\,U_\omega(X,\,Y,\,Z)[a_1(\omega)\boldsymbol{\alpha}_1(\omega)\mathrm{e}^{\mathrm{i}\delta_1(\omega)}\\ &\qquad + a_2(\omega)\boldsymbol{\alpha}_2(\omega)\mathrm{e}^{\mathrm{i}\delta_2(\omega)}]\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega,\\ \mathbf{H}(X,\,Y,\,Z,\,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{\omega^2}{c^2}\,U_\omega(X,\,Y,\,Z)[a_1(\omega)\boldsymbol{\beta}_1(\omega)\mathrm{e}^{\mathrm{i}\delta_1(\omega)}\\ &\qquad + a_2(\omega)\boldsymbol{\beta}_2(\omega)\mathrm{e}^{\mathrm{i}\delta_2(\omega)}]\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega.\end{aligned}\right\} \tag{13}$$

Here suffixes 1 and 2 refer to the contributions from oscillators which have their axes along the x_1 and x_2 directions.

In order to determine the intensity in the image region it will be convenient to write down separate expressions for each of the Cartesian components of $\mathbf{E}$ and $\mathbf{H}$. Let $\theta_1(\omega)$ and $\theta_2(\omega)$ denote the angles which the unit vectors $\boldsymbol{\alpha}_1(\omega)$ and $\boldsymbol{\alpha}_2(\omega)$ make with the x direction in the image space. Since $\boldsymbol{\alpha}_1(\omega)$ and $\boldsymbol{\beta}_1(\omega)$ and $\boldsymbol{\alpha}_2(\omega)$ and $\boldsymbol{\beta}_2(\omega)$ are real, mutually orthogonal vectors which lie in a plane perpendicular to the z direction, it follows from (13) that the components of $\mathbf{E}$ and $\mathbf{H}$ are approximately given by*

* It would be incorrect to conclude from (14) that the direction of the energy flow in the image region is necessarily everywhere parallel to z. For the relative errors in (14) may substantially affect the calculations of direction in regions where the intensity is small, e.g. in the neighbourhood of the dark rings in the Airy pattern.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$\left.\begin{aligned} E_x(X,\,Y,\,Z,\,t) = H_y(X,\,Y,\,Z,\,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega(X,\,Y,\,Z) f(\omega) \mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega, \\ E_y(X,\,Y,\,Z,\,t) = -H_x(X,\,Y,\,Z,\,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega(X,\,Y,\,Z) g(\omega) \mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega, \\ E_z(X,\,Y,\,Z,\,t) = H_z(X,\,Y,\,Z,\,t) &= 0, \end{aligned}\right\} \tag{14}$$

where

$$\left.\begin{aligned} f(\omega) &= \frac{\omega^2}{c^2}[a_1(\omega)\cos\theta_1(\omega)\mathrm{e}^{\mathrm{i}\delta_1(\omega)} + a_2(\omega)\cos\theta_2(\omega)\mathrm{e}^{\mathrm{i}\delta_2(\omega)}], \\ g(\omega) &= \frac{\omega^2}{c^2}[a_1(\omega)\sin\theta_1(\omega)\mathrm{e}^{\mathrm{i}\delta_1(\omega)} + a_2(\omega)\sin\theta_2(\omega)\mathrm{e}^{\mathrm{i}\delta_2(\omega)}]. \end{aligned}\right\} \tag{15}$$

It follows from (14) that the magnitude of the Poynting vector $\mathbf{S} = c[\mathbf{E} \times \mathbf{H}]/4\pi$ can be expressed approximately in the form

$$|\mathbf{S}| = \frac{c}{4\pi}[E_x^2 + E_y^2] = \frac{c}{4\pi}[H_x^2 + H_y^2]. \tag{16}$$

We must now determine the time average of this quantity.

For reasons of convergence we assume that the radiation field exists only between the instants $t = -T$ and $t = T$, where $T \gg 2\pi/\omega_0$; it is easy to pass to the limit $T \to \infty$ subsequently. It follows from (14), by the Fourier inversion theorem, that

$$U_\omega(X,\,Y,\,Z) f(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-T}^{T} E_x(X,\,Y,\,Z,\,t)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}t, \tag{17}$$

with similar expressions involving E_y, H_x and H_y. Now we have by (14)

$$\langle E_x^2\rangle = \frac{1}{2T}\int_{-T}^{T} E_x^2\,\mathrm{d}t = \frac{1}{2T}\int_{-T}^{T} E_x\,\mathrm{d}t\,\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega f(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega, \tag{18}$$

or, inverting the order of integration,

$$\begin{aligned} \langle E_x^2\rangle &= \frac{1}{2T}\int_{-\infty}^{+\infty} U_\omega f(\omega)\,\mathrm{d}\omega\,\frac{1}{\sqrt{2\pi}}\int_{-T}^{T} E_x\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t \\ &= \frac{1}{2T}\int_{-\infty}^{+\infty} U_\omega f_\omega U_\omega^\star f_\omega^\star\,\mathrm{d}\omega \qquad \text{by (17)} \\ &= \frac{1}{T}\int_0^{\infty} |U_\omega|^2 |f(\omega)|^2\,\mathrm{d}\omega, \end{aligned} \tag{19}$$

since $U_{-\omega} f(-\omega) = U_\omega^\star f^\star(\omega)$. Similarly

$$\langle E_y^2\rangle = \frac{1}{T}\int_0^{\infty} |U_\omega|^2 |g(\omega)|^2\,\mathrm{d}\omega. \tag{20}$$

Hence, the intensity $I(X,\,Y,\,Z)$, defined as the time average of the magnitude of the Poynting vector, is, according to (16), (19) and (20),

$$I(X,\,Y,\,Z) = \frac{c}{4\pi T}\int_0^{\infty} |U_\omega(X,\,Y,\,Z)|^2[|f(\omega)|^2 + |g(\omega)|^2]\,\mathrm{d}\omega. \tag{21}$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Now if $|\Delta\omega|$ is sufficiently small $|U_\omega|$ will be practically independent of ω over the effective frequency range, so that $|U_\omega|$ may then be replaced by $|U_{\omega_0}|$ and taken outside the integral. The remaining term

$$\frac{c}{4\pi T}\int_0^\infty \{|f(\omega)|^2 + |g(\omega)|^2\}\mathrm{d}\omega, \tag{22}$$

which is independent of X, Y, and Z, must also be independent of T (implicitly contained in f and g on account of (17)) if a stationary phenomenon is observed. Hence (22) must be a constant (C_0 say) and the intensity may therefore be finally written in the form

$$I(X, Y, Z) = C_0|U_{\omega_0}(X, Y, Z)|^2. \tag{23}$$

The constant C_0 depends in a complicated manner on the source and on the optical instrument; however, one is usually only interested in the relative distribution of the intensity and not in its absolute value. The intensity may then simply be measured by the quantity $|U_{\omega_0}|^2$. Thus the complex scalar function (12) is adequate for calculating the intensity distribution in the image formed with a source of natural light by an optical system of moderate numerical aperture.

8.5 Fraunhofer diffraction at apertures of various forms

We shall now investigate the Fraunhofer diffraction pattern for apertures of various forms.

8.5.1 The rectangular aperture and the slit

Consider first a rectangular aperture of sides $2a$ and $2b$. With origin O at the centre of the rectangle and with $O\xi$ and $O\eta$ axes parallel to the sides (Fig. 8.8), the Fraunhofer diffraction integral §8.3 (36) becomes

$$U(P) = C\int_{-a}^{a}\int_{-b}^{b} \mathrm{e}^{-\mathrm{i}k(p\xi+q\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta = C\int_{-a}^{a}\mathrm{e}^{-\mathrm{i}kp\xi}\,\mathrm{d}\xi\int_{-b}^{b}\mathrm{e}^{-\mathrm{i}kq\eta}\,\mathrm{d}\eta.$$

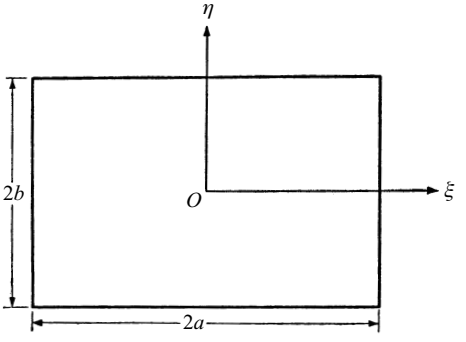


Fig. 8.8 Rectangular aperture.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Now

$$\int_{-a}^{a} \mathrm{e}^{-\mathrm{i}kp\xi}\,\mathrm{d}\xi = -\frac{1}{\mathrm{i}kp}[\mathrm{e}^{-\mathrm{i}kpa} - \mathrm{e}^{\mathrm{i}kpa}] = 2\frac{\sin kpa}{kp},$$

with a similar expression for the other integral. Hence the intensity is given by

$$I(P) = |U(P)|^2 = \left(\frac{\sin kpa}{kpa}\right)^2 \left(\frac{\sin kqb}{kqb}\right)^2 I_0, \tag{1}$$

where by §8.3 (44) $I_0 = C^2D^2 = \mathcal{P}D/\lambda^2R^2$ is the intensity at the centre of the pattern, $\mathcal{P}$ being the total power incident upon the aperture and $D = 4ab$ the area of the rectangle.

The function $y = (\sin x/x)^2$ is displayed in Fig. 8.9. It has a principal maximum $y = 1$ at $x = 0$ and zero minima at $x = \pm\pi, \pm2\pi, \pm3\pi, \ldots$. The minima separate the secondary maxima whose positions are given by the roots of the equation $\tan x - x = 0$ (see Table 8.1). The roots asymptotically approach the values $x = (2m + 1)\pi/2$, m being an integer.

We see that the intensity $I(P)$ is zero along two sets of lines parallel to the sides of the rectangle, given by

$$kpa = \pm u\pi, \qquad kqb = \pm v\pi \qquad (u, v = 1, 2, 3, \ldots) \tag{2}$$

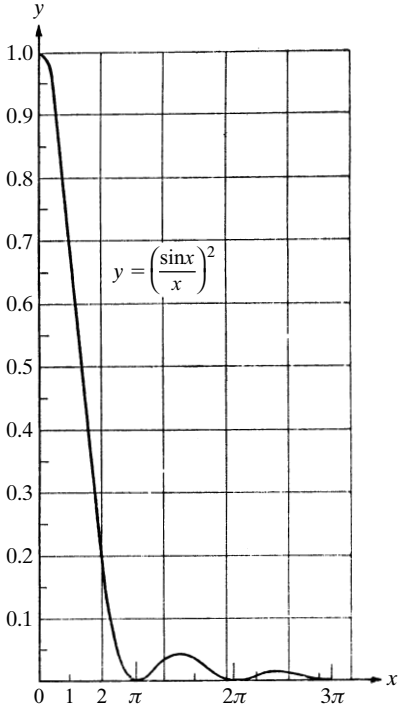


Fig. 8.9 Fraunhofer diffraction at a rectangular aperture. The function

$$y = \left(\frac{\sin x}{x}\right)^2.$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Table 8.1. *The first five maxima of the function*

$$y = \left(\frac{\sin x}{x}\right)^2.$$

x	$y = \left(\frac{\sin x}{x}\right)^2$
0	1
$1.430\pi = 4.493$	0.04718
$2.459\pi = 7.725$	0.01648
$3.470\pi = 10.90$	0.00834
$4.479\pi = 14.07$	0.00503

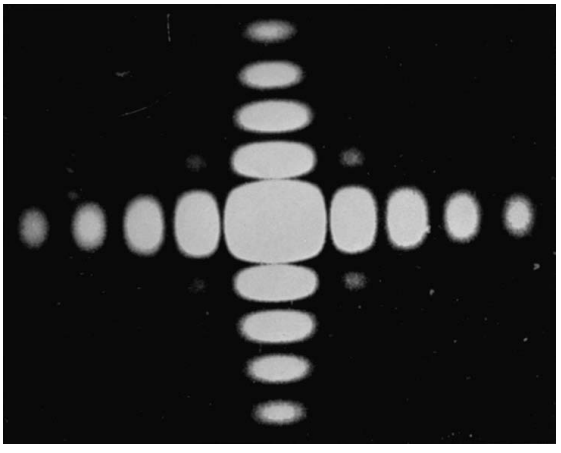


Fig. 8.10 Fraunhofer diffraction pattern of a rectangular aperture 8 mm × 7 mm, magnification 50×, mercury yellow light $\lambda = 5790$ Å. To show the existence of the weak secondary maxima the central portion was overexposed. (Photograph courtesy of H. Lipson, C. A. Taylor, and B. J. Thompson.)

or, since $p = l - l_0$, $q = m - m_0$, $k = 2\pi/\lambda$,

$$l - l_0 = \pm\frac{u\lambda}{2a}, \qquad m - m_0 = \pm\frac{v\lambda}{2b}. \tag{3}$$

Within each rectangle formed by pairs of consecutive dark lines the intensity rises to a maximum; all these maxima are, however, only a small fraction of the central maximum, and decrease rapidly with increasing distance from the centre (Fig. 8.10). The larger the opening, the smaller is the effective size of the diffraction pattern.

From the elementary diffraction pattern formed by coherent light from a point source, the diffraction pattern due to light from an extended source may be found by integration. If the source is coherent, it is the complex amplitude, and if it is incoherent, it is the intensity that must be integrated. The pattern due to a partially coherent source may also be determined from this elementary solution by a process of integration, taking into account the correlation which exists between the light from

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

the different elements of the source (see Chapter X). A case of particular importance is that of a very long incoherent line source (e.g. a luminous wire), the light from which is diffracted by a narrow slit, parallel to the source. For simplicity of calculations we assume that the luminous wire as well as the slit are effectively infinitely long, and take the y-axis in the direction of the source. Since $q = m - m_0$ where m_0 specifies the position of a point source, it follows that the intensity I' due to the line source is obtained by integrating (1) with respect to q:

$$I' = \int_{-\infty}^{+\infty} I(P)\mathrm{d}q = \frac{1}{kb}\left(\frac{\sin kpa}{kpa}\right)^2 I_0 \int_{-\infty}^{+\infty}\left(\frac{\sin t}{t}\right)^2 \mathrm{d}t.$$

Now*

$$\int_{-\infty}^{+\infty}\left(\frac{\sin t}{t}\right)^2 \mathrm{d}t = \pi,$$

so that

$$I' = \left(\frac{\sin kpa}{kpa}\right)^2 I_0', \tag{4}$$

where

$$I_0' = \frac{\lambda}{2b} I_0 = \frac{2a\mathcal{P}}{\lambda R^2}. \tag{5}$$

The pattern is again characterized by the function $[\sin(x)/x]^2$, and consists of a succession of bright and dark fringes parallel to the line source and the slit. The constant $I_0{}'$ is the intensity at the central position $p = 0$.

8.5.2 The circular aperture

In a similar way we may investigate Fraunhofer diffraction at a circular aperture. It is now appropriate to use polar instead of rectangular coordinates. Let (ρ, θ) be the polar coordinates of a typical point in the aperture:

$$\rho\cos\theta = \xi, \qquad \rho\sin\theta = \eta; \tag{6}$$

and let (w, ψ) be the coordinates of a point P in the diffraction pattern, referred to the geometrical image of the source:

$$w\cos\psi = p, \qquad w\sin\psi = q. \tag{7}$$

From the definition of p and q it follows that $w = \sqrt{p^2 + q^2}$ is the sine of the angle which the direction (p, q) makes with the central direction $p = q = 0$. The diffraction integral §8.3 (36) now becomes, if a is the radius of the circular aperture,

$$U(P) = C\int_0^a \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}k\rho w\cos(\theta-\psi)}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta. \tag{8}$$

* See, for example, W. Gröbner and N. Hofreiter, *Integraltafel*, Vol. II (Wien, Springer, 1950), p. 333.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Now we have the well-known integral representation of the Bessel functions* $J_n(z)$:

$$\frac{\mathrm{i}^{-n}}{2\pi}\int_0^{2\pi} \mathrm{e}^{\mathrm{i}x\cos\alpha}\mathrm{e}^{\mathrm{i}n\alpha}\,\mathrm{d}\alpha = J_n(x). \tag{9}$$

Eq. (8) therefore reduces to

$$U(P) = 2\pi C\int_0^a J_0(k\rho w)\rho\,\mathrm{d}\rho. \tag{10}$$

Also, there is the well-known recurrence relation†

$$\frac{\mathrm{d}}{\mathrm{d}x}[x^{n+1}J_{n+1}(x)] = x^{n+1}J_n(x), \tag{11}$$

giving, for $n = 0$, on integration

$$\int_0^x x'J_0(x')\mathrm{d}x' = xJ_1(x). \tag{12}$$

From (10) and (12) it follows that

$$U(P) = CD\left[\frac{2J_1(kaw)}{kaw}\right], \tag{13}$$

where $D = \pi a^2$. Hence the intensity is given by

$$I(P) = |U(P)|^2 = \left[\frac{2J_1(kaw)}{kaw}\right]^2 I_0, \tag{14}$$

where by §8.3 (44) $I_0 = C^2D^2 = \mathcal{P}D/\lambda^2R^2$. This is a celebrated formula first derived in a somewhat different form by Airy.‡

The intensity distribution in the neighbourhood of the geometrical image is characterized by the function $y = (2J_1(x)/x)^2$ shown in Fig. 8.11. It has its principal maximum $y = 1$ at $x = 0$, and with increasing x it oscillates with gradually diminishing amplitude, in a similar way to the function $[\sin(x)/x]^2$ which we discussed in §8.5.1. The intensity is zero (minimum) for values of x given by $J_1(x) = 0$. The minima are no longer strictly equidistant (see Table 8.2). The positions of the secondary maxima are given by the values of x that satisfy the equation

$$\frac{\mathrm{d}}{\mathrm{d}x}[J_1(x)/x] = 0,$$

* See, for example, E. Jahnke and F. Emde, *Tables of Functions with Formulae and Curves* (Leipzig, Teubner, 1933; reprinted by Dover Publications, New York, 4th edition, 1945), p. 149; or G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge, Cambridge University Press, 1922), p. 20, equation 5 (with an obvious substitution).

† See, for example, E. Jahnke and F. Emde, *loc. cit*., p. 145 or E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge, Cambridge University Press, 4th edition, 1952), pp. 360–361.

‡ G. B. Airy, *Trans. Camb. Phil. Soc.*, **5** (1835), 283. Almost at the same time as Airy, Schwerd obtained an approximate solution by replacing the circle by a regular polygon with 180 sides.

Vectorial treatments of diffraction of a convergent spherical wave at a circular aperture, which take into account polarization properties of the field were published by W. S. Ignatowski, *Trans. Opt. Inst. Petrograd.*, **1** (1919) No 4, 36; V. A. Fock, *ibid*., **3** (1924), 24; H. H. Hopkins, *Proc. Phys. Soc*. **55** (1943), 116; R. Burtin, *Optica Acta*, **3** (1956), 104; B. Richards and E. Wolf, *Proc. Roy. Soc.*, A, **253** (1959), 358; A. Boivin and E. Wolf, *Phys. Rev.*, **138** (1965), B 1561; A. Boivin, J. Dow and E. Wolf, *J. Opt. Soc. Amer.*, **57** (1967), 1171.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Table 8.2. *The first few maxima and minima of the function* $y = \left[\frac{2J_1(x)}{x}\right]^2$.

x	$\left[\frac{2J_1(x)}{x}\right]^2$	
0	1	Max
$1.220\pi = 3.833$	0	Min
$1.635\pi = 5.136$	0.0175	Max
$2.233\pi = 7.016$	0	Min
$2.679\pi = 8.417$	0.0042	Max
$3.238\pi = 10.174$	0	Min
$3.699\pi = 11.620$	0.0016	Max

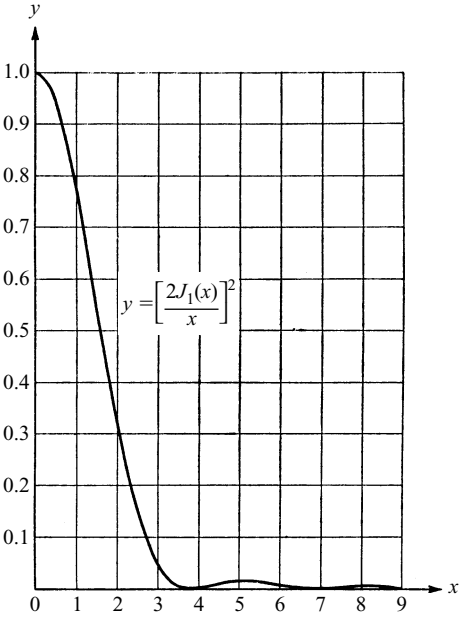


Fig. 8.11 Fraunhofer diffraction at a circular aperture. The function $y = \left[\frac{2J_1(x)}{x}\right]^2$.

or using the formula* (analogous to (11))

$$\frac{\mathrm{d}}{\mathrm{d}x}[x^{-n}J_n(x)] = -x^{-n}J_{n+1}(x), \tag{15}$$

by the roots of the equations $J_2(x) = 0$. With increasing x the separation between two

* See for example E. Jahnke and F. Emde, *loc. cit.*, p. 145, or E. T. Whittaker and G. N. Watson, *loc. cit.*, p. 361.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

successive minima or two successive maxima approaches the value π, as in the previous case.

The results show that the pattern consists of a bright disc, centred on the geometrical image $p = q = 0$ of the source, surrounded by concentric bright and dark rings (see Figs. 8.11 and 8.12). The intensity of the bright rings decreases rapidly with their radius and normally only the first one or two rings being bright enough to be visible to the naked eye. From Table 8.2 it follows, since $x = 2\pi aw/\lambda$, that the angular radii of the dark rings are

$$w = \sqrt{p^2 + q^2} = 0.610\frac{\lambda}{a}, \qquad 1.116\frac{\lambda}{a}, \qquad 1.619\frac{\lambda}{a}, \qquad \ldots . \tag{16}$$

The angular separation between two neighbouring rings approaches asymptotically the value $\lambda/2a$. The effective size of the diffraction pattern is again seen to be inversely proportional to the linear dimensions of the aperture.

It is also of interest to examine what fraction of the total incident energy is contained within the central core of the diffraction pattern. Let $L(w_0)$ denote the fraction of the total energy contained within a circle in the image plane, centred on the geometrical image and subtending a small angular radius w_0 at the center of the aperture. Then

$$\begin{aligned} L(w_0) &= \frac{R^2}{\mathcal{P}} \int_0^{w_0} \int_0^{2\pi} I(w) w \, \mathrm{d}w \, \mathrm{d}\psi \\ &= \frac{D}{\lambda^2} \int_0^{w_0} \int_0^{2\pi} \left[\frac{2J_1(kaw)}{kaw}\right]^2 w \, \mathrm{d}w \, \mathrm{d}\psi \\ &= 2\int_0^{kaw_0} \frac{J_1^2(x)}{x} \, \mathrm{d}x. \end{aligned} \tag{17}$$

Now from (11) for $n = 0$, we have, on multiplying by $J_1(x)$ and using (15) with $n = 0$,

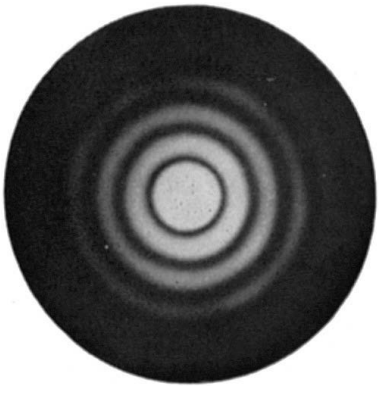


Fig. 8.12 Fraunhofer diffraction pattern of a circular aperture (the Airy pattern) 6 mm in diameter, magnification 50×, mercury yellow light $\lambda = 5790$ Å. To show the existence of the weak subsidiary maxima, the central portion was overexposed. (Photograph courtesy of H. Lipson, C. A. Taylor, and B. J. Thompson.)

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

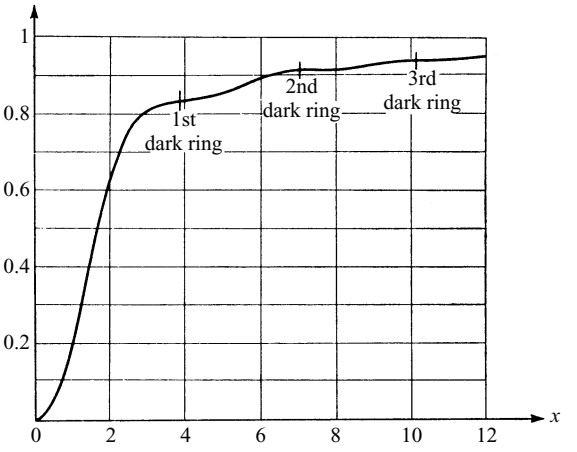


Fig. 8.13 The function $1 - J_0^2(x) - J_1^2(x)$ representing the fraction of the total energy contained within circles of prescribed radii in the Fraunhofer diffraction pattern of a circular aperture.

$$\frac{J_1^2(x)}{x} = J_0(x)J_1(x) - \frac{\mathrm{d}J_1(x)}{\mathrm{d}x}J_1(x)$$
$$= -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}[J_0^2(x) + J_1^2(x)].$$

The expression (17) now becomes, remembering that $J_0(0) = 1$, $J_1(0) = 0$,

$$L(w_0) = 1 - J_0^2(kaw_0) - J_1^2(kaw_0), \tag{18}$$

a formula due to Rayleigh.* This function is shown in Fig. 8.13. For the dark rings $J_1(kaw_0) = 0$, so that the fraction of the total energy outside any dark ring is simply $J_0^2(kaw_0)$. For the first, second and third dark rings, $J_0^2(kaw_0)$ is equal to 0.162, 0.090, and 0.062 respectively. Thus more than 90 per cent of the light is contained within the circle bounded by the second dark ring.

8.5.3 Other forms of aperture

Fraunhofer diffraction at apertures of other forms may be studied in a similar manner, the calculations being particularly simple when curvilinear coordinates can be chosen so that one of the coordinate lines coincides with the boundary of the aperture. We cannot discuss other cases in detail here,† but we shall derive a useful theorem concerning the modification of the pattern when the aperture is uniformly extended (or

* Lord Rayleigh, *Phil. Mag.* (5), **11** (1881), 214. Also *Scientific papers by John William Strutt, Baron Rayleigh*, Vol. 1 (Cambridge, Cambridge University Press, 1899–1920), p. 513.

† Fraunhofer diffraction at an annular aperture is briefly considered in connection with resolving power in §8.6.2.

Photographs of Fraunhofer diffraction patterns for apertures of various forms can be found in a paper by J. Scheiner and S. Hirayama, *Abh. d. Königl. Akad. Wissensch.*, Berlin (1894), Anhang I. Photographs of Fresnel patterns were published by Y. V. Kathavate, *Proc. Ind. Acad. Sci.*, **21** (1945), 177–210.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

contracted) in one direction, and also consider Fraunhofer diffraction at a screen containing a large number of openings of the same size and shape.

Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two apertures such that the extension of $\mathcal{A}_2$ in a particular direction ($O\xi$) is μ times that of $\mathcal{A}_1$. For Fraunhofer diffraction at $\mathcal{A}_1$, we have

$$U_1(p, q) = C\iint_{\mathcal{A}_1} \mathrm{e}^{-\mathrm{i}k(p\xi+q\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta. \tag{19}$$

Similarly for Fraunhofer diffraction at $\mathcal{A}_2$,

$$U_2(p, q) = C\iint_{\mathcal{A}_2} \mathrm{e}^{-\mathrm{i}k(p\xi+q\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta. \tag{20}$$

If in (20) we change the variables of integration from (ξ, η) to (ξ', η'), where

$$\xi' = \frac{1}{\mu}\xi, \qquad \eta' = \eta, \tag{21}$$

we obtain

$$U_2(p, q) = \mu C\iint_{\mathcal{A}_1} \mathrm{e}^{-\mathrm{i}k(\mu p\xi'+q\eta')}\,\mathrm{d}\xi'\,\mathrm{d}\eta' = \mu U_1(\mu p, q). \tag{22}$$

This shows that *when the aperture is uniformly extended in the ratio* μ:1 *in a particular direction, the Fraunhofer pattern contracts in the same direction in the ratio* 1:μ; *and the intensity in the new pattern is* μ^2 *times the intensity at the corresponding point of the original pattern*. Using this result we may, for example, immediately determine the Fraunhofer pattern of an aperture which has the form of an ellipse or a parallelogram from that of a circle or rectangle respectively. Fig. 8.14 illustrates the case of an elliptical aperture.

We now consider the important case of a screen that contains a large number of identical and similarly oriented apertures. (According to Babinet's principle the results will also apply to the complementary distribution of obstacles.) Let $O_1, O_2, \ldots, O_N$ be a set of similarly situated points, one in each aperture, and let the coordinates of those points referred to a fixed set of axes in the plane of the apertures be (ξ_1, η_1), $(\xi_2, \eta_2), \ldots, (\xi_N, \eta_N)$. The light distribution in the Fraunhofer diffraction pattern is then given by

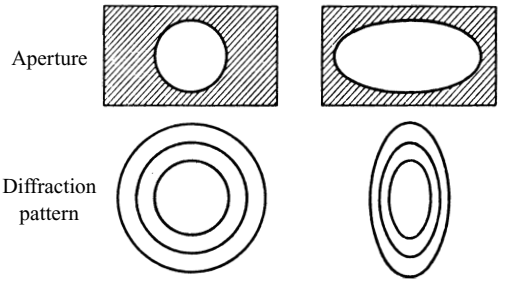


Fig. 8.14 Comparison of Fraunhofer diffraction at a circular and an elliptical aperture.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$U(p, q) = C \sum_n \iint_{\mathcal{A}} \mathrm{e}^{-\mathrm{i}k[(\xi_n+\xi')p+(\eta_n+\eta')q]}\, \mathrm{d}\xi'\, \mathrm{d}\eta'$$

$$= C \sum_n \mathrm{e}^{-\mathrm{i}k[p\xi_n+q\eta_n]} \iint_{\mathcal{A}} \mathrm{e}^{-\mathrm{i}k(p\xi'+q\eta')}\, \mathrm{d}\xi'\, \mathrm{d}\eta', \tag{23}$$

where the integration extends over any one opening $\mathcal{A}$ of the set. The integral expresses the effect of a single aperture, whilst the sum represents the superposition of the coherent diffraction patterns. If $I^{(0)}(p, q)$ is the intensity distribution arising from a single aperture, then, according to (23), the total intensity is given by

$$I(p, q) = I^{(0)}(p, q) \left| \sum_n \mathrm{e}^{-\mathrm{i}k(p\xi_n+q\eta_n)} \right|^2$$

$$= I^{(0)}(p, q) \sum_n \sum_m \mathrm{e}^{-\mathrm{i}k[p(\xi_n-\xi_m)+q(\eta_n-\eta_m)]}. \tag{24}$$

The simplest case, that of two openings, was considered earlier in §7.2, in connection with the theory of interference. However, we neglected there the dependence of $I^{(0)}$ on p and q (i.e. the effect of diffraction at each opening) and only studied the effect of superposition. It is easily seen that the earlier result (§7.2 (17)) is in agreement with (24). For if $N = 2$, (24) reduces to

$$I = I^{(0)}\{2 + \mathrm{e}^{-\mathrm{i}k[p(\xi_1-\xi_2)+q(\eta_1-\eta_2)]} + \mathrm{e}^{-\mathrm{i}k[p(\xi_2-\xi_1)+q(\eta_2-\eta_1)]}\}$$

$$= 4I^{(0)} \cos^2 \tfrac{1}{2}\delta,$$

with

$$\delta = k[p(\xi_2 - \xi_1) + q(\eta_2 - \eta_1)].$$

Let us now consider the effect of a large number of apertures. We shall see that quite different results are obtained, depending on whether the apertures are distributed regularly or irregularly over the screen.

When the apertures are distributed irregularly over the screen, terms with different values of m and n in the double sum will fluctuate rapidly between $+1$ and -1 as m and n take on different values, and in consequence the sum of such terms will have zero mean value. Each remaining term ($m = n$) has the value unity. Hence it follows that except for local fluctuations* the total intensity is N times the intensity of the light diffracted by a single aperture:

$$I(p, q) \sim NI^{(0)}(p, q). \tag{25}$$

Diffraction effects of this type or, more often still, complementary effects (in the sense of Babinet's principle) may be easily observed, for example when a glass plate, dusted with lycopodium powder or covered with other particles of equal size and shape, is

* Fluctuations of a somewhat different type arise when the apertures are not of the same form, but are distributed regularly or according to some statistical law (M. v. Laue, *Berl. Ber.*, (1914), 1144). Similar effects are observed in connection with diffraction of X-rays by liquids (J. A. Prins, *Naturwiss.*, **19** (1931), 435).

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

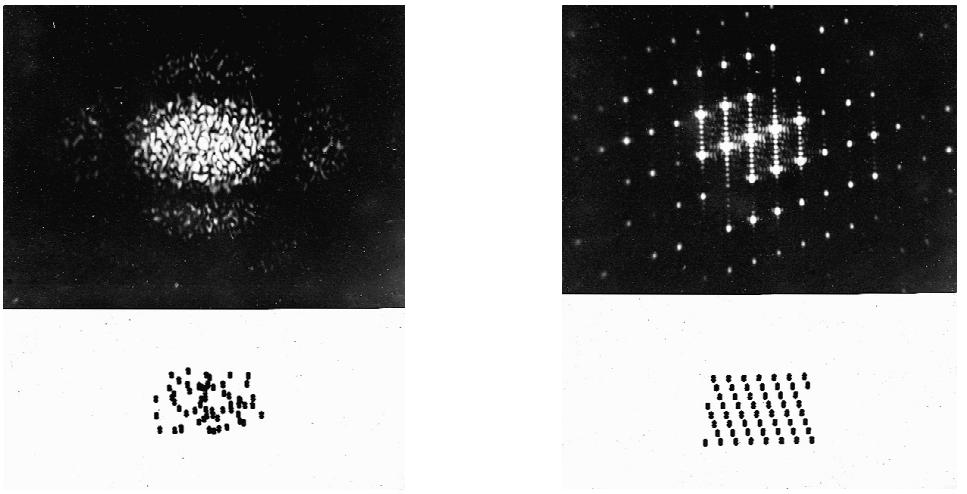


Fig. 8.15 Fraunhofer diffraction pattern from an irregular distribution (a), and a regular distribution (b), of 56 identical and similarly situated apertures in a plane screen. The form and distribution of the apertures is shown in the lower portions of the figures. Light: mercury yellow, $\lambda = 5790$ Å. (Photograph courtesy of H. Lipson, C. A. Taylor, and B. J. Thompson.)

held in the path of light from a distant source. A piece of tin-foil pierced indiscriminately by a pin will also act as a diffracting screen of the type just considered.

The results are quite different when the openings are distributed regularly, for the terms with $m \neq n$ may now give appreciable contributions for certain values of p and q. For example, if the points O_n are so situated that for certain values of p and q the phases of all the terms for which $m \neq n$ are exact multiple of 2π their sum will be equal to $N(N-1)$ and so for large N will be of the order of N^2. This enormous increase in intensity for particular directions, clearly illustrated in Fig. 8.15, is, as we shall see in the next section, of great importance in practice.

8.6 Fraunhofer diffraction in optical instruments

8.6.1 Diffraction gratings

(a) The principle of the diffraction grating

A diffraction grating may be defined as any arrangement which imposes on an incident wave a periodic variation of amplitude or phase, or both. We may characterize any particular grating by its *transmission function*, defined as follows:

Let a transparent or semitransparent object (not necessarily periodic) cover a portion of a fictitious reference plane $\xi\eta$, and let it be illuminated by a plane monochromatic wave incident in a direction specified by the direction cosines l_0, m_0. Fig. 8.16 illustrates the arrangement, the η-axis being perpendicular to the plane of the drawing. If no object were present, the disturbance in the ξ, η-plane would be represented by the

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

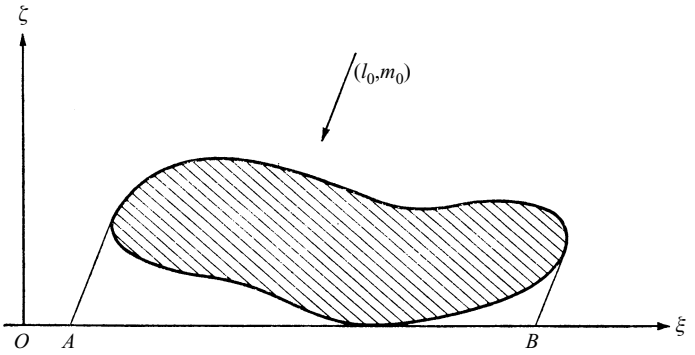


Fig. 8.16 Definition of the transmission function.

function $V_0(\xi, \eta) = A \exp[ik(l_0\xi + m_0\eta)]$, the factor $\exp(-i\omega t)$ being, as usual, omitted. Because of the presence of the object the disturbance will be modified and may be represented by some other function, which we denote by $V(\xi, \eta)$. The *transmission function* of the object is then defined as

$$F(\xi, \eta) = \frac{V(\xi, \eta)}{V_0(\xi, \eta)}. \tag{1}$$

In general F depends, of course, not only on ξ and η but also on the direction (l_0, m_0) of illumination. The transmission function is in general complex, since both the amplitude and the phase of the light may be altered on passing through the object. In the special case when the object alters the amplitude but not the phase of the incident wave (i.e. if $\arg F \equiv 0$), we speak of an *amplitude object*; if it alters the phase but not the amplitude (i.e. $|F| = 1$) we speak of a *phase object*.

If we are concerned with reflected light rather than with light that is transmitted by an object, it is more appropriate to speak of a *reflection function*, defined in a similar way, the only difference being that the reference plane is on the same side of the object as the incident light.

The ratio $|V/V_0|$ is practically unity for points outside the geometrical shadow (whose boundary is represented by points A and B in Fig. 8.16) cast by the object. If the portion outside the shadow region is covered by an opaque screen, the arrangements act as a diffracting aperture $\mathcal{A}$ with a nonuniform pupil function (see §8.3 (39)). If the linear dimensions of $\mathcal{A}$ are large compared to the wavelength and if F remains sensibly constant over regions whose dimensions are of the same order as the wavelength, the diffraction formula §8.3 (23) remains valid under the same conditions as before, provided that the integrand of the diffraction integral is multiplied by F.

Let us now consider a one-dimensional grating consisting of N parallel grooves of arbitrary profile, ruled on one surface of a plane-parallel glass plate. Let the ξ, η-plane coincide with the plane face of the plate, η being the direction of the grooves and let d be the period in the ξ direction (see Fig. 8.17).

Assume that the direction of propagation of the wave incident upon the grating is in the plane of the figure, making an angle θ_0 with $O\zeta$, and let θ denote the angle which $O\zeta$ makes with the line joining a very distant point of observation P with the grating.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

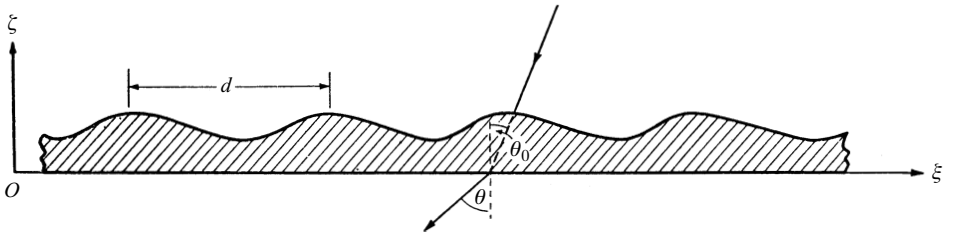


Fig. 8.17 Illustrating the theory of the diffraction grating.

As before we set $l_0 = \sin\theta_0$, $l = \sin\theta$, $p = l - l_0 = \sin\theta - \sin\theta_0$. The complex amplitude at P is then immediately obtained from §8.5 (23), where the integrand must be multiplied by the transmission function F of one periodic element. We may set $q = 0$ and

$$\xi_n = nd, \qquad \eta_n = 0 \qquad (n = 0, 1, \ldots, N-1). \tag{2}$$

We then obtain

$$U(p) = U^{(0)}(p) \sum_{n=0}^{N-1} \mathrm{e}^{-\mathrm{i}kndp} = U^{(0)}(p) \frac{1 - \mathrm{e}^{-\mathrm{i}Nkdp}}{1 - \mathrm{e}^{-\mathrm{i}kdp}}, \tag{3}$$

where*

$$U^{(0)}(p) = C \int_A F(\xi) \mathrm{e}^{-\mathrm{i}kp\xi}\, \mathrm{d}\xi. \tag{4}$$

Hence

$$\begin{aligned} I(p) = |U(p)|^2 &= \frac{(1 - \mathrm{e}^{-\mathrm{i}Nkdp})}{(1 - \mathrm{e}^{-\mathrm{i}kdp})} \cdot \frac{(1 - \mathrm{e}^{\mathrm{i}Nkdp})}{(1 - \mathrm{e}^{\mathrm{i}kdp})} |U^{(0)}(p)|^2 \\ &= \frac{1 - \cos Nkdp}{1 - \cos kdp} I^{(0)}(p), \end{aligned} \tag{5}$$

where $I^{(0)}(p) = |U^{(0)}(p)|^2$. If we introduce the function

$$H(N, x) = \left(\frac{\sin Nx}{\sin x}\right)^2, \tag{6}$$

the formula (5) for the intensity may be written as

$$I(p) = H\left(N, \frac{kdp}{2}\right) I^{(0)}(p). \tag{5a}$$

Before discussing the implications of this basic formula we note that according to (3) the light distribution is the same as that due to a set of coherent secondary sources each characterized by the same amplitude function $|U^{(0)}(p)|$ and with phases that differ from each other by integral multiples of $k\,d\,p$. To see the significance of this phase difference consider two corresponding points A and B in neighbouring grooves

* Since F depends on l_0, the quantities $U^{(0)}$ and $I^{(0)}$ now depend on both l and l_0 and not on the difference $l - l_0$ only. As we are only interested in effects for a fixed direction of incidence, we may regard l_0 as a constant and retain the previous notation.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

of the grating (Fig. 8.18). Since the effect of the grating is to impress a periodic variation onto the incident wave, it follows that the path difference between the light arriving at A and at B is the same as in the absence of the grating, i.e. it is equal to $AK = d\sin\theta_0$, K denoting the foot of the perpendicular from B on to the ray incident at A. Further, the light path from B in the direction θ exceeds the light path from A by $BL = d\sin\theta$, L being the foot of the perpendicular from A on to the ray diffracted at B in the direction θ. Hence the total path difference between light arriving at the distant point of observation from corresponding points in two neighbouring grooves is

$$BL - AK = d(\sin\theta - \sin\theta_0) = dp, \tag{7}$$

and the corresponding phase difference is $2\pi dp/\lambda = kdp$.

Formula (5a) expresses $I(p)$ as the product of two functions: one of them, $I^{(0)}$, represents the effect of a single period of the grating; the other, H, represents the effect of interference of light from different periods. The function $H(N, x)$ has maxima, each of height N^2, at all points where the denominator $\sin^2 x$ vanishes, i.e. where x is zero or an integral multiple of π. Hence $H(N,\ kdp/2)$ has maxima of height N^2 when

$$p \equiv \sin\theta - \sin\theta_0 = \frac{m\lambda}{d} \qquad (m = 0, \pm1, \pm2, \ldots). \tag{8}$$

The integer m represents, according to (7), the path difference in wavelengths between light diffracted in the direction of the maximum, from corresponding points in two neighbouring grooves. In agreement with our earlier definition (§7.3.1), we call m the *order of interference*. Between these principal maxima there are weak secondary maxima (see Fig. 8.19(a)), the first secondary maximum being only a few per cent of the principal maximum when N is large. The maxima are separated by points of zero intensity at $x = kdp/2 = \pm n\pi/N$, i.e. in directions given by

$$p \equiv \sin\theta - \sin\theta_0 = \frac{n\lambda}{Nd} \qquad (n = \pm1, \pm2, \ldots), \tag{9}$$

the case where n/N is an integer being excluded.

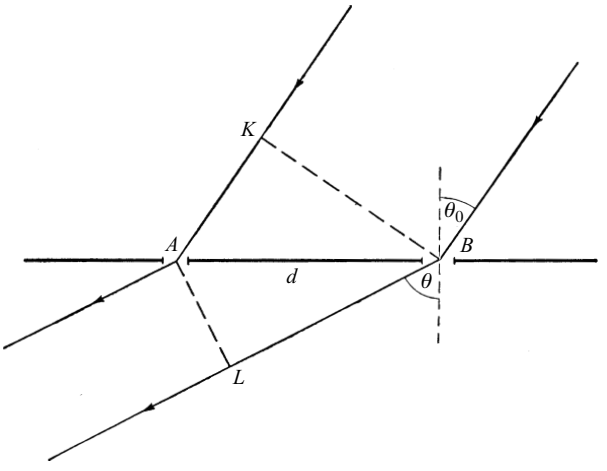


Fig. 8.18 Illustrating the theory of the diffraction grating.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

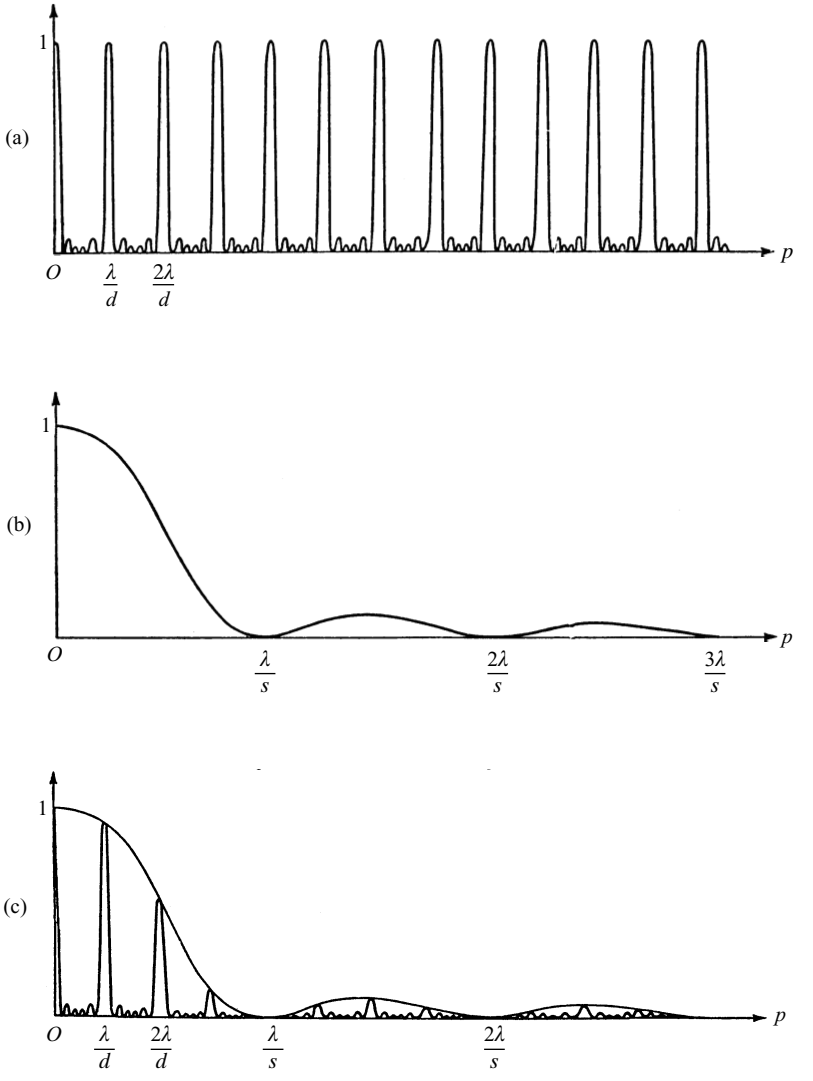


Fig. 8.19 (a) The normalized interference function

$$\frac{1}{N^2} H(N,\, kdp/2) = \left[\frac{\sin(Nkdp/2)}{N \sin(kdp/2)}\right]^2.$$

(b) The normalized intensity function of a slit

$$I^{(0)}(p) = \left[\frac{\sin ksp/2}{ksp/2}\right]^2.$$

(c) The normalized intensity function of a grating consisting of N similar equidistant parallel slits

$$\frac{1}{N^2} I(p) = \left[\frac{\sin(Nkdp/2)}{N \sin(kdp/2)}\right]^2 \left[\frac{\sin ksp/2}{ksp/2}\right]^2.$$

Only the range $p \geqslant 0$ is shown, all the curves being symmetrical about the vertical axis $p = 0$.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

The function $I^{(0)}(p)$ depends on the form of the grooves. Suppose that it has a principal maximum for some direction $p = p'$ and that on both sides of the maximum it falls off slowly in comparison with H. Then $I(p)$ will have the general form of the interference function H, but will be 'modulated' by $I^{(0)}$. Thus $I(p)$ will still have fairly sharp maxima near the directions $p = m\lambda/d$. Since these directions (except for $m = 0$) depend on the wavelength, we see that the grating will decompose a beam of nonmonochromatic light into *spectral orders*.

To illustrate these remarks let us consider a grating consisting of a succession of long equidistant slits (Fig. 8.20), each of width s and length L, in an opaque screen. If the grating is illuminated from a very distant line source parallel to the slits, the intensity $I^{(0)}$ is given by the expression §8.5 (4) (with $2a = s$, $2b = L$) and we obtain

$$I(p) = \frac{sE}{\lambda R^2}\left(\frac{\sin\frac{Nkdp}{2}}{\sin\frac{kdp}{2}}\right)^2\left(\frac{\sin\frac{ksp}{2}}{\frac{ksp}{2}}\right)^2. \tag{10}$$

Curves representing the two factors in (10) and their product are shown in Fig. 8.19. The last factor in (10), which represents the effect of a single slit, has a principal maximum at $p = 0$ and minima given by $ksp/2 = n\pi$, i.e. at

$$p = \frac{n\lambda}{s}, \qquad (n = \pm 1, \pm 2, \ldots) \tag{11}$$

separated by weak secondary maxima. We see that if $\lambda/s \gg \lambda/d$, i.e. if the width of each slit is small compared to d, the intensity $I(p)$ has in addition to a principal maximum at $p = 0$ a series of sharp, but progressively decreasing, maxima on either side of it, near directions given by (8).

Returning to the general case, it is evident that if the width of each groove is very small, of the order of a wavelength (as is often the case in practice) the formula (4), derived on the basis of Kirchhoff's approximation, can evidently no longer be expected to hold. In such cases more refined considerations must be made to determine the detailed distribution of the intensity. We may, however, expect that the main qualitative features indicated by our elementary theory, namely the existence of sharp maxima whose positions are substantially determined by the interference function H, remain even when the grooves are very narrow, provided, of course, that the intensity function of a single period varies slowly in an interval of the order $\Delta p = \lambda/d$.

Let us now consider the resolution that may be attained with a grating. The separation between a primary maximum of order m and a neighbouring minimum is, according to (9), given by

$$\Delta p = \frac{\lambda}{Nd}. \tag{12}$$

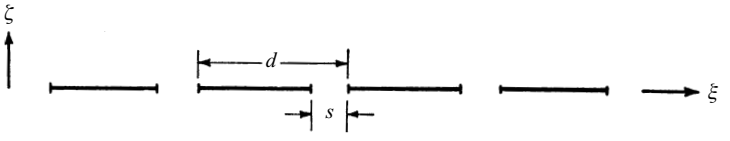


Fig. 8.20 Profile of a simple line grating.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

If the wavelength is changed by an amount $\Delta\lambda$, the mth-order maximum is, according to (8), displaced by an amount

$$\Delta' p = \frac{|m|}{d}\Delta\lambda. \tag{13}$$

Assuming that the lines of wavelength $\lambda \pm \frac{1}{2}\Delta\lambda$ will just be resolved when the maximum of the one wavelength coincides with the first minimum of the other (see p. 371) we have on the limit of resolution in the mth order, $\Delta p = \Delta' p$, i.e.

$$\frac{\lambda}{\Delta\lambda} = |m|N. \tag{14}$$

Thus, *the resolving power is equal to the product of the order number m and the number N of the grooves*. For the mth order we have, according to (8), that $d(\sin\theta - \sin\theta_0) = m\lambda$, so that we may also express the resolving power in the form

$$\frac{\lambda}{\Delta\lambda} = \frac{Nd|\sin\theta - \sin\theta_0|}{\lambda}. \tag{14a}$$

Because of (7) this implies that *the resolving power is equal to the number of wavelengths in the path difference between rays that are diffracted in the direction θ from the two extreme ends* (*separated by distance Nd*) *of the grating*. It is to be noted that since $|\sin\theta - \sin\theta_0|$ cannot exceed 2, the resolving power that can be attained with a grating of overall width w can never exceed the value $2w/\lambda$.

Let us illustrate the formula (14) by determining the number of grooves that a grating must have, in order to separate two lines which are a tenth of an ångström unit apart, near the centre of the visible region of the spectrum. In this case $\lambda \sim 5500$ Å, $\Delta\lambda = 10^{-1}$ Å, and if we observe in the second order ($m = 2$), we must have, according to (14) $N \geqslant 5.5 \times 10^3/2 \times 10^{-1} = 27{,}500$, i.e. the grating must have at least 27,500 grooves.

For comparison let us consider the resolving power of a prism, in the position of minimum deviation, with a line source that is parallel to the edge A of the prism (slit of the spectrograph). A pencil of parallel rays will be incident upon the prism and will be diffracted at a rectangle of width $l_1 = l_2$ (see Fig. 4.28). According to §8.5 (2), the first minimum of the intensity is at an angular distance (assumed to be small)

$$p = \frac{\lambda}{l_1} \tag{15}$$

from the geometrical image of the slit. The change in the angular dispersion corresponding to a change of wavelength by amount $\Delta\lambda$ is, according to §4.7 (36),

$$\Delta\varepsilon = \frac{t}{l_1}\frac{\mathrm{d}n}{\mathrm{d}\lambda}\Delta\lambda, \tag{16}$$

where t is the greatest thickness of the glass through which one of the extreme rays has passed, and n is the refractive index of the material of the prism. Since at the limit of resolution $p \sim \Delta\varepsilon$ the resolving power is given by

$$\frac{\lambda}{\Delta\lambda} = t\left|\frac{\mathrm{d}n}{\mathrm{d}\lambda}\right|. \tag{17}$$

Eq. (17) shows that, *with a given glass, the resolving power of a prism depends only*

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

on the greatest thickness of the glass traversed by the rays; in particular the resolving power is independent of the angle of the prism. As an example, suppose that the length of the base of the prism is equal to 5 cm and that it is made of heavy flint glass, for which $\mathrm{d}n/\mathrm{d}\lambda \sim 1000\ \mathrm{cm}^{-1}$ at wavelength $\lambda = 5500$ Å. If the prism is completely filled with light, then according to (17) it will resolve lines near the centre of the visible region which are not less than $\Delta\lambda$ apart, where $\Delta\lambda \sim 5.5 \times 10^{-5}\ \mathrm{cm}/5 \times 10^{3} = 1.1$ Å. Thus a prism of this considerable size has a resolving power 10 times smaller than the grating of 27,500 grooves discussed before.

We have so far considered one-dimensional gratings only, but the analysis may easily be extended to two- and three-dimensional periodic arrangements of diffracting bodies. Two-dimensional gratings (sometimes called cross gratings) find no practical applications, though their effects can often be observed, for example when looking at a bright source through a finely woven material (e.g. a handkerchief). The theory of three-dimensional gratings (sometimes called space gratings) is, on the other hand, of great importance, such gratings being formed by a regular arrangement of atoms in a crystal. The lattice distances (distances between neighbouring atoms) are of the order of an ångström unit (10^{-8} cm), this being also the order of magnitude of the wavelengths of X-rays. Hence, by sending a beam of X-rays through a crystal, diffraction patterns are produced, and from their analysis information about the structure of the crystal may be deduced. We will briefly discuss this subject in §13.1.3.

Another example of a grating-like structure is presented by ultra-sonic waves in liquids. These are elastic waves produced by a piezo-electric oscillator, differing from ordinary sound waves only in having a frequency well above the upper limit of audibility. Such waves give rise to rarefactions and condensations in the liquid which then act on the incident light like a grating. The theory of this phenomenon is discussed in Chapter XII. For the rest of this section we shall restrict our attention to one-dimensional gratings as used in spectroscopic work.

(b) Types of grating[*]

The principle of the diffraction grating was discovered by Rittenhouse in 1785,† but this discovery attracted practically no attention. The principle was re-discovered by Fraunhofer‡ in 1819. Fraunhofer's first gratings were made by winding very fine wire round two parallel screws. Because of the relative ease with which wire gratings may be constructed these are occasionally used even today, particularly in the long-wavelength (infra-red) range. Later Fraunhofer made gratings with the help of a machine, by ruling through gold films deposited on a glass plate; also, using a diamond as a ruling point, he ruled the grooves directly onto the surface of glass.

Great advances in the technique of production of gratings were made by Rowland§ who constructed several excellent ruling machines and also invented the so-called

[*] For a fuller account of methods of production of gratings and their development, see G. R. Harrison, *J. Opt. Soc. Amer.*, **39** (1949), 413.

† D. Rittenhouse, *Trans. Amer. Phil. Soc.*, **2** (1786), 201. See also the article by T. D. Cope in *Journ. Franklin Inst.*, **214** (1932), 99.

‡ J. Fraunhofer, *Denkschr. Akad. Wiss. München*, **8** (1821–1822), 1. *Ann. d. Physik*, **74** (1823), 337. Reprinted in his collected works (Munich, 1888), pp. 51, 117.

§ H. A. Rowland, *Phil. Mag.* (5), **13** (1882), 469. *Nature*, **26** (1882), 211. *Phil. Mag.*, **16** (1883), 297.

concave grating (discussed on pp. 459–461). Rowland's machine was able to rule gratings with grooves over 4 in long over a length of 6 in, and his first machine ruled about 14,000 grooves per inch, giving a resolving power in excess of 150,000. Later Michelson ruled gratings considerably wider than 6 in, with a resolving power approaching 400,000.

Most of the early gratings were ruled on speculum metal and glass, but the more recent practice is to rule the grooves on evaporated layers of aluminium. Since aluminium is a soft metal it causes less wear on the ruling point (diamond) and it also reflects better in the ultra-violet.

A perfect grating would have all the grooves strictly parallel and of identical form, but in practice errors will naturally occur. Quite irregular errors lead to a blurring of the spectrum and are not so serious as systematic errors, such as periodic errors of spacing. These errors give rise to spurious lines in the spectrum, known as *ghosts*. Often they can be distinguished from true lines only with difficulty.

High resolving power is not always the only important requirement in spectroscopic applications. When little energy is available, as for example in the study of spectra of faint stars or nebulae, or for work in the infra-red region of the spectrum, it is essential that as much light as possible should be diffracted into one particular order. Moreover, for precise wavelength measurements, a grating that gives high dispersion must be used. According to (8), the angular dispersion (with a fixed angle of incidence) is given by

$$\frac{\mathrm{d}\theta}{\mathrm{d}\lambda} = \frac{1}{\cos\theta}\frac{m}{d}, \tag{18}$$

so that to obtain high dispersion the spacing d should be small or the observations must be made in high orders (m large). If, however, the grating is formed by a succession of opaque and transparent (or reflecting) strips, only a small fraction of the incident light is thrown into any one order. This drawback is overcome in modern practice by ruling the grooves to controlled shape. With a grating which consists of grooves of the form shown in Fig. 8.21, most of the light may be directed into one or two orders on one side of the central image. Gratings of this type, with fairly coarse grooves, are called *echelette gratings*, because they may be regarded as being intermediate between the older types of grating and the so-called echelon gratings which will be described later. Echelette gratings were first ruled by Wood* on copper

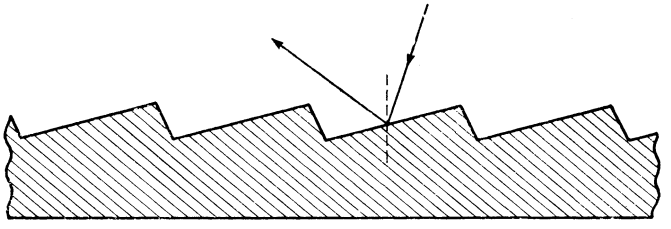


Fig. 8.21 Reflection grating with controlled groove form.

* R. W. Wood, *Phil. Mag.*, **20** (1910), 770; *Ibid.* **23** (1912), 310; A. Trowbridge and R. W. Wood, *ibid.*, **20** (1910), 886, 898.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

plates, using the natural edge of a selected carborundum crystal as a ruling point. Later they were ruled with diamond edges ground to the desired shape. They had 2000–3000 grooves per inch, and when used with visible light they sent the greater part of the light into a group of spectra near the 15th or 30th order. Echelette gratings have considerable value in infra-red spectroscopy.

More recently methods have been developed for controlling the groove shape for gratings with much smaller groove spacing.* These *blazed gratings*, as they are called, have grooves of similar form as the echelettes, but form the most intense spectra in much lower orders (usually the first or second).

It appears that the resolving power of gratings of the type described is limited by practical considerations of manufacture to about 400,000. For some applications (e.g. for the study of Zeeman effects and the hyperfine and isotope structure patterns), a resolving power that exceeds this value is required. For the attainment of such a high resolving power, Harrison† proposed the so-called *echelle grating*, which has wide, shallow grooves and is designed for use at an angle of incidence greater than 45°, the direction of incidence being normal to the narrow side of the step. These gratings operate with relatively high orders ($m \sim 1000$). A 10 in echelle with 100 grooves per inch, designed for observation in the 1000th order, has a resolving power of 1,000,000‡.

Because a grating of good quality is very difficult to produce, *replicas* of original rulings are often used.§ These are obtained by moulding from an original ruled master grating.

Finally we must mention a 'grating' of an entirely different construction, the *echelon* invented by Michelson.‖ It consists of a series of strictly similar plane-parallel glass plates arranged in the form of a flight of steps (hence the name), as shown in Fig. 8.22. Each step retards the beam of light which passes through it by the same amount with respect to its neighbour. Because the breadth of each step is large compared to the wavelength, the effect of diffraction is confined to small angles, so that most of the light is concentrated in one or two spectra near the direction $\theta = 0$, and these correspond to very high orders, since the retardation introduced between successive beams is very many wavelengths.

The resolving power of the echelon depends not only on the path difference between the rays from the extreme ends of the grating but also (though to a much lesser extent) on the dispersion of the glass. If n is the refractive index, t the thickness of each step, and d its breadth (see Fig. 8.22), the path difference between rays diffracted from neighbouring steps is evidently $pd + (n-1)t$, it being assumed that p is small. Hence the positions of the principal maxima are given by

$$pd + (n-1)t = m\lambda, \qquad (m = 0, 1, 2, \ldots). \tag{19}$$

* R. W. Wood, *Nature*, **140** (1937), 723; *J. Opt. Soc. Amer.*, **34** (1944), 509; H. Babcock, *ibid.*, **34** (1944), 1.
† G. R. Harrison, *J. Opt. Soc. Amer.*, **39** (1949), 522.
‡ For a review of the theory and production of high-resolution gratings, see G. W. Stroke, *Progress in Optics*, Vol. 2, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1963), p. 1.
§ First replicas were made by T. Thorp, British Patent No. 11,460 (1899); and later by R. J. Wallace, *Astrophys. J.*, **22** (1905), 123; *ibid.*, **23** (1906), 96. Improved methods have been described by T. Merton, *Proc. Roy. Soc. A*, **201** (1950), 187.
‖ A. A. Michelson, *Astrophys. J.*, **8** (1898), 37; *Proc. Amer. Acad. Arts Sci.*, **35** (1899), 109.

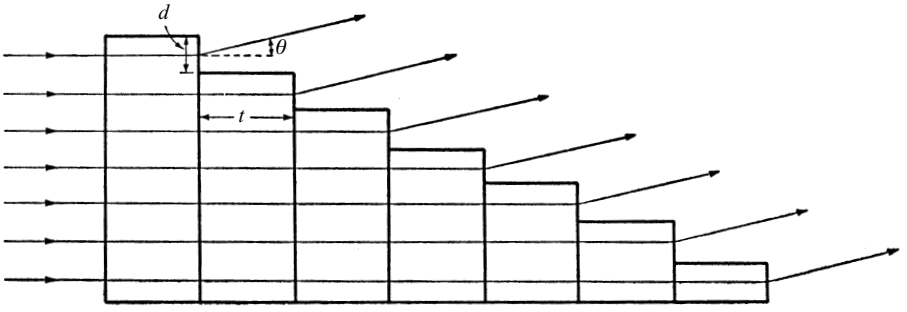


Fig. 8.22 Michelson's echelon.

If the wavelength is changed by an amount $\Delta\lambda$, the mth-order maximum is displaced by

$$\Delta' p = \left| m - t\frac{\mathrm{d}n}{\mathrm{d}\lambda} \right| \frac{\Delta\lambda}{d}. \tag{20}$$

The separation Δp between a principal maximum of order m and a neighbouring minimum is again given by (12), so that the condition $\Delta p = \Delta' p$ for the limit of resolution gives

$$\frac{\lambda}{\Delta\lambda} = N\left| m - \frac{\mathrm{d}n}{\mathrm{d}\lambda} t \right|. \tag{21}$$

Here we may substitute for m the value $(n-1)t/\lambda$ obtained from (19) by neglecting the term pd, for the p values for which the intensity is appreciable are of the order of λ/d, i.e. pd is of the order of a wavelength, whilst $(n-1)t$ is of the order of many thousand wavelengths. We thus obtain the following expression for the resolving power of the echelon:

$$\frac{\lambda}{\Delta\lambda} \sim N\left| \frac{n-1}{\lambda} - \frac{\mathrm{d}n}{\mathrm{d}\lambda} \right| t. \tag{22}$$

The ratio $(\mathrm{d}n/\mathrm{d}\lambda)/[(n-1)/\lambda]$ is small. For flint glass near the centre of the visible region it has a value near -0.05 to -0.1. Hence, under these circumstances, the resolving power of an echelon exceeds, by about 5–10 per cent, the resolving power of a line grating with N grooves, when observation is made in the order $m = (n-1)t/\lambda$. One of Michelson's echelons consisted of twenty plates, each having a thickness $t = 18$ mm, and the breadth d of each step was about 1 mm. Taking $n = 1.5$, the retardation between two successive beams measured in wavelengths of green light $\lambda = 5 \times 10^{-5}$ cm was $m \sim 0.5 \times 1.8/5 \times 10^{-5} \sim 20{,}000$. Assuming $(\mathrm{d}n/\mathrm{d}\lambda)/[(n-1)/\lambda] = -0.1$, this gives a resolving power of about 20 $(20{,}000 + 0.1 \times 20{,}000) = 440{,}000$.

More important is the *reflection echelon*. Here each step is made highly reflecting by means of metallic coating, and the spectra formed by reflected light are observed. With a reflection echelon the resolving power is 3–4 times larger than with a transmission echelon of corresponding dimensions, since each step introduces a retardation between successive beams of amount $2t$ instead of $(n-1)t \sim t/2$. Like

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

the echelle grating the reflection echelon is capable of giving resolving power of over one million. Another advantage of the reflection echelon over the transmission echelon is that it may be used in the ultra-violet region of the spectrum, where glass absorbs. Although Michelson realized that advantages would be gained by using the instrument with reflected rather than transmitted light, technical difficulties prevented the production of a satisfactory reflection echelon for nearly thirty years until they were overcome by Williams.* Because of difficulties in assembling a large number of plates of equal thickness within the narrow permissible tolerance, the number of steps is limited in practice to about forty.

Finally, a few remarks must be made about overlapping of orders. Restricting ourselves to the visible region, i.e. considering wavelengths in the range $\lambda_1 = 0.4\ \mu$m to $\lambda_2 = 0.75\ \mu$m, we see that the first-order spectrum does not quite reach to the spectrum of the second order: for the first-order spectrum covers the range from $p = \lambda_1/d$ to $p = \lambda_2/d = 0.75\lambda_1/0.4d = 1.8\lambda_1/d$, whilst the second order begins at $p = 2\lambda_1/d$. On the other hand the spectrum of the second order extends across a part of the third-order spectrum, namely from $p = 2\lambda_1/d$ to $p = 2\lambda_2/d$, whilst the third order begins already at $2 \times 1.8\lambda_1/d$. As the order increases the successive spectra overlap more and more (see Fig. 8.23). If the lines of wavelength λ and $\lambda + \delta\lambda$ coincide in two successive orders $(m + 1)$th and mth, then

$$(m + 1)\lambda = m(\lambda + \delta\lambda),$$

i.e.

$$\frac{\delta\lambda}{\lambda} = \frac{1}{m}. \tag{23}$$

Thus the '*free spectral range*' *is inversely proportional to the order.*

The overlapping of orders was formerly used to compare wavelengths, in the so-called method of coincidences (see p. 379); this method has been superseded by simple interpolation between standard wavelengths determined interferometrically.

In conclusion let us summarize the main distinguishing features of the different types of gratings. We recall that, according to (14), high resolving power may be

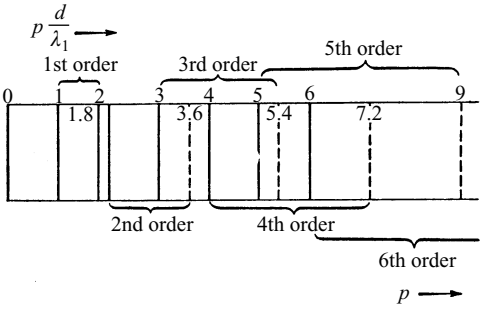


Fig. 8.23 The overlapping of grating spectra.

* W. E. Williams, British Patent No. 312,534 (1926); *Proc. Opt. Conv.*, **2** (1926), 982; *Proc. Phys. Soc.*, **45** (1933), 699.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

attained with either a large number of periods and relatively low orders, or with a moderate number of periods and large orders. Ordinary ruled gratings represent the low-order extreme ($m \sim 1$ to 5), whilst the echelons represent the extreme of high orders ($m \sim 20{,}000$). In between are the echellettes ($m \sim 15$ to 30) and the echelles ($m \sim 1000$). For particular applications one must bear in mind that the angular dispersion is directly proportional to the spectral order and inversely proportional to the period whilst the free spectral range is inversely proportional to the order.

(c) Grating spectrographs

In a grating spectrograph coloured images of a slit source are produced in the various orders into which the grating separates the incident light. A simple arrangement is shown in Fig. 8.24. Collimated light from a slit source S in the focal plane of a lens L is incident on a reflection grating G, and the images of S formed by the diffracted rays are observed in the focal plane F of a telescope T. A modification of this arrangement, known as *Littrow's mounting*, which has the advantage of compactness, is shown in Fig. 8.25. This is an autocollimation device, which needs only one lens. The slit is just

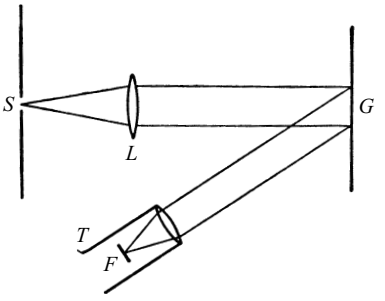


Fig. 8.24 A grating spectrograph.

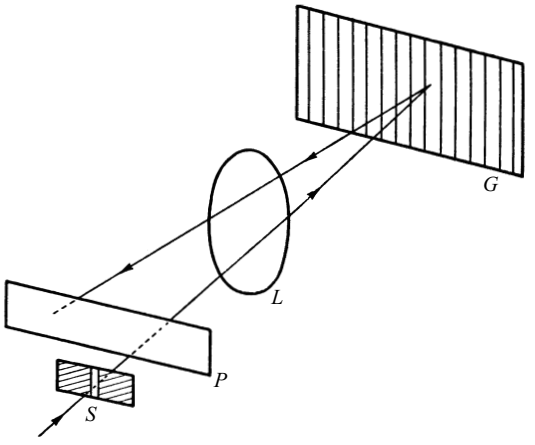


Fig. 8.25 A grating spectrograph: Littrow's mounting.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

below the plate P and the lens is near the grating, which can be turned through a prescribed angle with respect to the direction of the incident beam.

In order to avoid losses of light which necessarily arise when the diffracted rays are focused by means of lenses, Rowland introduced the *concave grating*. Here the grooves are ruled on a concave highly reflecting metal surface, i.e. on a concave mirror, in such a way that their projections on a chord of the mirror surface are equidistant. A simple geometrical theorem indicates the possible positions of the slit and the plane of observation relative to the grating:

Let Q be the midpoint of the surface of the grating and C its centre of curvature, and describe a circle K with centre at the midpoint O of QC and with radius $r = OQ = OC$ (Fig. 8.26). We shall prove that light from any point S of the circle K will be approximately reflected to a point P and diffracted to points P', P'', ... on the circle, each of these points being a focus for diffracted rays of a particular order. To show this, construct the reflected ray QP corresponding to the incident ray SQ. If $\alpha = \angle SQC$ is the angle of incidence, the angle $\angle CQP$ of reflection is also equal to α and, moreover, the arc SC is equal to the arc CP. Consider now another ray from S incident upon the grating at a different point R. If the diameter of the circle is sufficiently large (in practice it is usually several feet), then no appreciable error is introduced by assuming R to lie on the circle K. Hence, since C is the centre of curvature of the grating, the angle of incidence $\angle SRC$ and consequently also the angle of reflection are again equal to α. Moreover, since the arc CP is equal to the arc SC it follows that the ray reflected at R again passes through P.

Similar considerations apply to the diffracted rays. Let β be the angle which a ray diffracted at Q makes with QP. The corresponding ray of the same order, diffracted at R, will make the same angle (β) with RP. Hence the ray diffracted at Q makes the same angle with SQ as the ray diffracted at R makes with SR, namely $2\alpha + \beta$. The two diffracted rays, therefore, meet in a point P' of the circle K. Thus, *to obtain sharp lines, the slit, the grating and the plane of observation (photographic plate) should be situated on a circle, whose diameter is equal to the radius of curvature of the concave grating.*

There are several mountings based on this principle. Rowland himself used the arrangement shown in Fig. 8.27. Here the grating G and the plate holder P are fixed to opposite ends of a movable girder, whose length is equal to the radius of curvature of

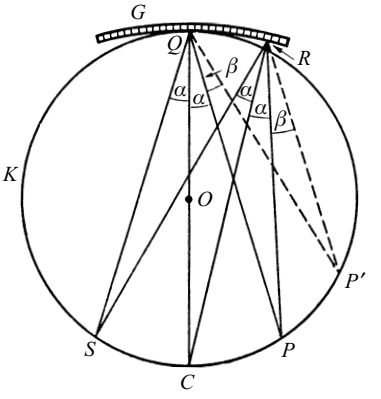


Fig. 8.26 Focusing with a concave grating (Rowland's circle).

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

the grating. The two ends of the girder are free to move along fixed tracks which are at right angles to each other. The slit S is mounted immediately above their intersection in such a way that light falling normally on the slit proceeds along SG. The slit is thus situated on a Rowland circle with diameter PG and the order of the spectrum appearing on the plate depends on the position of the girder.

A different arrangement, shown in Fig. 8.28, avoids the use of mobile parts. Here a circular steel rail to which the slit S and the grating G are permanently attached plays the part of Rowland's circle. Round the rail a series of plate holders P_0, P_1, P_{-1}, ... is set up, so that spectra of several orders can be photographed simultaneously. This arrangement, called *Paschen's mounting*, also has the advantage of great stability.

Another arrangement due to Eagle has, like the Littrow mounting for a plane grating, the advantage of compactness. Here the slit is immediately above or below the centre of the plate holder (Fig. 8.29), or it may be mounted at the side and the light

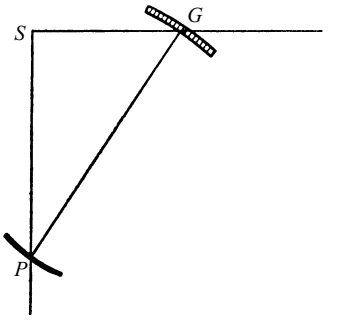


Fig. 8.27 Rowland's mounting for a concave grating.

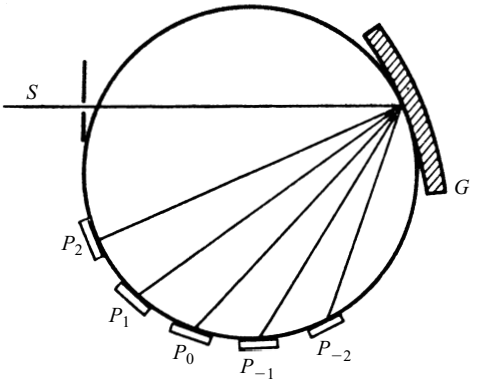


Fig. 8.28 Paschen's mounting for a concave grating.

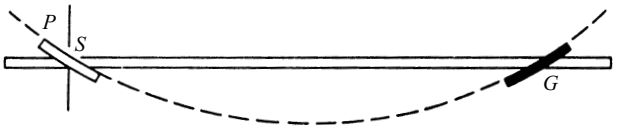


Fig. 8.29 Eagle's mounting for a concave grating.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

may be thrown in the required direction by means of a small reflecting surface. To observe different portions of the spectrum, both the plate and the grating have to be rotated through the same amounts and in opposite senses, and their separation has to be changed, so that they are always tangential to the Rowland circle. One observes that part of the spectrum which is diffracted back at angles nearly equal to the angle of incidence. For this mounting to be strictly autocollimating, the slit S should be at the centre of the plate.

Spectral lines obtained with a concave grating show the same aberrations as images obtained with a concave mirror, chiefly astigmatism. If, however, the concave grating is used in parallel light, the astigmatism may be made to vanish on the grating normal and to be very small over the whole usable spectrum.*

8.6.2 Resolving power of image-forming systems

The Fraunhofer diffraction formula §8.3 (36) finds important applications in the calculation of the *resolving power* of optical systems. We have already introduced the concept of resolving power in connection with interference spectroscopes in §7.6.3, and in the preceding section we have estimated the resolving power that can be attained with gratings and prisms. We shall now extend this concept to image-forming systems.

In a spectral apparatus (e.g. a line grating or the Fabry–Perot interferometer), the resolving power is a measure of the ability of the instrument to separate two neighbouring spectral lines of slightly different wavelengths. In an image-forming system, it is a measure of the ability to separate images of two neighbouring object points. In the absence of aberrations each point object would, according to geometrical optics, give rise to a sharp point image. Because of diffraction the actual image will nevertheless always be a finite path of light. And if two such image patches (diffraction patterns) overlap, it will be more and more difficult to detect the presence of two objects, the closer the central intensity maxima are to each other. The limit down to which the eye can detect the two objects is, of course, to some extent a matter of practical experience. With a photographic plate the contrast may be enhanced and so the limit of resolution decreased by suitable development. Nevertheless it is desirable to have some simple criterion which permits a rough comparison of the relative efficiency of different systems, and for this purpose Rayleigh's criterion discussed in §7.6.3 may again be employed. According to this criterion two images are regarded as just resolved when the principal maximum of one coincides with the first minimum of the other. For a spectral apparatus, where the limit of resolution is a certain wavelength difference $\Delta\lambda$, the resolving power is defined as the quantity $\lambda/\Delta\lambda$. For an image-forming system the limit of resolution is some distance δx or angle $\delta\theta$ and the resolving power is defined as the reciprocal (i.e. $1/\delta x$ or $1/\delta\theta$) of this quantity.

Let us consider first the limit of resolution of a telescope. For a distant object, the edge of the entrance pupil coincides with the circular boundary of the objective, and

* Such an astigmatic-free mounting was described by F. L. O. Wadsworth, *Astrophys. J.*, **3** (1896), 54. For a discussion of the aberration theory of gratings and grating mountings see the article by W. T. Welford in *Progress in Optics*, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 241.

acts as the diffracting aperture. If a is the radius of the objective aperture then, according to §8.5 (16), the position of the first minimum of intensity referred to the central maximum is given by*

$$w = 0.61\frac{\lambda}{a}. \tag{24}$$

Now $w = \sqrt{p^2 + q^2}$ represents the sine of the angle ϕ which the direction (p, q) makes with the central direction $p = q = 0$. This angle is usually so small that its sine may be replaced by the angle itself, and it then follows that (on the basis of Rayleigh's criterion) *the angular separation of two stars that can just be resolved is* $0.61\lambda/a$.

With a given objective, the angular size of the image as seen by the eye depends on the magnification of the eyepiece. It is impossible, however, to bring out detail not present in the primary image by increasing the power of the eyepiece, for each element of the primary image is a small diffraction pattern, and the actual image, as seen by the eyepiece, is only the ensemble of the magnified images of these patterns.

Consider the well-known large telescope at Mount Palomar which has a diameter $2a \sim 5$ m. Neglecting for the moment the effect of the central obstruction in the telescope, the theoretical limit of resolution for light near the centre of the visible range ($\lambda \sim 5.6 \times 10^{-5}$ cm) is seen to be

$$\phi \sim 0.61\frac{5.6 \times 10^{-5}\ \text{cm}}{2.5 \times 10^{2}\ \text{cm}} \sim 1.4 \times 10^{-7}$$

or, in seconds of arc,

$$\phi \sim 0.028''.$$

In §6.1 we quoted the value of 1 minute of arc for the limit of resolution of the eye. We can now give a more precise estimate. Since the diameter of the pupil of the eye varies from about 1.5 mm to about 6 mm (depending on the intensity of the light), it follows that the limit of resolution lies in the range (again taking $\lambda = 5.6 \times 10^{-5}$ cm)

$$0.61\frac{5.6 \times 10^{-5}}{0.75 \times 10^{-1}} > \phi > 0.61\frac{5.6 \times 10^{-5}}{3 \times 10^{-1}},$$

i.e.

$$4.55 \times 10^{-4} > \phi > 1.14 \times 10^{-4}$$

or, in minutes and seconds,

$$1'34'' > \phi > 0'24''.$$

So far we have assumed the aperture to be circular. Of considerable interest is also the case of an annular aperture, since in many telescopes, for example, the central portion of the circular aperture is obstructed by the presence of a secondary mirror. Suppose that the annular aperture is bounded by two concentric circles of radii a and εa, where ε is some positive number less than unity. The light distribution in the Fraunhofer pattern is then represented by an integral of the form §8.5 (8), but with the

* According to §7.6.3 the saddle-to-peak intensity ratio at the limit of resolution for diffraction at a slit aperture is $8/\pi^2 = 0.811$. The corresponding value for the present case (circular aperture) is 0.735.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

ρ integration extending only over the domain $\varepsilon a \leqslant \rho \leqslant a$. In place of §8.5 (13) we then obtain

$$U(P) = C\pi a^2\left[\frac{2J_1(kaw)}{kaw}\right] - C\pi\varepsilon^2 a^2\left[\frac{2J_1(k\varepsilon aw)}{k\varepsilon aw}\right], \tag{25}$$

so that the intensity is given by

$$I(P) = \frac{1}{(1-\varepsilon^2)^2}\left[\left(\frac{2J_1(kaw)}{kaw}\right) - \varepsilon^2\left(\frac{2J_1(k\varepsilon aw)}{k\varepsilon aw}\right)\right]^2 I_0, \tag{26}$$

where $I_0 = |C|^2\pi^2 a^4(1-\varepsilon^2)^2$ is the intensity at the centre $w = 0$ of the pattern. The positions of the minima (zeros) of intensity are now given by the roots of the equation

$$J_1(kaw) - \varepsilon J_1(k\varepsilon aw) = 0 \qquad (w \neq 0), \tag{27}$$

whilst the maxima are given by the roots of

$$J_2(kaw) - \varepsilon^2 J_2(k\varepsilon aw) = 0. \tag{28}$$

In deriving (28), §8.5 (15) was used, as in the case of circular aperture. For the unobstructed aperture ($\varepsilon = 0$) the first root of (27) is, of course, the value given by (27), namely $w = 3.83/ka = 0.61\lambda/a$. As ε is increased, the first root of (27) decreases,* and with $\varepsilon = \frac{1}{2}$, for example, it is slightly less than $3.15/ka = 0.50\lambda/a$. As the principal maximum remains at $w = 0$ independently of ε, we see that on obstructing the central portion of the aperture the resolving power is increased. This improvement is, however, accompanied by a decrease in the brightness of the image. Also the secondary maxima become more pronounced so that the contrast is reduced. With $\varepsilon = \frac{1}{2}$, the first secondary maximum (at $w = 4.8/ka$) is 0.10 of the principal maximum, as compared with the value 0.018 (at $w = 5.14/ka$) for a circular aperture (see Fig. 8.30).

The obstruction of the central part of the circular aperture corresponds to the replacement of the pupil function [see §8.3 (39)] $G(\xi, \eta) = C$ or 0 according as $0 \leqslant \rho \leqslant a$ or $\rho > a$ by $G(\xi, \eta) = C$ or 0 according as $\varepsilon a \leqslant \rho \leqslant a$ or $\rho < \varepsilon a$, $\rho > a$. Naturally it is possible to alter the pupil function in other ways. A general method for modifying a pupil function consists in depositing on one or more surfaces of the system a thin, partially transmitting film of suitable substance. The same effect may be achieved by means of a specially constructed 'filter', for example, a hollow lens of appropriate form filled with an absorbing liquid. The problem arises of determining the form of the pupil function which, in some agreed sense, would give the best possible image. This problem has been investigated by a number of workers.† Of particular interest is a result due to Toraldo di Francia,‡ that the pupil function may be

* From (26) it follows by a simple calculation that as $\varepsilon \to 1$, $I/I_0 \to J_0^2(kaw)$. Since the first zero of the equation $J_0(x) = 0$ is at $x = 2.40$, it follows that with increasing ε the radius of the first dark ring approaches the value given by $w = 2.40/ka = 0.38\lambda/a$.

† See, for example, R. Straubel, *P. Zeeman Verh.* (Den Haag, Martinus Nijhoff, 1935), p. 302; R. K. Luneburg, *Mathematical Theory of Optics* (Berkeley and Los Angeles, University of California Press, 1964), §50; H. Osterberg and J. E. Wilkins, *J. Opt. Soc. Amer.*, **39** (1949), 553; G. Lansraux, *Rev. d'Optique*, **32** (1953), 475.

A brief review of investigations in this field is given in an article by E. Wolf, *Rep. Progr. Phys.* (London, Physical Society), **14** (1951), 109.

‡ G. Toraldo di Francia, *Suppl. Nuovo Cimento*, **9** (1952), 426.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

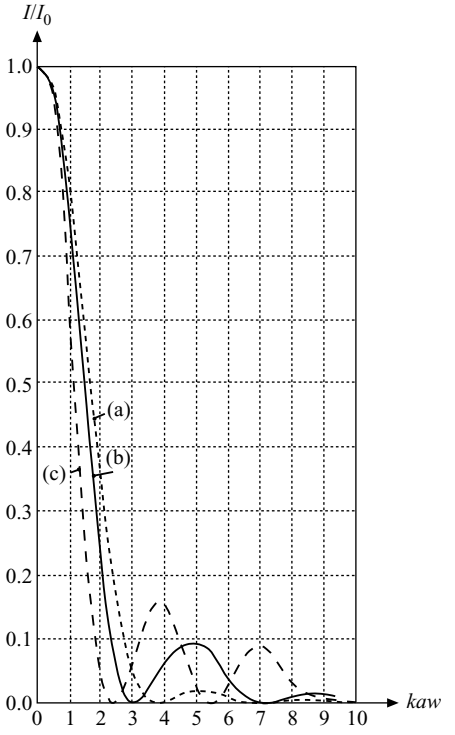


Fig. 8.30 Illustrating the effect of central obstruction on the resolution. Normalized intensity curves for Fraunhofer patterns of: (a) circular aperture; (b) annular aperture with $\varepsilon = \frac{1}{2}$; and (c) annular aperture with $\varepsilon \to 1$. (After G. C. Steward, *The Symmetrical Optical System* (Cambridge, Cambridge University Press, 1928), p. 89.)

so chosen as to make the radius of the first dark ring arbitrarily small and at the same time the dark zone surrounding the central ring arbitrarily large. However, by gradually decreasing the radius of the first dark ring, less and less light is utilized in the central disc, so that the smallest practicable size of the disc and hence the resolving power is limited by the amount of light available.

A partial suppression of the secondary maxima by an appropriate modification of the pupil function is known as *apodization*.* In spectroscopic analysis it facilitates the detection of satellites of spectral lines, whilst in astronomical applications it facilitates the resolution of double stars of appreciably different apparent brightness.

The conventional theory of resolving power, as outlined in this section, is particularly appropriate to direct visual observations. With other methods of detection (e.g. photometric) the presence of two objects of much smaller angular separation than indicated by Rayleigh's criterion may often be revealed. In this connection it is also of

* A thorough review of investigations on this subject was given by P. Jacquinot and B. Roizen-Dossier in *Progress in Optics*, Vol. 3, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1964), p. 29.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

interest to compare the relative resolving efficiency of a telescope and a Michelson stellar interferometer (§7.3.6). If the presence of two stars is judged by means of the first vanishing of the fringes formed by the interferometer, and if d is the maximum distance by which its outer mirrors may be separated, then according to §7.3 (34) double stars down to angular separation $\phi \sim \frac{1}{2}\lambda/d$ may be detected with this instrument. Comparison of this value with (24) shows that, to detect double stars of this separation with a telescope used visually, the diameter $2a$ of its objective would have to be about $2.4d$.

8.6.3 Image formation in the microscope

In the elementary theory of resolving power which we have just outlined, light from the two object points was assumed to be incoherent. This assumption is justified when the two objects are self-luminous, e.g. with stars viewed by a telescope. The intensity observed at any point in the image plane is then equal to the sum of the intensities due to each of the object points.

In a microscope the situation is, as a rule, much more complicated. The object is usually nonluminous and must, therefore, be illuminated with the help of an auxiliary system. Owing to diffraction on the aperture of the illuminating system (condenser), each element of the source gives rise to a diffraction pattern in the object plane of the microscope. The diffraction patterns which have centres on points that are sufficiently close to each other partly overlap, and in consequence the light vibrations at neighbouring points of the object plane are in general partially correlated. Some of this light is transmitted through the object with or without a change of phase, whilst the rest is scattered, reflected or absorbed. In consequence, it is in general impossible to obtain, by means of a single observation, or even by the use of one particular arrangement, a faithful enlarged picture showing all the small-scale structural variations of the object. Various methods of observation have, therefore, been developed, each suitable for the study of certain types of objects, or designed to bring out particular features.

We shall briefly outline the theory of image formation in a microscope, confining our attention first of all to the two extreme cases of completely incoherent and perfectly coherent illumination. Partially coherent illumination will be discussed in §10.5.2.

(a) Incoherent illumination

We first consider a self-luminous object (e.g. an incandescent filament of an electric bulb). Let P be the axial point of the object and Q a neighbouring point in the object plane, at a distance Y from P, and let P' and Q' be the images of these points (Fig. 8.31). Further let θ and θ' be the angles which the marginal rays of the axial pencils make with the axis.

If a' is the radius of the region (assumed to be circular) in which the beam of light converging on P' intersects the back focal plane $\mathcal{F}'$ and if D' is the distance between the back focal plane and the image plane, then, since θ' is small,

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

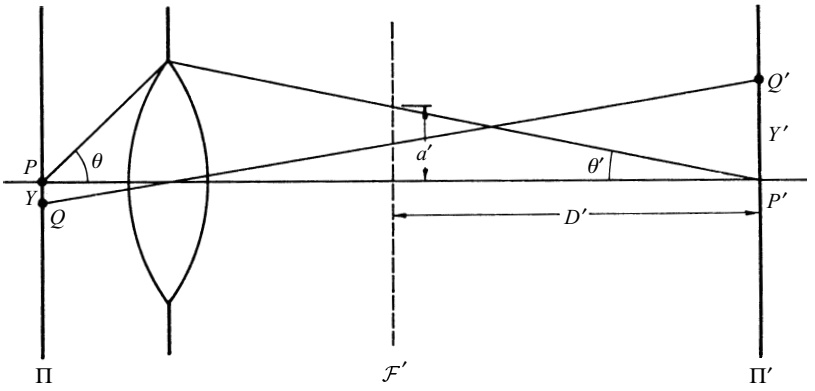


Fig. 8.31 Illustrating the theory of the resolving power of the microscope.

$$\theta' = \frac{a'}{D'}. \tag{29}$$

Further, if $w = \sqrt{p^2 + q^2}$ is the separation of Q' from P' measured in 'diffraction units' [see §8.3 (35) and §8.5 (7)], i.e. the sine of the angle which the two points subtend at the centre of the diffracting aperture, then we have to a good approximation,

$$Y' = w\mathrm{D}'. \tag{30}$$

Let n and n' be the refractive indices, λ and λ' the wavelengths in the object and image spaces, and λ_0 the wavelength in vacuum. Then, since according to §8.5 (16) the first minimum of the diffraction pattern of P is given by $w = 0.61\lambda'/a'$, we have, at the limit of resolution

$$Y' = 0.61\lambda' \frac{\mathrm{D}'}{a'} = 0.61 \frac{\lambda'}{\theta'} = 0.61 \frac{\lambda_0}{n'\theta'}. \tag{31}$$

A microscope must, of course, be so designed that it gives a sharp image not only of an axial point but also of neighbouring points of the object plane. According to §4.5.1 the sine condition must therefore be satisfied, i.e.*

$$nY \sin\theta = -n' Y' \sin\theta'.$$

Since θ' is small we may replace $\sin\theta'$ by θ'. On substituting for Y' into (31), we finally obtain

$$|Y| \sim 0.61 \frac{\lambda_0}{n \sin\theta}. \tag{32}$$

This formula gives the distance between two object points which a microscope can just resolve when the illumination is *incoherent* and the aperture is circular.

The quantity $n \sin\theta$ which enters into (32) is the *numerical aperture* [see §4.8 (13)] and must be large if a high resolving power is to be achieved. Means for obtaining a large numerical aperture were discussed in §6.6.

* The minus sign appears here because θ' corresponds to $-\gamma_1$ of §4.5.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

(b) Coherent illumination — Abbe's theory

We now consider the other extreme case, namely when the light emerging from the object may be treated as strictly coherent. This situation is approximately realized when a thin object of relatively simple structure is illuminated by light from a sufficiently small source via a condenser of low aperture (see §10.5.2).

The first satisfactory theory of resolution with coherent illumination was formulated and also illustrated with beautiful experiments, by E. Abbe.* According to Abbe, the object acts as a diffraction grating, so that not only every element of the aperture of the objective, but also every element of the object must be taken into account in determining the complex disturbance at any particular point in the image plane. Expressed mathematically, the transition from the object to the image involves two integrations, one extending over the object plane, the other extending over the aperture. In Abbe's theory, diffraction by the object is first considered and the effect of the aperture is taken into account in the second stage. An alternative procedure, in which the order is reversed, is also permissible and leads naturally to the same result.†

To illustrate Abbe's theory we consider first the imaging of a grating-like object which is illuminated by a plane wave incident normally on to the object plane (Köhler's central illumination). The wave is diffracted by the object and gives rise to a Fraunhofer diffraction pattern of the grating (see §8.6.1), in the back focal plane $\mathcal{F}'$ of the objective. In Fig. 8.32 the maxima (spectra of successive orders) of this pattern are denoted by $\ldots, S_{-2}, S_{-1}, S_0, S_1, S_2, \ldots$. Every point in the focal plane may be considered to be a centre of a coherent secondary disturbance, whose strength is proportional to the amplitude at that point. The light waves that proceed from these secondary sources will then interfere with each other and will give rise to the image of

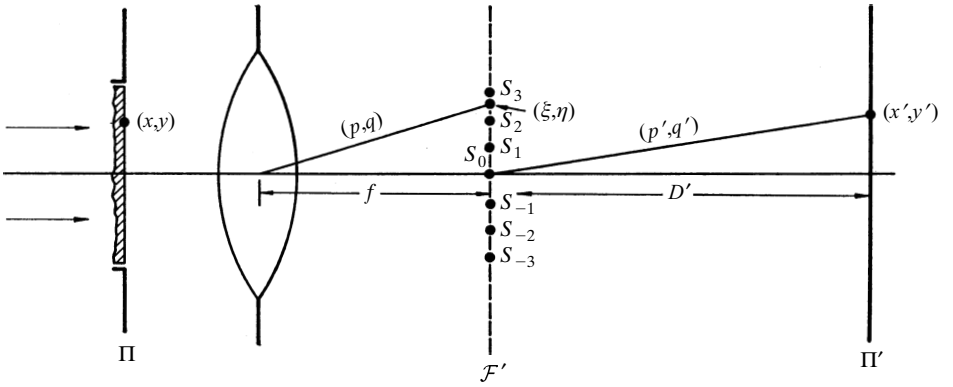


Fig. 8.32 Illustrating the Abbe theory of image formation in a microscope with coherent illumination.

* Ernst Abbe, *Archiv. f. Mikroskopische Anat.*, **9** (1873), 413. Also his *Gesammelte Abhandlungen*, Vol. 1 (Jena, G. Fischer, 1904), p. 45 and O. Lummer and F. Reiche, *Die Lehre von der Bildtstehung im Mikroskop von Ernst Abbe* (Braunschweig, Vieweg, 1910). A good account of Abbe's theory was also given in A. B. Porter, *Phil. Mag.* (6), **11** (1906), 154.

† A theory of image formation in the microscope which is equivalent to this alternative approach was formulated by Lord Rayleigh in *Phil. Mag.* (5), **42** (1896), 167; also his *Scientific Papers*, Vol. 4 (Cambridge, Cambridge University Press, 1903), p. 235.

the object in the image plane Π' of the objective. To obtain a faithful image it is necessary that all the spectra contribute to the formation of the image. Strictly this is never possible because of the finite aperture of the objective. We shall see later that the exclusion of some of the spectra may result in completely false detail appearing in the image. For practical purposes it is evidently sufficient that the aperture shall be large enough to admit all those spectra that carry an appreciable amount of energy.

Let us express these considerations in more precise terms without restricting ourselves to a grating-like object. If x, y are the coordinates of a typical point in the object plane and f is the distance of the focal plane $\mathcal{F}'$ from the lens objective, the disturbance at a point

$$\xi = pf, \qquad \eta = qf, \tag{33}$$

of the $\mathcal{F}'$ plane (see Fig. 8.32) is given by the Fraunhofer formula

$$U(\xi, \eta) = C_1 \iint_{\mathcal{A}} F(x, y) \mathrm{e}^{-\mathrm{i}k[\frac{\xi}{f}x + \frac{\eta}{f}y]} \,\mathrm{d}x \,\mathrm{d}y, \tag{34}$$

where F is the transmission function of the object, C_1 is a constant, and the integration is taken over the area $\mathcal{A}$ of the object plane Π covered by the object.

Next consider the transition from the back focal plane $\mathcal{F}'$ to the image plane Π'. If, as before, D' denotes the distance between $\mathcal{F}'$ and Π', and $V(x', y')$ is the disturbance at a typical point

$$x' = p'\mathrm{D}', \qquad y' = q'\mathrm{D}', \tag{35}$$

of the image plane, we have for Fraunhofer diffraction on the aperture $\mathcal{B}$ in $\mathcal{F}'$

$$V(x', y') = C_2 \iint_{\mathcal{B}} U(\xi, \eta) \mathrm{e}^{-\mathrm{i}k[\frac{x'}{\mathrm{D}'}\xi + \frac{y'}{\mathrm{D}'}\eta]} \,\mathrm{d}\xi \,\mathrm{d}\eta, \tag{36}$$

it being assumed that $a'/D' \ll 1$ (see Fig. 8.31). Substitution from (34) into (36) gives

$$V(x', y') = C_1 C_2 \iint_{\mathcal{A}} \iint_{\mathcal{B}} F(x, y) \mathrm{e}^{-\mathrm{i}\frac{k}{f}[(x + \frac{f}{\mathrm{D}'}x')\xi + (y + \frac{f}{\mathrm{D}'}y')\eta]} \,\mathrm{d}x \,\mathrm{d}y \,\mathrm{d}\xi \,\mathrm{d}\eta. \tag{37}$$

Now if $F(x, y)$ is defined as zero for all points of the object plane that lie outside $\mathcal{A}$, the integration with respect to x and y may formally be extended from $-\infty$ to $+\infty$. Also, if the aperture $\mathcal{B}$ is so large that $|U(\xi, \eta)|$ is negligible for points of the $\mathcal{F}'$-plane that lie outside $\mathcal{B}$, the integrations with respect to ξ and η may likewise each be extended over the range from $-\infty$ to $+\infty$. Noting also that (cf. §4.3 (10) where f' and Z' correspond to our f and $-D'$ respectively)

$$\frac{f}{\mathrm{D}'} = -\frac{1}{M}, \tag{38}$$

where $M(<0)$ is the magnification between Π and Π', we obtain by the application of the Fourier integral theorem*

$$V(x', y') = CF\left(\frac{x'}{M}, \frac{y'}{M}\right) = CF(x, y), \tag{39}$$

* See for example R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1 (New York, Interscience Publishers, 1953), p. 79.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

where $(x,\ y)$ is the object point whose image is at $(x',\ y')$, and

$$C = C_1 C_2 \lambda^2 f^2$$

is a constant. Hence to the accuracy here implied* the image is strictly similar to the object (but inverted), provided the aperture is large enough.

To show that completely false detail may appear in the image if some of the spectra that carry appreciable energy are excluded, we consider a one-dimensional grating-like object consisting of N equidistant congruent slits of width s, separated by opaque regions, with period d. For simplicity the aperture will be assumed to be rectangular with two of its sides parallel to the strips.

According to §8.6 (3)

$$U(\xi) = C_1' \left(\frac{\sin \dfrac{k\xi s}{2f}}{\dfrac{k\xi s}{2f}} \right) \frac{1 - \mathrm{e}^{-\mathrm{i}Nkd\xi/f}}{1 - \mathrm{e}^{-\mathrm{i}kd\xi/f}}, \tag{40}$$

where for $U^{(0)}$ there has been substituted the expression relating to diffraction on a rectangular aperture and $C_1{}'$ is a constant (see §8.5.1). If the rectangular aperture extends in the ξ direction throughout the range

$$-a \leqslant \xi \leqslant a,$$

the disturbance in the image plane is, by (36) and (40), given by (C' denoting a constant)

$$V(x') = C' \int_{-a}^{a} \frac{\sin \dfrac{k\xi s}{2f}}{\dfrac{k\xi s}{2f}} \frac{1 - \mathrm{e}^{-\mathrm{i}Nkd\xi/f}}{1 - \mathrm{e}^{-\mathrm{i}kd\xi/f}} \mathrm{e}^{-\mathrm{i}kx'\xi/\mathrm{D}'} \, \mathrm{d}\xi. \tag{41}$$

The position of principal maxima of the integrand are given by the roots of the equation $1 - \exp[-\mathrm{i}kd\xi/f] = 0$, i.e. by $\xi = mf\lambda/d$, where m is an integer. Between these principal maxima there are weak secondary maxima. If N is large, the principal maxima are very sharp and the secondary maxima negligible in comparison. To a good approximation we may then replace the integral by a sum of integrals, each extending from the midpoint Q_m of the interval between two successive principal maxima to the next midpoint Q_{m+1}. In each interval we may replace the argument by the central value $\xi = mf\lambda/d = 2\pi mf/kd$, and obtain for V the following expression:

$$V(x') \sim V_0 \sum_{-\overline{m} < m < \overline{m}} \frac{\sin \dfrac{m\pi s}{d}}{\dfrac{m\pi s}{d}} \mathrm{e}^{\frac{2\pi \mathrm{i} m x'}{d'}}. \tag{42}$$

Here

$$\overline{m} = \frac{ad}{\lambda f}, \qquad d' = Md = -\frac{\mathrm{D}'}{f} d, \tag{43}$$

* As pointed out on p. 427, the Fraunhofer approximation used here is restricted to the case when the object points as well as the image points are sufficiently close to the axis.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

and V_0 is the integral

$$V_0 = C' \int_{Q_m}^{Q_{m+1}} \frac{1 - e^{-iNkd\xi/f}}{1 - e^{-ikd\xi/f}} \, d\xi, \tag{44}$$

which, apart from small correction terms in the high orders, is practically independent of m. The series (42) may be re-written in real form as

$$\frac{V(x')}{V_0} = 1 + 2 \sum_{1 < m < \overline{m}} \frac{\sin \dfrac{m\pi s}{d}}{\dfrac{m\pi s}{d}} \cos \frac{2\pi m x'}{d'}. \tag{45}$$

Suppose first that the length a of the aperture is very large. The summation may then formally be extended over the whole infinite range ($\overline{m} = \infty$), and we can easily verify that the image is then strictly similar to the object. For this purpose we expand the transmission function F of the grating-like object (see Fig. 8.33),

$$\left.\begin{aligned} F(x) &= F_0 \qquad 0 < |x| < s/2 \\ &= 0 \qquad s/2 < |x| < d/2 \end{aligned}\right\} \tag{46}$$

into a Fourier series

$$F(x) = c_0 + 2 \sum_{m=1}^{\infty} c_m \cos \frac{2\pi m x}{d}. \tag{47}$$

Then

$$c_0 = \frac{F_0 s}{d}, \qquad c_m = F_0 \frac{\sin \dfrac{\pi m s}{d}}{\pi m} \qquad (m = 1, 2, 3, \ldots). \tag{48}$$

We see that apart from a constant factor this series is the same as (45).

Suppose now that the length a of the aperture is decreased. If a is so small that only the zero-order spectrum contributes to the image, i.e. if $\overline{m} = ad/\lambda f$ is only a fraction of unity, then according to (45) $V(x') =$ constant, so that the image plane is uniformly illuminated. (This result is, of course, not strictly true, as we have neglected certain error terms; in reality there is a weak drop in intensity towards the edge.)

If in addition to the zero-order spectrum the two spectra of the first order (S_1, S_{-1}) are also admitted by the aperture, i.e. if $\overline{m} = ad/\lambda f$ is slightly greater than unity, then we see from (45) that

$$\frac{V(x')}{V_0} = 1 + 2 \frac{\sin \dfrac{\pi s}{d}}{\dfrac{\pi s}{d}} \cos \frac{2\pi x'}{d'}. \tag{49}$$

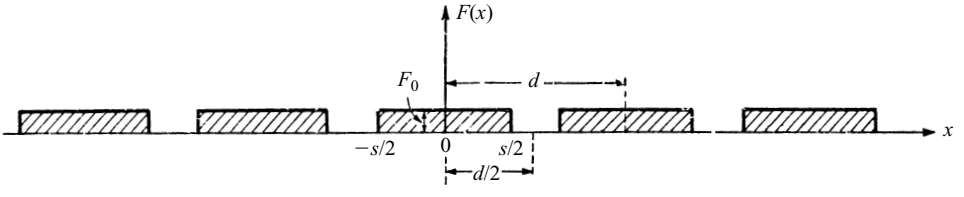


Fig. 8.33 A grating-like object.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

The image has now the correct periodicity $x' = d'$, but a considerably flattened intensity distribution. By increasing the aperture more and more the image is seen to resemble the object more and more closely.

A completely false image is obtained when the lower orders are excluded. If for example all orders except the second are excluded, then

$$\frac{V(x')}{V_0} = 2\frac{\sin\dfrac{2\pi s}{d}}{\dfrac{2\pi s}{d}}\cos\frac{4\pi x'}{d'}, \tag{50}$$

so that the image has the period $x' = d'/2$; the 'image' shows twice the number of lines that are in fact present in the object.

Finally let us estimate the resolving power. Consider again the situation illustrated in Fig. 8.31, but assume now that the light from P and Q is coherent. Then the distribution in the image plane arises essentially from the coherent superposition of the two Airy diffraction patterns, one centred on P', the other on Q'. The complex amplitude at a point situated between P' and Q' at distance w_1 (measured in 'diffraction units') from P' is given by

$$U(w_1) = \left[\frac{2J_1(kaw_1)}{kaw_1} + \frac{2J_1[ka(w-w_1)]}{ka(w-w_1)}\right]U_0, \tag{51}$$

w being the distance between P' and Q' and the other symbols having the same meaning as before. The intensity is, therefore, given by

$$I(w_1) = \left[\frac{2J_1(kaw_1)}{kaw_1} + \frac{2J_1[ka(w-w_1)]}{ka(w-w_1)}\right]^2 I_0. \tag{52}$$

Now in the case of incoherent illumination, P' and Q' were considered as resolved when the principal intensity maximum of the one pattern coincided with the first minimum of the other. The intensity at the midpoint ($kaw \sim 1.92$) between the two maxima is then equal to $2[2J_1(1.92)/1.92]^2 \sim 0.735$ of the maximum intensity of either, i.e. the combined intensity curve has a dip of about 26.5 per cent between the principal maxima. (This corresponds to the value 19 per cent for a slit aperture — see Fig. 7.62.) If we consider a dip of this amount as again substantially determining the limit of resolution, the critical separation $w = 2w_1$ is obtained from the relation

$$\frac{I(w_1)}{I(0)} = 0.735 \qquad (w = 2w_1). \tag{53}$$

The first root of this transcendental equation is $w_1 \sim 2.57/ka$, so that the critical separation measured in ordinary units is

$$Y' = 2w_1\mathrm{D}' \sim \frac{2.57}{\pi}\frac{\mathrm{D}'\lambda'}{a} = \frac{0.82\lambda'}{\theta'} = \frac{0.82\lambda_0}{n'\theta'}. \tag{54}$$

To relate Y' to the corresponding separation Y of the object points we use the sine condition (with the approximation $\sin\theta' \sim \theta'$), and finally obtain for the *limit of resolution with coherent illumination* the expression

$$|Y| = 0.82\frac{\lambda_0}{n\sin\theta}. \tag{55}$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Apart from a larger numerical factor (which in any case is somewhat arbitrary as it depends on the form of the object and aperture and on the sensitivity of the receptor), we obtain the same expression as in the case of incoherent illumination ((32)). Thus with light of a given wavelength the resolving power is again substantially determined by the numerical aperture of the object.

(c) Coherent illumination — Zernike's phase contrast method of observation[*]

We have defined a *phase object* as one which alters the phase but not the amplitude of the incident wave. An object of this type is of nonuniform optical thickness, but does not absorb any of the incident light. Such objects are frequently encountered in biology, crystallography and other fields. It is evident from the preceding discussion that with ordinary methods of observation little information about phase objects can be obtained. For the complex amplitude function that specifies the disturbance in the image plane is then similar to the transmission function of the object† and, as the eye (or any other observing instrument) only distinguishes changes in intensity, one can only draw conclusions about the amplitude changes but not about the phase changes introduced by the object.

To obtain information about phase objects, special methods of observation must be used, for example, the so-called *central dark ground method of observation* where the central order is excluded by a stop, or the *Schlieren method*, where all the spectra on one side of the central order are excluded. The most powerful method, which has the advantage that it produces an intensity distribution which is directly proportional to the phase changes introduced by the object, is due to Zernike‡ and was first described by him in 1935. It is known as *the phase contrast method*. In 1953 Zernike was awarded the Nobel Prize in Physics for this invention.

To explain the principle of the phase contrast method, consider first a transparent object in the form of a one-dimensional phase grating. The transmission function of such an object is by definition (see p. 447) of the form

$$F(x) = \mathrm{e}^{\mathrm{i}\phi(x)}, \tag{56}$$

where $\phi(x)$ is a real periodic function, whose period (d say) is equal to the period of the grating. We assume that the magnitude of ϕ is small compared to unity, so that we may write

$$F(x) \sim 1 + \mathrm{i}\phi(x). \tag{57}$$

[*] For fuller discussion of the phase contrast method see, for example, M. Françon, *Le contraste de phase en optique et en microscopie* (Paris, Revue d'Optique, 1950); and A. H. Bennett, H. Jupnik, H. Osterberg, and O. W. Richards, *Phase Microscopy* (New York, J. Wiley & Sons, 1952).

† Strict similarity would actually be attained only if the objective had an infinite aperture. Because the aperture is always finite, some details of the phase structures can be seen. In some cases the visibility of such 'images' is enhanced, at the expense of resolution, by a slight defocusing of the instrument (see H. H. Hopkins, contribution in M. Françon, *Le contraste de phase et le contraste par interférences* (Paris, Revue d'Optique, 1952), p. 142).

‡ F. Zernike, *Z. Tech. Phys.*, **16** (1935), 454; *Phys. Z.*, **36** (1935), 848; *Physica*, **9** (1942), 686, 974.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

If we develop F into a Fourier series

$$F(x) = \sum_{m=-\infty}^{\infty} c_m e^{\frac{2\pi i mx}{d}}, \tag{58}$$

then, since F is of the form (56) and ϕ is real and numerically small compared to unity,

$$c_0 = 1, \qquad c_{-m} = -c_m^{\star} \qquad (m \neq 0). \tag{59}$$

The intensity of the mth-order spectra is proportional to $|c_m|^2$.

In the phase contrast method of observation a thin plate of transparent material called the *phase plate* is placed in the back focal plane $\mathcal{F}'$ of the objective and by means of it the phase of the central order (S_0 in Fig. 8.32) is retarded or advanced with respect to the diffraction spectra (S_1, S_{-1}, S_2, S_{-2}, ...) by one-quarter of a period. This means that the complex amplitude distribution in the focal plane is altered from a distribution characterized by the coefficients c_m, to a distribution characterized by coefficients c'_m, where

$$c'_0 = c_0 e^{\pm i\pi/2} = \pm i, \qquad c'_m = c_m \qquad (m \neq 0), \tag{60}$$

the positive or negative sign being taken according as the phase of the central order is retarded or advanced. The resulting light distribution in the image plane will now no longer represent the phase grating (57), but rather a fictitious amplitude grating

$$G(x) = \pm i + i\phi(x). \tag{61}$$

Hence the intensity in the image plane will now be proportional to (neglecting ϕ^2 in comparison to unity)

$$I(x') = |G(x)|^2 = 1 \pm 2\phi(x), \tag{62}$$

where as before $x' = Mx$, M being the magnification. This relation shows that *with the phase contrast method of observation, phase changes introduced by the object are transformed into changes in intensity, the intensity at any point of the image plane being* (*apart from an additive constant*) *directly proportional to the phase change due to the corresponding element of the object*.* When the phase of the central order is retarded with respect to the diffraction spectra [upper sign in (61)], regions of the object which have greater optical thickness will appear brighter than the mean illumination, and one then speaks of a *bright phase contrast*; when the phase of the central order is advanced, regions of greater spectral thickness will appear darker and one then speaks of a *dark phase contrast* (Figs. 8.34 and 8.35).

To obtain good resolution, the aperture of the illuminating system is often of annular rather than circular form (see §8.6.2). In this case the annular region of $\mathcal{F}'$ through which the direct (undiffracted) light passes plays the role of the central order S_0 of Fig. 8.32, and it is this light which must then be retarded or advanced by a quarter period.

The phase-changing plate may be produced by evaporating a thin layer of a suitable dielectric substance on to a glass substrate. If n is the refractive index of the substance

* The approximations implicit in the elementary theory of the phase contrast method are discussed by J. Picht, *Zeitschr. f. Instrkde*, **56** (1936), 481; *ibid.* **58** (1938), 1 and by F. D. Kahn, *Proc. Phys. Soc.*, B **68** (1955), 1073.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

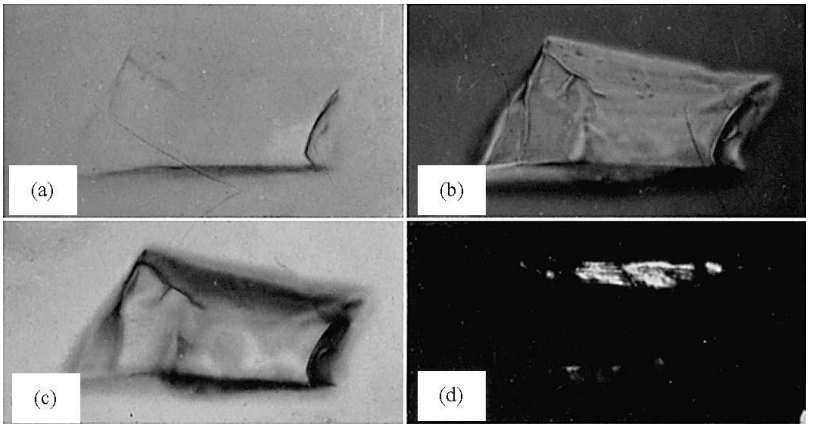


Fig. 8.34 Microscope images of glass fragments ($n = 1.52$) mounted in clarite ($n = 1.54$), 100×: (a) bright field image; (b) and (c) phase contrast images; (d) dark field image. (After A. H. Bennett, H. Jupnik, H. Osterberg and O. W. Richards, *Trans. Amer. Microscop. Soc.*, **65** (1946), 126.)

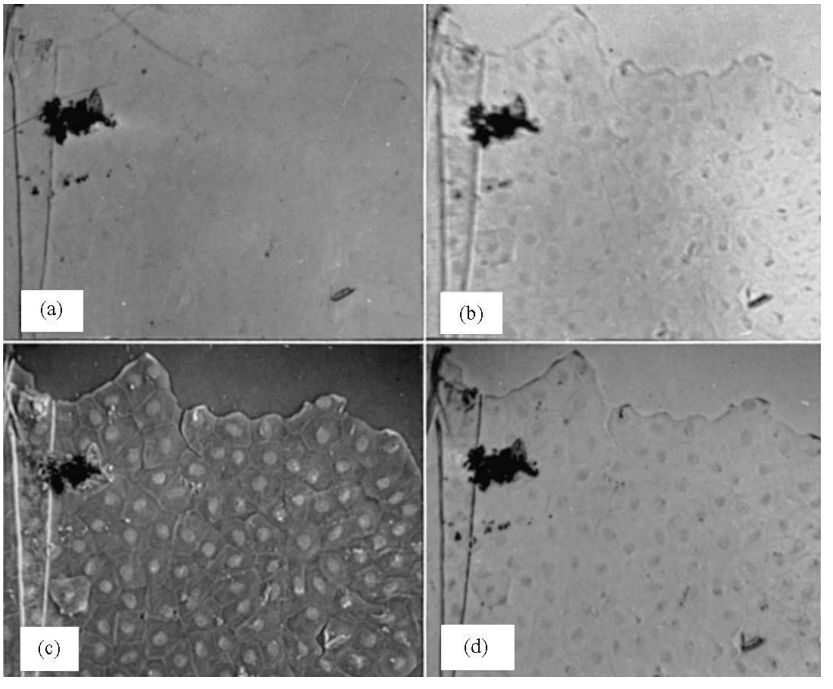


Fig. 8.35 Microscope images of epithelium from frog nictitating membrane, 100×: (a) bright field image at full aperture, NA = 0.25; (b) bright field image with aperture half filled; (c) phase-contrast image (bright contrast) at full aperture; (d) phase-contrast image (dark contrast) at full aperture. (After A. H. Bennett, H. Jupnik, H. Osterberg and O. W. Richards, *Trans. Amer. Microscop. Soc.*, **65** (1946), 119.)

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

and d the thickness of the layer, then for a retardation of a quarter of a period one must have $d = \lambda/4(n-1)$. A retardation of the central order by this amount is, of course, equivalent to an advance of the diffracted spectra by three-quarters of a period, and vice versa. It is possible to increase the sensitivity of the method by using slightly absorbing instead of a dielectric coating. We shall return to this point later.

It remains to show that the phase contrast method is not restricted to phase objects of periodic structure. For this purpose we divide the integral (34) into two parts:

$$U(\xi, \eta) = U_0(\xi, \eta) + U_1(\xi, \eta), \tag{63}$$

where

$$\left.\begin{aligned} U_0 &= C_1 \iint_{\mathcal{A}} \mathrm{e}^{-\frac{\mathrm{i}k}{f}[\xi x + \eta y]}\, \mathrm{d}x\, \mathrm{d}y, \\ U_1 &= C_1 \iint_{\mathcal{A}} [F(x, y) - 1]\mathrm{e}^{-\frac{\mathrm{i}k}{f}[\xi x + \eta y]}\, \mathrm{d}x\, \mathrm{d}y. \end{aligned}\right\} \tag{64}$$

U_0 represents the light distribution that would be obtained in the plane $\mathcal{F}'$ if no object were present, whilst U_1 represents the effect of diffraction. Now the 'direct light' U_0 (corresponding to the central order S_0 of Fig. 8.32), will be concentrated in only a small region $\mathcal{B}_0$ of the $\mathcal{F}'$-plane, around the axial point $\xi = \eta = 0$. On the other hand a very small fraction of the diffracted light will, in general, reach this region, most of it being diffracted to other parts of this plane.*

Suppose that the region $\mathcal{B}_0$ through which the direct light passes is covered by a phase plate. The effect of the plate may be described by a transmission function

$$A = a\mathrm{e}^{\mathrm{i}\alpha}. \tag{65}$$

For a plate that only retards or advances the light which is incident upon it, $a = 1$; for a plate that also absorbs light, $a < 1$. The light emerging from the aperture will be represented by

$$U'(\xi, \eta) = AU_0(\xi, \eta) + U_1(\xi, \eta) \tag{66}$$

so that, according to (36), the distribution of the complex amplitude in the image is given by

$$V(x', y') = V_0(x', y') + V_1(x', y'), \tag{67}$$

where

$$\left.\begin{aligned} V_0 &= AC_2 \iint_{\mathcal{B}} U_0(\xi, \eta)\mathrm{e}^{-\frac{\mathrm{i}k}{D'}[x'\xi + y'\eta]}\, \mathrm{d}\xi\, \mathrm{d}\eta, \\ V_1 &= C_2 \iint_{\mathcal{B}} U_1(\xi, \eta)\mathrm{e}^{-\frac{\mathrm{i}k}{D'}[x'\xi + y'\eta]}\, \mathrm{d}\xi\, \mathrm{d}\eta. \end{aligned}\right\} \tag{68}$$

Now the aperture $\mathcal{B}$ greatly exceeds in size the region $\mathcal{B}_0$, and since U_0 was seen to be practically zero outside $\mathcal{B}_0$, on appreciable error is introduced by extending the domain of integration in V_0 over the whole $\mathcal{F}'$-plane. Moreover, if $\mathcal{B}$ is assumed to be so large

* This point was investigated in detail by J. Picht, *Zeitschr. f. Instrkde*, **58** (1938), 1. See also F. Zernike, *Mon. Not. Roy. Astr. Soc.*, **94** (1934), 382–383, where it is discussed in a somewhat different connection.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

as to admit all the diffracted rays that carry any appreciable energy, the integral for V_1 may likewise be given infinite limits. Finally, if as before the transmission function $F(x, y)$ is defined as zero at points of the object plane outside the region covered by the object, the integrals (64) may also be taken with infinite limits. We then obtain, on substituting from (64) into (68), and using the Fourier integral theorem and the relation (38),

$$\left.\begin{aligned} V_0(x', y') &= CA, \\ V_1(x', y') &= C\left[F\left(\frac{x'}{M}, \frac{y'}{M}\right) - 1\right] = C[F(x, y) - 1]. \end{aligned}\right\} \tag{69}$$

From (67) and (69) it follows that the intensity in the image plane is given by

$$I(x', y') = |V(x', y')|^2 = |C|^2|A + F(x, y) - 1|^2. \tag{70}$$

With a phase object

$$F(x, y) = \mathrm{e}^{\mathrm{i}\phi(x,y)}, \tag{71}$$

and (70) reduces to*

$$I(x', y') = |C|^2(a^2 + 2\{1 - a\cos\alpha - \cos\phi(x, y) + a\cos[\alpha - \phi(x, y)]\}). \tag{72}$$

Since ϕ was assumed to be small, (72) may be written as

$$I(x', y') = |C|^2[a^2 + 2a\phi(x, y)\sin\alpha], \tag{73}$$

and, if the phase difference introduced by the plate represents a retardation or advance by a quarter of a period, then $\alpha = \pm\pi/2$ and (73) reduces to

$$I(x', y') = |C|^2[a^2 \pm 2a\phi(x, y)]. \tag{74}$$

When the plate does not absorb any of the incident light ($a = 1$) we have again the expression (62). The intensity changes are then directly proportional to the phase variations of the object. With a plate that absorbs a fraction a^2 of the direct light the ratio of the second term to the first term in (73) has the value $\pm\phi/a$, so that the contrast of the image is enhanced. For example, by weakening the direct light to one-ninth of its original value, the sensitivity of the method is increased three times.

8.7 Fresnel diffraction at a straight edge

8.7.1 The diffraction integral

Having considered various cases of Fraunhofer diffraction, we now turn our attention to the more general case of Fresnel diffraction.

The basic diffraction integral §8.3 (28) may be written in the form

$$U(P) = B(C + \mathrm{i}S), \tag{1}$$

where

* The special case when $a = 0$ corresponds to the dark-ground method of observation. According to (72), the intensity distribution is then given by

$$I(x', y') = 2C^2[1 - \cos\phi(x, y)].$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$B = -A\frac{\mathrm{i}}{\lambda}\cos\delta\,\frac{\mathrm{e}^{\mathrm{i}k(r'+s')}}{r's'}, \tag{2}$$

$$\left.\begin{aligned} C &= \iint_{\mathcal{A}} \cos[kf(\xi,\eta)]\mathrm{d}\xi\,\mathrm{d}\eta, \\ S &= \iint_{\mathcal{A}} \sin[kf(\xi,\eta)]\mathrm{d}\xi\,\mathrm{d}\eta. \end{aligned}\right\} \tag{3}$$

The intensity $I(P) = |U(P)|^2$ at the point P of observation is then given by

$$I(P) = |B|^2(C^2 + S^2). \tag{4}$$

We must now retain in the expansion §8.3 (31) for $f(\xi, \eta)$ terms in ξ and η at least up to the second order.

As before we take the plane of the aperture $\mathcal{A}$ as the x, y-plane. To simplify the calculations we choose as the x direction the projection of the line P_0P onto the plane of the aperture (Fig. 8.36). Thus with a source in a prescribed position our reference system will in general be different for different points of observation.

According to §8.3 (30), we now have $l = l_0$, $m = m_0$, so that the linear terms in $f(\xi, \eta)$ disappear. The direction cosines of the rays P_0O and OP are

$$\left.\begin{aligned} l &= l_0 = \sin\delta, \\ m &= m_0 = 0, \\ n &= n_0 = \cos\delta, \end{aligned}\right\} \tag{5}$$

where, as before, δ denotes the angle between the line P_0P and the z-axis. The expression §8.3 (31) for $f(\xi, \eta)$ reduces to

$$f(\xi,\eta) = \frac{1}{2}\left(\frac{1}{r'} + \frac{1}{s'}\right)(\xi^2\cos^2\delta + \eta^2) + \cdots. \tag{6}$$

If we neglect terms of third and higher order in ξ and η, the integrals (3) become

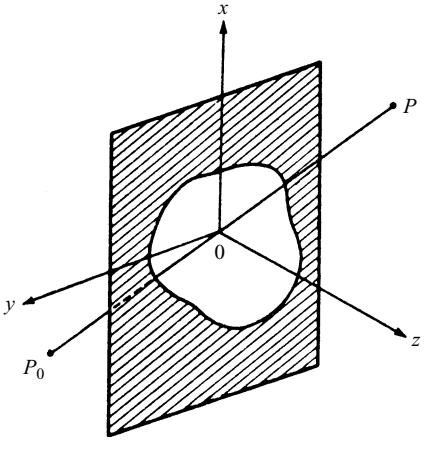


Fig. 8.36 Fresnel diffraction at an opening in a plane opaque screen.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$\left.\begin{aligned} C &= \iint_{\mathcal{A}} \cos\left[\frac{\pi}{\lambda}\left(\frac{1}{r'}+\frac{1}{s'}\right)(\xi^2\cos^2\delta+\eta^2)\right]\mathrm{d}\xi\,\mathrm{d}\eta, \\ S &= \iint_{\mathcal{A}} \sin\left[\frac{\pi}{\lambda}\left(\frac{1}{r'}+\frac{1}{s'}\right)(\xi^2\cos^2\delta+\eta^2)\right]\mathrm{d}\xi\,\mathrm{d}\eta. \end{aligned}\right\} \tag{7}$$

It is convenient to introduce new variables of integration u, v, defined by

$$\left.\begin{aligned} \frac{\pi}{\lambda}\left(\frac{1}{r'}+\frac{1}{s'}\right)\xi^2\cos^2\delta &= \frac{\pi}{2}u^2, \\ \frac{\pi}{\lambda}\left(\frac{1}{r'}+\frac{1}{s'}\right)\eta^2 &= \frac{\pi}{2}v^2. \end{aligned}\right\} \tag{8}$$

Then

$$\mathrm{d}\xi\,\mathrm{d}\eta = \frac{\lambda}{2}\frac{\mathrm{d}u\,\mathrm{d}v}{\left(\frac{1}{r'}+\frac{1}{s'}\right)\cos\delta}$$

and our integrals become

$$\left.\begin{aligned} C &= b\iint_{\mathcal{A}'} \cos\left[\frac{\pi}{2}(u^2+v^2)\right]\mathrm{d}u\,\mathrm{d}v, \\ S &= b\iint_{\mathcal{A}'} \sin\left[\frac{\pi}{2}(u^2+v^2)\right]\mathrm{d}u\,\mathrm{d}v, \end{aligned}\right\} \tag{9}$$

where

$$b = \frac{\lambda}{2\left(\frac{1}{r'}+\frac{1}{s'}\right)\cos\delta}. \tag{10}$$

The integration now extends over the region $\mathcal{A}'$ of the u, v-plane into which the region $\mathcal{A}$ of the aperture is transformed by means of (8).

8.7.2 Fresnel's integrals

If $\mathcal{A}'$ is a rectangle with sides parallel to the axes of u and v, the integrals are simplified still further by means of the identities

$$\left.\begin{aligned} \cos\left[\frac{\pi}{2}(u^2+v^2)\right] &= \cos\left(\frac{\pi}{2}u^2\right)\cos\left(\frac{\pi}{2}v^2\right) - \sin\left(\frac{\pi}{2}u^2\right)\sin\left(\frac{\pi}{2}v^2\right), \\ \sin\left[\frac{\pi}{2}(u^2+v^2)\right] &= \sin\left(\frac{\pi}{2}u^2\right)\cos\left(\frac{\pi}{2}v^2\right) + \cos\left(\frac{\pi}{2}u^2\right)\sin\left(\frac{\pi}{2}v^2\right). \end{aligned}\right\} \tag{11}$$

To evaluate (9) in this case we must consider the integrals

$$\left.\begin{aligned} \mathcal{C}(w) &= \int_0^w \cos\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau, \\ \mathcal{S}(w) &= \int_0^w \sin\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau. \end{aligned}\right\} \tag{12}$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$\mathcal{C}(w)$ and $\mathcal{S}(w)$ are known as *Fresnel's integrals*. They are of importance in connection with many diffraction problems and have been extensively studied. We must briefly consider some of their properties.*

First we derive series expressions for $\mathcal{C}(w)$ and $\mathcal{S}(w)$. Expanding the cosine and sine under the integral signs into power series and integrating term by term we find that

$$\left.\begin{aligned}\mathcal{C}(w) &= w\left[1-\frac{1}{2!5}\left(\frac{\pi}{2}w^2\right)^2+\frac{1}{4!9}\left(\frac{\pi}{2}w^2\right)^4-\cdots\right],\\ \mathcal{S}(w) &= w\left[\frac{1}{1!3}\left(\frac{\pi}{2}w^2\right)-\frac{1}{3!7}\left(\frac{\pi}{2}w^2\right)^3+\frac{1}{5!11}\left(\frac{\pi}{2}w^2\right)^5-\cdots\right].\end{aligned}\right\} \tag{13}$$

The series (13) are convergent for all values of w but are suitable for computations only when w is small. When w is large the integrals may be evaluated from series in inverse powers of w. We re-write (12) as

$$\mathcal{C}(w) = \mathcal{C}(\infty) - \int_w^\infty \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\sin\frac{\pi}{2}\tau^2\right)\frac{\mathrm{d}\tau}{\pi\tau}. \tag{14}$$

Integration by parts gives

$$\mathcal{C}(w) = \mathcal{C}(\infty) + \frac{1}{\pi w}\sin\left(\frac{\pi}{2}w^2\right) + \int_w^\infty \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\cos\frac{\pi}{2}\tau^2\right)\frac{\mathrm{d}\tau}{\pi^2\tau^3}.$$

Integrating again by parts and continuing this process we obtain

$$\left.\begin{aligned}\mathcal{C}(w) &= \mathcal{C}(\infty) - \frac{1}{\pi w}\left[P(w)\cos\left(\frac{\pi}{2}w^2\right) - Q(w)\sin\left(\frac{\pi}{2}w^2\right)\right],\\ \text{and similarly}\quad \mathcal{S}(w) &= \mathcal{S}(\infty) - \frac{1}{\pi w}\left[P(w)\sin\left(\frac{\pi}{2}w^2\right) + Q(w)\cos\left(\frac{\pi}{2}w^2\right)\right],\end{aligned}\right\} \tag{15}$$

where

$$\left.\begin{aligned}Q(w) &= 1-\frac{1\times3}{(\pi w^2)^2}+\frac{1\times3\times5\times7}{(\pi w^2)^4}-\cdots,\\ P(w) &= \frac{1}{\pi w^2}-\frac{1\times3\times5}{(\pi w^2)^3}+\frac{1\times3\times5\times7\times9}{(\pi w^2)^5}\cdots.\end{aligned}\right\} \tag{16}$$

To evaluate the integrals $\mathcal{C}(\infty)$ and $\mathcal{S}(\infty)$, we combine them into a complex integral

$$\mathcal{C}(\infty) + \mathrm{i}\mathcal{S}(\infty) = \int_0^\infty \mathrm{e}^{\mathrm{i}(\pi/2)\tau^2}\,\mathrm{d}\tau \tag{17}$$

* Of the numerous tables of Fresnel's integrals we may refer to the following: British Association Report (Oxford, 1926), 273–275. E. Jahnke and F. Emde, *Tables of Functions with Formulae and Curves* (Leipzig and Berlin, Teubner; reprinted by Dover Publications, New York, 4th edition, 1945), p. 35. G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge, Cambridge University Press, 2nd edition, 1944), p. 744. T. Pearcey, *Tables of Fresnel Integrals to Six Decimal Places* (Cambridge, Cambridge University Press, 1956).

and introduce a new variable of integration

$$\zeta = \tau\sqrt{-\frac{\mathrm{i}\pi}{2}} = \tau\frac{\mathrm{i}-1}{2}\sqrt{\pi}, \qquad \tau = -\zeta\frac{\mathrm{i}+1}{\sqrt{\pi}}.$$

The real path of integration $0 \leqslant \tau \leqslant \infty$ goes over into a path along a line through the origin and inclined at 45° to the real axis in the complex ζ-plane. Now it is easy to see that if the integral is taken along a line parallel to the imaginary axis then, with increasing distance from the origin, the integral tends to zero. It then follows from Cauchy's residue theorem that the integral taken along any oblique line through the origin is equal to the value of the integral taken along the real axis. Hence

$$\mathcal{C}(\infty) + \mathrm{i}\mathcal{S}(\infty) = \frac{\mathrm{i}+1}{\sqrt{\pi}}\int_0^\infty \mathrm{e}^{-\zeta^2}\,\mathrm{d}\zeta = \frac{\mathrm{i}+1}{2}.$$

(The real integral with respect to ζ is the well-known Gaussian error integral,* and has the value $\sqrt{\pi}/2$.) Thus

$$\left.\begin{aligned}\mathcal{C}(\infty) &= \int_0^\infty \cos\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau = \tfrac{1}{2},\\ \mathcal{S}(\infty) &= \int_0^\infty \sin\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau = \tfrac{1}{2}.\end{aligned}\right\} \tag{18}$$

The relations (15), together with (16) and (18), express Fresnel's integrals in series of inverse powers of w. These are divergent (asymptotic) series, which provide a good approximation to the integrals when w is large by taking only a limited number of terms into account (see Appendix III).

The behaviour of Fresnel's integrals may be illustrated by means of an elegant geometrical representation due to Cornu.† $\mathcal{C}$ and $\mathcal{S}$ are regarded as rectangular Cartesian coordinates of a point P. As w takes on all the possible values, the point P describes a curve. Since $\mathcal{C}(0) = \mathcal{S}(0) = 0$ the curve passes through the origin, and since

$$\mathcal{C}(-w) = -\mathcal{C}(w), \qquad \mathcal{S}(-w) = -\mathcal{S}(w), \tag{19}$$

it is antisymmetric with respect to both axes. If $\mathrm{d}l$ is an element of arc of the curve then

$$\begin{aligned}\mathrm{d}l^2 = \mathrm{d}\mathcal{C}^2 + \mathrm{d}\mathcal{S}^2 &= \left[\left(\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}w}\right)^2 + \left(\frac{\mathrm{d}\mathcal{S}}{\mathrm{d}w}\right)^2\right](\mathrm{d}w)^2\\ &= \left[\cos^2\left(\frac{\pi}{2}w^2\right) + \sin^2\left(\frac{\pi}{2}w^2\right)\right](\mathrm{d}w)^2,\end{aligned}$$

i.e.

$$(\mathrm{d}l)^2 = (\mathrm{d}w)^2. \tag{20}$$

* See, for example, R. Courant, *Differential and Integral Calculus*, Vol. 1 (London and Glasgow, Blackie and Sons Ltd, 2nd edition, 1942), p. 496.

† A. Cornu, *Journ. de Phys.*, **3** (1874), 5, 44.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

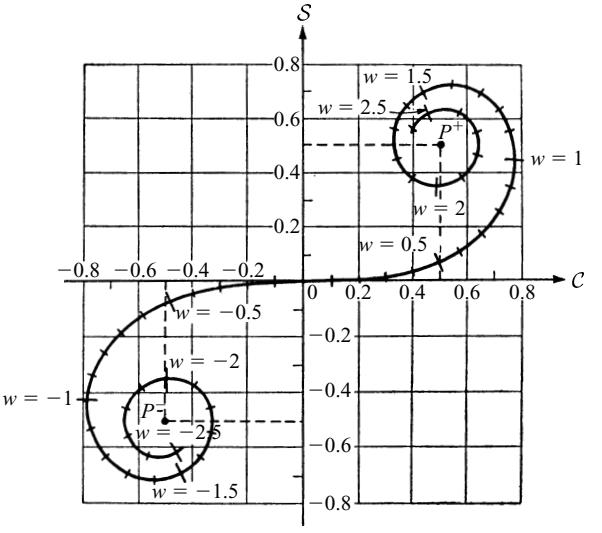


Fig. 8.37 The Cornu spiral.

Hence if l is measured in the sense of increasing w, the parameter w represents the length of arc of the curve measured from the origin.

Let θ be the angle which the tangent to the curve makes with the $\mathcal{C}$-axis. Then

$$\tan\theta = \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mathcal{C}} = \frac{\dfrac{\mathrm{d}\mathcal{S}}{\mathrm{d}w}}{\dfrac{\mathrm{d}\mathcal{C}}{\mathrm{d}w}} = \frac{\sin\left(\dfrac{\pi}{2}w^2\right)}{\cos\left(\dfrac{\pi}{2}w^2\right)} = \tan\left(\frac{\pi}{2}w^2\right)$$

i.e.

$$\theta = \frac{\pi}{2}w^2. \tag{21}$$

Thus θ increases monotonically with $|w|$. Since $\theta = 0$ when $w = 0$ the tangent touches the $\mathcal{C}$-axis at the origin. When $w^2 = 1$ then $\theta = \pi/2$, so that the tangent is then perpendicular to the $\mathcal{C}$ axis. When $w^2 = 2$, $\theta = \pi$ and the tangent is then parallel to the $\mathcal{C}$ axis again but is oriented in the negative direction. Since according to (18) and (19) $\mathcal{C}(\infty) = -\mathcal{C}(-\infty) = \frac{1}{2}$, $\mathcal{S}(\infty) = -\mathcal{S}(-\infty) = \frac{1}{2}$, the two branches of the curve approach the points P^+ and P^- with coordinates $(\frac{1}{2}, \frac{1}{2})$ and $(-\frac{1}{2}, -\frac{1}{2})$ respectively. This curve is known as the *Cornu spiral* (see Fig. 8.37) and is useful in discussions of the general properties of Fresnel diffraction patterns.

8.7.3 Fresnel diffraction at a straight edge

We now consider Fresnel diffraction at a semiinfinite plane bounded by a sharp straight edge. This problem is of special interest with regard to the behaviour of the field near the boundaries of geometrical shadows. We restrict our attention to the case where the line P_0P and also its projection (our x-axis) on to the half-plane is perpendicular to the

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

edge (Fig. 8.38). If x is the distance of the edge from the origin (which lies on the line P_0P), the integration extends throughout the region

$$-\infty < \xi < x, \qquad -\infty < \eta < \infty,$$

or, in terms of u and v,

$$-\infty < u < w, \qquad -\infty < v < \infty, \tag{22}$$

where

$$w = \sqrt{\frac{2}{\lambda}\left(\frac{1}{r'} + \frac{1}{s'}\right)}\, x \cos\delta. \tag{23}$$

The point of observation P lies in the illuminated region or in the geometrical shadow according as $x > 0$ or $x < 0$.

The diffraction integrals (9) become

$$\left.\begin{aligned} C &= b\int_{-\infty}^{w} \mathrm{d}u \int_{-\infty}^{+\infty} \mathrm{d}v \left[\cos\left(\frac{\pi}{2}u^2\right)\cos\left(\frac{\pi}{2}v^2\right) - \sin\left(\frac{\pi}{2}u^2\right)\sin\left(\frac{\pi}{2}v^2\right)\right], \\ S &= b\int_{-\infty}^{w} \mathrm{d}u \int_{-\infty}^{+\infty} \mathrm{d}v \left[\sin\left(\frac{\pi}{2}u^2\right)\cos\left(\frac{\pi}{2}v^2\right) + \cos\left(\frac{\pi}{2}u^2\right)\sin\left(\frac{\pi}{2}v^2\right)\right]. \end{aligned}\right\} \tag{24}$$

We have violated here a condition used in the derivation of (9), namely that the linear dimensions of the domain of integration shall be small compared to the distances P_0O and OP. To justify the approximate validity of these formulae also in the present case, a more careful discussion of the error terms is necessary. We shall omit it here, since the diffraction by a half-plane will be treated again later by rigorous methods (§11.5).

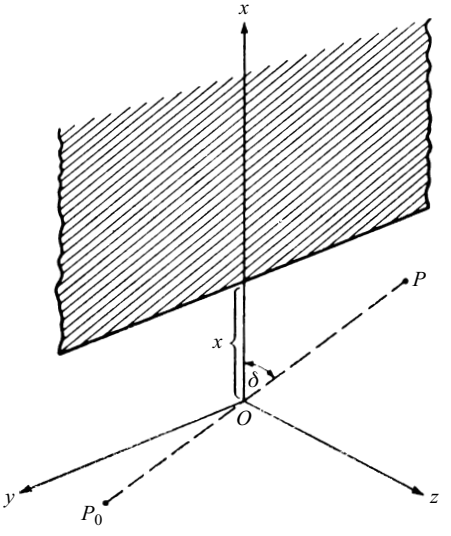


Fig. 8.38 Fresnel diffraction at a straight edge.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

From the relations (18) and (19), we have

$$\left.\begin{aligned}\int_{-\infty}^{w}\cos\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau &= \int_{-\infty}^{0}+\int_{0}^{w}=\mathcal{C}(\infty)+\mathcal{C}(w)=\tfrac{1}{2}+\mathcal{C}(w),\\ \int_{-\infty}^{+\infty}\cos\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau &= 1,\end{aligned}\right\} \tag{25}$$

and similarly

$$\left.\begin{aligned}\int_{-\infty}^{w}\sin\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau &= \tfrac{1}{2}+\mathcal{S}(w),\\ \int_{-\infty}^{+\infty}\sin\left(\frac{\pi}{2}\tau^2\right)\mathrm{d}\tau &= 1.\end{aligned}\right\} \tag{26}$$

Hence (24) becomes

$$\left.\begin{aligned}C &= b\{[\tfrac{1}{2}+\mathcal{C}(w)]-[\tfrac{1}{2}+\mathcal{S}(w)]\},\\ S &= b\{[\tfrac{1}{2}+\mathcal{C}(w)]+[\tfrac{1}{2}+\mathcal{S}(w)]\},\end{aligned}\right\} \tag{27}$$

and substitution into (4) gives finally an expression for the intensity:

$$I=\tfrac{1}{2}\{[\tfrac{1}{2}+\mathcal{C}(w)]^2+[\tfrac{1}{2}+\mathcal{S}(w)]^2\}I^{(0)}, \tag{28}$$

where

$$I^{(0)}=4|B|^2b^2=\frac{|A|^2}{(r'+s')^2}. \tag{29}$$

The behaviour of the intensity function (28) can be deduced from the Cornu spiral. The quantity $2I/I^{(0)}$ is seen to be equal to the square of the distance of the point w of the Cornu spiral from the 'asymptotic point' P^- $(-\frac{1}{2}, -\frac{1}{2})$. Thus if the point of observation is in the illuminated region $(w>0)$, $I/I^{(0)}$ oscillates with diminishing amplitudes as the distance from the edge increases and approaches asymptotically the value unity, as may be expected on the basis of geometrical optics. The maximum value of the intensity is not at the edge of the geometrical shadow, but some distance away from it, in the directly illuminated region (Fig. 8.39). On the edge of the shadow

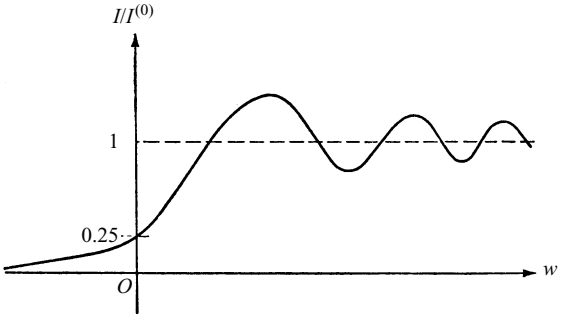


Fig. 8.39 Intensity distribution in the Fresnel diffraction pattern of a straight edge.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

($w = 0$), $I/I^{(0)} = \frac{1}{4}$. In the shadow region, $I/I^{(0)}$ decreases monotonically towards zero. These predictions are found to be in good agreement with experimental results.

8.8 The three-dimensional light distribution near focus

In §8.3.3 it was shown that the light distribution in the focal plane of a well-corrected lens arises essentially from Fraunhofer diffraction on the aperture of the lens, and in §8.5 the Fraunhofer diffraction patterns for apertures of various forms were studied in detail. To obtain a fuller knowledge of the structure of an optical image, we must study the light distribution not only in the geometrical focal plane but also in the neighbourhood of this plane. Knowledge of the three-dimensional (Fresnel) distribution near focus is of particular importance in estimating the tolerance in the setting of the receiving plane in an image-forming system.

The properties of the out-of-focus monochromatic images of a point source by a circular aperture were first discussed in detail by E. Lommel in a classical memoir.* Starting from the Huygens–Fresnel integral, Lommel succeeded in expressing the complex disturbance in terms of convergent series of Bessel functions and also confirmed experimentally the phenomena predicted on the basis of these calculations. Almost at the same time as Lommel, H. Struve† published a similar though less comprehensive analysis relating to the circular aperture. He did not work out the numerical consequences in such detail but gave useful approximations for the intensity near the edge of the geometrical shadow, where the series expansions are rather slowly convergent. Asymptotic approximations relating to points of observations at distances of many wavelengths from the focus were derived some years later by K. Schwarzschild.‡

The investigations of Lommel and Struve attracted relatively little attention, and in 1909 the problem was treated again by P. Debye,§ whose discussion established certain general features of the diffracted field both near and far away from the focus. In more recent times the analysis of these authors was extended and diagrams were published which show in detail the structure of the field in this complex region, and the results were broadly confirmed by experiment, both with light and with microwaves (short radio waves).‖

In discussing the light distribution near focus, we shall take as our starting point the analysis of Lommel and Struve, but it will be convenient and instructive to begin from the integral representation of the field in the form employed by Debye.

8.8.1 Evaluation of the diffraction integral in terms of Lommel functions

Consider a spherical monochromatic wave emerging from a circular aperture and converging towards the axial focal point O. We shall consider the disturbance $U(P)$ at

* E. Lommel, *Abh. Bayer. Akad.*, **15**, Abth. 2, (1885), 233; a later paper (*ibid.* **15**, Abth. 3, (1886), 531) deals with diffraction by a slit, by an opaque strip and by a straight edge.
† H. Struve, *Mém. de l'Acad. de St. Petersbourgh* (7), **34** (1886), 1.
‡ K. Schwarzschild, *Sitzb. München. Akad. Wiss., Math.-Phys. Kl.*, **28** (1898), 271.
§ P. Debye, *Ann. d. Physik.* (4), **30** (1909), 755.
‖ A comprehensive treatment of this subject is given in J. J. Stamnes, *Waves in Focal Regions* (Bristol, Adam Hilger, 1986).

a typical point P in the neighbourhood of O. The point P will be specified by a position vector $\mathbf{R}$ relative to O, and it will be assumed that the distance $R = OP$ as well as the radius $a(\gg \lambda)$ of the aperture are small compared to the radius $f = CO$ of the wave-front W that momentarily fills the aperture (Fig. 8.40).

If s denotes the distance from the point of observation P to a point Q on W and A/f is the amplitude at Q of the incident wave we have, by the application of the Huygens–Fresnel principle,

$$U(P) = -\frac{\mathrm{i}}{\lambda}\frac{A\mathrm{e}^{-\mathrm{i}kf}}{f}\iint_W \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\,\mathrm{d}S, \tag{1}$$

where, since only small angles are involved, the variation of the inclination factor over the wave-front has been neglected. If $\mathbf{q}$ denotes the unit vector in the direction OQ, we have, to a good approximation

$$s - f = -\mathbf{q}\cdot\mathbf{R}. \tag{2}$$

Also, the element $\mathrm{d}S$ of the wave-front is given by

$$\mathrm{d}S = f^2\mathrm{d}\Omega, \tag{3}$$

where $\mathrm{d}\Omega$ is the element of the solid angle that $\mathrm{d}S$ subtends at O. Moreover, we may replace s by f in the denominator of the integrand without introducing an appreciable error. Eq. (1) then becomes

$$U(P) = -\frac{\mathrm{i}}{\lambda}A\iint_\Omega \mathrm{e}^{-\mathrm{i}k\mathbf{q}\cdot\mathbf{R}}\,\mathrm{d}\Omega, \tag{4}$$

the integration now extending over the solid angle Ω which the aperture subtends at the focus. Eq. (4) is the *Debye integral* and expresses the field as a superposition of plane waves of different directions of propagation (specified by the vectors $\mathbf{q}$ which fill Ω).

Before discussing the evaluation of the Debye integral we note the interesting fact that, being a sum of elementary solutions (plane waves), it represents a rigorous solution of the wave equation which in the limit $f \to \infty$ (aperture at infinite distance) is valid throughout the whole of space. Of course, (4) is not a rigorous solution of our

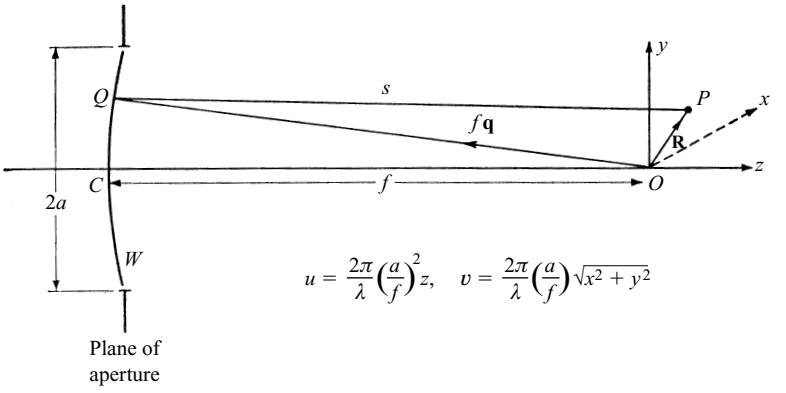


Fig. 8.40 Diffraction of a converging spherical wave at a circular aperture: notation.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

original problem, since no account has been taken of the nature of the screen, the exact boundary conditions being approximated by those of the Kirchhoff diffraction theory. The true solution to our problem would include not only contributions of plane waves that are propagated in the directions of the incident geometrical rays, but of waves propagated in all possible directions.* However, if $f \gg a \gg \lambda$ and if, in addition, $a^2/\lambda f \gg 1$, the Debye integral can be expected to give a good approximation to the light distribution in the region of focus.†

To evaluate (4) we first express the integrand in a more explicit form. We take Cartesian axes at O, with the z direction along Oz. Let (x, y, z) be the coordinates of P and (ξ, η, ζ) those of Q. We set

$$\left.\begin{aligned} \xi &= a\rho \sin\theta, \quad & x &= r\sin\psi, \\ \eta &= a\rho\cos\theta, \quad & y &= r\cos\psi. \end{aligned}\right\} \tag{5}$$

Since Q lies on the spherical wave-front W,

$$\zeta = -\sqrt{f^2 - a^2\rho^2} = -f\left[1 - \frac{1}{2}\frac{a^2\rho^2}{f^2} + \cdots\right]. \tag{6}$$

Then

$$\begin{aligned} \mathbf{q}\cdot\mathbf{R} &= \frac{x\xi + y\eta + z\zeta}{f} \\ &= \frac{a\rho r\cos(\theta - \psi)}{f} - z\left[1 - \frac{1}{2}\frac{a^2\rho^2}{f^2} + \cdots\right]. \end{aligned} \tag{7}$$

It is useful at this stage to introduce dimensionless variables u and v, which together with ψ specify the position of P:

$$u = \frac{2\pi}{\lambda}\left(\frac{a}{f}\right)^2 z, \qquad v = \frac{2\pi}{\lambda}\left(\frac{a}{f}\right) r = \frac{2\pi}{\lambda}\frac{a}{f}\sqrt{x^2 + y^2}. \tag{8}$$

We note that the point P lies in the direct beam of light or in the geometrical shadow according as $|v/u| \lesseqgtr 1$.

From (7) and (8) it follows that if terms above the second power in $a\rho/f$ are neglected in comparison to unity,

$$k\mathbf{q}\cdot\mathbf{R} = v\rho\cos(\theta - \psi) - \left(\frac{f}{a}\right)^2 u + \tfrac{1}{2}u\rho^2. \tag{9}$$

Further, the element of the solid angle is

$$\mathrm{d}\Omega = \frac{\mathrm{d}S}{f^2} = \frac{a^2\rho\,\mathrm{d}\rho\,\mathrm{d}\theta}{f^2}. \tag{10}$$

* This corresponds to the representation of the field in terms of a so-called angular spectrum of plane waves. See §11.4.2 and §13.2.1.

† Y. Li and E. Wolf, *Opt. Commun.*, **39** (1981), 205. The changes which the intensity distribution in the vicinity of the focus undergoes as the Fresnel number $N = a^2/\lambda f$ (see §8.2) is decreased are discussed by Y. Li and E. Wolf in *J. Opt. Soc. Amer. A*, **1** (1984), 801. The corresponding changes in the phase are discussed in Y. Li, *J. Opt. Soc. Amer. A*, **10** (1985), 1667. An experimental investigation of diffraction patterns in focusing systems of low Fresnel numbers was described by Y. Li and H. Platzer in *Optica Acta*, **30** (1983), 1621.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Hence (4) becomes

$$U(P) = -\frac{\mathrm{i}}{\lambda}\frac{a^2 A}{f^2}\mathrm{e}^{\mathrm{i}(\frac{f}{a})^2 u}\int_0^1\int_0^{2\pi}\mathrm{e}^{-\mathrm{i}[v\rho\cos(\theta-\psi)+\frac{1}{2}u\rho^2]}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta. \tag{11}$$

The integral with respect to θ is the same as one that we encountered in connection with Fraunhofer diffraction at a circular aperture (§8.5.2). It is equal to $2\pi J_0(v\rho)$, where $J_0(v\rho)$ is the Bessel function of zero order. Hence (11) becomes

$$U(P) = -\frac{2\pi\mathrm{i}a^2 A}{\lambda f^2}\mathrm{e}^{\mathrm{i}(\frac{f}{a})^2 u}\int_0^1 J_0(v\rho)\mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}\,\rho\,\mathrm{d}\rho. \tag{12}$$

It is convenient to consider separately the real and imaginary parts of the integral. We set

$$2\int_0^1 J_0(v\rho)\mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}\,\rho\,\mathrm{d}\rho = C(u,v) - \mathrm{i}S(u,v), \tag{13}$$

where

$$\left.\begin{aligned} C(u,v) &= 2\int_0^1 J_0(v\rho)\cos(\tfrac{1}{2}u\rho^2)\rho\,\mathrm{d}\rho,\\ S(u,v) &= 2\int_0^1 J_0(v\rho)\sin(\tfrac{1}{2}u\rho^2)\rho\,\mathrm{d}\rho. \end{aligned}\right\} \tag{14}$$

These integrals may be evaluated in terms of the *Lommel functions*

$$\left.\begin{aligned} U_n(u,v) &= \sum_{s=0}^{\infty}(-1)^s\left(\frac{u}{v}\right)^{n+2s}J_{n+2s}(v),\\ V_n(u,v) &= \sum_{s=0}^{\infty}(-1)^s\left(\frac{v}{u}\right)^{n+2s}J_{n+2s}(v), \end{aligned}\right\} \tag{15}$$

introduced by Lommel for this purpose.* Using the relation §8.5 (11)

$$\frac{\mathrm{d}}{\mathrm{d}x}[x^{n+1}J_{n+1}(x)] = x^{n+1}J_n(x),$$

$C(u,v)$ may be written as

$$\begin{aligned} C(u,v) &= \frac{2}{v}\int_0^1\frac{\mathrm{d}}{\mathrm{d}\rho}[\rho J_1(v\rho)]\cos(\tfrac{1}{2}u\rho^2)\mathrm{d}\rho\\ &= \frac{2}{v}\left[J_1(v)\cos\tfrac{1}{2}u + u\int_0^1\rho^2 J_1(v\rho)\sin(\tfrac{1}{2}u\rho^2)\mathrm{d}\rho\right], \end{aligned} \tag{16}$$

on integrating by parts. Again using the relation §8.5 (11), integrating by parts and continuing this process, we obtain

* For fuller discussions of these functions we refer to Lommel's memoirs (*loc. cit.*) and the following books: G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge, Cambridge University Press, 1922), pp. 537–550; A. Gray, G. B. Mathews and T. M. MacRobert, *A Treatise on Bessel Functions* (London, Macmillan, 2nd edition, 1922), Chapter XIV; and J. Walker, *The Analytical Theory of Light* (Cambridge, Cambridge University Press, 1904), p. 396.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$C(u,v)=\frac{\cos\frac{1}{2}u}{\frac{1}{2}u}\left[\left(\frac{u}{v}\right)J_1(v)-\left(\frac{u}{v}\right)^3J_3(v)+\cdots\right]$$
$$+\frac{\sin\frac{1}{2}u}{\frac{1}{2}u}\left[\left(\frac{u}{v}\right)^2J_2(v)-\left(\frac{u}{v}\right)^4J_4(v)+\cdots\right]$$
$$=\frac{\cos\frac{1}{2}u}{\frac{1}{2}u}U_1(u,v)+\frac{\sin\frac{1}{2}u}{\frac{1}{2}u}U_2(u,v). \tag{17a}$$

In a similar way we find

$$S(u,v)=\frac{\sin\frac{1}{2}u}{\frac{1}{2}u}U_1(u,v)-\frac{\cos\frac{1}{2}u}{\frac{1}{2}u}U_2(u,v). \tag{17b}$$

These formulae are valid at all points in the neighbourhood of the focus, but are only convenient for computations when $|u/v|<1$, i.e. when the point of observation lies in the geometrical shadow. When $|u/v|>1$, i.e. when the point of observation is in the illuminated region, it is more appropriate to use expansions involving positive powers of v/u. These may be derived in a similar manner by integrating by parts with respect to the trigonometric term. The first step gives

$$C(u,v)=\frac{2}{u}\int_0^1 J_0(v\rho)\frac{\mathrm{d}}{\mathrm{d}\rho}[\sin(\tfrac{1}{2}u\rho^2)]\mathrm{d}\rho$$
$$=\frac{2}{u}\left[J_0(v)\sin\tfrac{1}{2}u+v\int_0^1 J_1(v\rho)\sin(\tfrac{1}{2}u\rho^2)\mathrm{d}\rho\right], \tag{18}$$

where the relation §8.5 (15)

$$\frac{\mathrm{d}}{\mathrm{d}x}[x^{-n}J_n(x)]=-x^{-n}J_{n+1}(x)$$

has been used. Integrating by parts again and using the last relation together with the well-known formula (which may be deduced from the series expansion for $J_n(x)$)

$$\lim_{x\to0}\frac{J_n(x)}{x^n}=\frac{1}{2^n n!}, \tag{19}$$

we obtain

$$C(u,v)=\frac{\sin\frac{1}{2}u}{\frac{1}{2}u}\left[J_0(v)-\left(\frac{v}{u}\right)^2J_2(v)+\cdots\right]$$
$$-\frac{\cos\frac{1}{2}u}{\frac{1}{2}u}\left[\left(\frac{v}{u}\right)J_1(v)-\left(\frac{v}{u}\right)^3J_3(v)+\cdots\right]$$
$$+\frac{2}{u}\left[\frac{v^2}{2u}-\frac{1}{3!}\left(\frac{v^2}{2u}\right)^3+\cdots\right].$$

The series in the first two lines are two of the Lommel V_n functions and the series in the third line will be recognized as the expansion of $\sin(v^2/2u)$. Hence

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$C(u, v) = \frac{2}{u} \sin \frac{v^2}{2u} + \frac{\sin \frac{1}{2}u}{\frac{1}{2}u} V_0(u, v) - \frac{\cos \frac{1}{2}u}{\frac{1}{2}u} V_1(u, v). \tag{20a}$$

In a similar way we obtain for the other integral the expression

$$S(u, v) = \frac{2}{u} \cos \frac{v^2}{2u} - \frac{\cos \frac{1}{2}u}{\frac{1}{2}u} V_0(u, v) - \frac{\sin \frac{1}{2}u}{\frac{1}{2}u} V_1(u, v). \tag{20b}$$

This completes the formal solution of our problem. We shall now discuss some of the implications of these formulae.

8.8.2 The distribution of intensity

According to (12), (13), (17) and (20), the intensity $I = |U|^2$ in the neighbourhood of focus is given by the two equivalent expressions

$$I(u, v) = \left(\frac{2}{u}\right)^2 [U_1^2(u, v) + U_2^2(u, v)] I_0, \tag{21a}$$

and

$$I(u, v) = \left(\frac{2}{u}\right)^2 \left\{1 + V_0^2(u, v) + V_1^2(u, v) - 2V_0(u, v)\cos\left[\tfrac{1}{2}\left(u + \frac{v^2}{u}\right)\right] - 2V_1(u, v)\sin\left[\tfrac{1}{2}\left(u + \frac{v^2}{u}\right)\right]\right\} I_0, \tag{21b}$$

where

$$I_0 = \left(\frac{\pi a^2 |A|}{\lambda f^2}\right)^2 \tag{22}$$

is the intensity at the geometrical focus $u = v = 0$.

It follows from (15) that $U_1(-u, v) = -U_1(u, v)$, $U_2(-u, v) = U_2(u, v)$, $V_0(-u, v) = V_0(u, v)$, $V_1(-u, v) = -V_1(u, v)$. Accordingly $I(u, v)$ remains unchanged when u is replaced by $-u$. Hence, *in the neighbourhood of the focus the intensity distribution is symmetrical about the geometrical focal plane*.* Naturally, the distribution is also symmetrical about the axis $v = 0$.

From formulae (21), Lommel computed the intensity distribution in a number of selected receiving planes near focus, and verified experimentally some of the predictions.† The lines of equal intensity (called *isophotes*) near focus, constructed from Lommel's data are shown in Fig. 8.41.‡

* Symmetry properties of fields in the focal region, under more general conditions than considered here, were discussed by E. Collett and E. Wolf, *Opt. Lett.*, **5** (1980), 264.

† Related experiments were also described by M. E. Hufford and H. T. Davis, *Phys. Rev.*, **33** (1926), 589 and C. A. Taylor and B. J. Thompson, *J. Opt. Soc. Amer.*, **48** (1958), 844. Similar experiments with microwaves were carried out by M. P. Bachynski and G. Bekefi, *J. Opt. Soc. Amer.*, **47** (1957), 428, for the case of a circular aperture, and by P. A. Mathews and A. L. Cullen, *Proc. Inst. Elect. Engrs*, Pt. C, **103** (1956), 449, for a rectangular aperture.

‡ A similar, but less detailed, diagram was published by M. Berek, *Z. Phys.*, **40** (1926), 421. It is reproduced in some books with errors (incorrect position of the geometrical shadow and interchange of axes). Another

Footnote continued on page 490

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

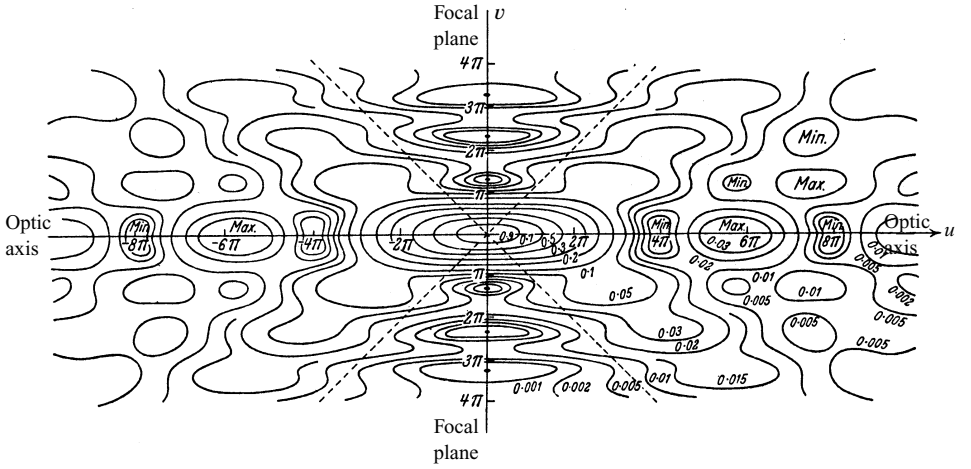


Fig. 8.41 Isophotes (contour lines of the intensity $I(u, v)$) in a meridional plane near focus of a converging spherical wave diffracted at a circular aperture. The intensity is normalized to unity at focus. The dashed lines represent the boundary of the geometrical shadow. When the figure is rotated about the u-axis, the minima on the v-axis generate the Airy dark rings. [Adapted from E. H. Linfoot and E. Wolf, *Proc. Phys. Soc.*, B, **69** (1956), 823.]

Of particular interest is the tubular structure of the bright central portion of the diffraction image seen clearly in the figure and already postulated on experimental grounds in 1894 by Taylor.* It is this structure that is responsible for the tolerance in the setting of a receiving plane in an image-forming system.

We shall now consider several special cases of interest.

(a) Intensity in the geometrical focal plane

For points in the geometrical focal plane, $u = 0$ and (21a) reduces to

$$I(0, v) = 4 \lim_{u \to 0} \left[\frac{U_1^2(u, v) + U_2^2(u, v)}{u^2} \right] I_0. \tag{23}$$

From the defining equation for the U_n functions it follows that

$$\lim_{u \to 0} \left[\frac{U_1(u, v)}{u} \right] = \frac{J_1(v)}{v}, \qquad \lim_{u \to 0} \left[\frac{U_2(u, v)}{u} \right] = 0, \tag{24}$$

version of the diagram was given by F. Zernike and B. R. A. Nijboer in their contribution to *La Théorie des Images Optiques* (Paris, Revue d'Optique, 1949), p. 227. Corresponding diagrams, calculated on the basis of electromagnetic theory, showing contours of the electric energy density and of the energy flow, were published by A. Boivin and E. Wolf, *Phys. Rev.*, **138** (1965), B 1561 and A. Boivin, J. Dow and E. Wolf, *J. Opt. Soc. Amer.*, **57** (1967), 1171.

Similar diagrams for the annular aperture were published by E. H. Linfoot and E. Wolf, *Proc. Phys. Soc.*, B, **66** (1953), 145. Some extensions of Lommel's analysis to diffraction by concentric arrays of ring-shaped apertures were discussed by A. Boivin, *J. Opt. Soc. Amer.*, **42** (1952), 60.

* H. D. Taylor, *Mon. Not. Roy. Astr. Soc.*, **54** (1894), 67.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

so that

$$I(0, v) = \left[\frac{2J_1(v)}{v}\right]^2 I_0. \tag{25}$$

We thus obtain the Airy formula §8.5 (14) for Fraunhofer diffraction at a circular aperture, as was to be expected.

(b) Intensity along the axis

For points on the axis, $v = 0$ and the two V_n functions entering the expression (21b) reduce to

$$V_0(u, 0) = 1, \qquad V_1(u, 0) = 0.$$

Hence

$$\begin{aligned} I(u, 0) &= \frac{4}{u^2}(2 - 2\cos\tfrac{1}{2}u)I_0 \\ &= \left(\frac{\sin u/4}{u/4}\right)^2 I_0. \end{aligned} \tag{26}$$

Thus the intensity along the axis is characterized by the function $[\sin(x)/x]^2$ which we discussed in §8.5.1 in connection with Fraunhofer diffraction at a rectangular aperture. The first zero of intensity on the axis is given by $u/4 \equiv \pi a^2 z/2\lambda f^2 = \pm\pi$, i.e. it is at a distance $z = \pm 2f^2\lambda/a^2$ from the focus.

It is usual to regard a loss of about 20 per cent in intensity at the centre of the image patch as permissible. Since $[\sin(u/4)/(u/4)]^2$ decreases by this amount when the receiving plane is displaced from the central position ($u = 0$) to $u \sim 3.2$, it follows that the focal tolerance Δz is approximately

$$\Delta z = \pm 3.2\frac{\lambda}{2\pi}\left(\frac{f}{a}\right)^2 \sim \pm\tfrac{1}{2}\left(\frac{f}{a}\right)^2\lambda. \tag{27}$$

With an $f/10$ pencil for example ($f/a = 20$), and with light of wavelength $\lambda = 5 \times 10^{-5}$ cm, the focal tolerance is about $\pm 0.5 \times 20^2 \times 5 \times 10^{-5}$ cm $= \pm 0.1$ mm.

(c) Intensity along the boundary of the geometrical shadow

For points on the boundary of the geometrical shadow $u = \pm v$. Since the distribution is symmetrical with respect to the geometrical focal plane we may, without loss of generality, take $u = +v$. The U_n functions then reduce to

$$U_1(u, u) = \sum_{s=0}^{\infty}(-1)^s J_{2s+1}(u), \qquad U_2(u, u) = \sum_{s=0}^{\infty}(-1)^s J_{2s+2}(u). \tag{28}$$

We recall the well-known identities of Jacobi*

* See for example G. N. Watson, *A Treatise on the Theory of the Bessel Functions* (Cambridge, Cambridge University Press, 1922), p. 22.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

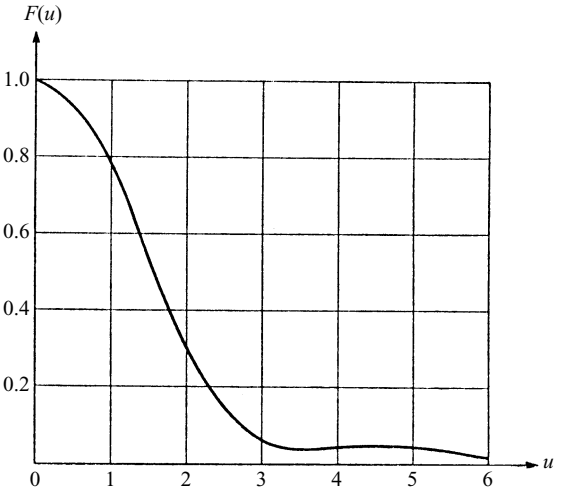


Fig. 8.42 The variation of intensity along the boundary of the geometrical shadow. The function

$$F(u) = \frac{1 - 2J_0(u)\cos u + J_0^2(u)}{u^2}.$$

$$\left.\begin{aligned} \cos(u\cos\theta) &= J_0(u) + 2\sum_{s=1}^{\infty}(-1)^s J_{2s}(u)\cos 2s\theta, \\ \sin(u\cos\theta) &= 2\sum_{s=0}^{\infty}(-1)^s J_{2s+1}(u)\cos(2s+1)\theta. \end{aligned}\right\} \tag{29}$$

Setting $\theta = 0$, it follows on comparison with (28) that

$$\left.\begin{aligned} U_1(u,\, u) &= \tfrac{1}{2}\sin u, \\ U_2(u,\, u) &= \tfrac{1}{2}[J_0(u) - \cos u], \end{aligned}\right\} \tag{30}$$

and (21a) reduces to

$$I(u,\, u) = \frac{1 - 2J_0(u)\cos u + J_0^2(u)}{u^2} I_0. \tag{31}$$

This function is shown in Fig. 8.42.

8.8.3 The integrated intensity

It is also desirable to determine the fraction L of the (time averaged) total energy that falls within a small circle of prescribed radius r_0 about the axial point in the receiving plane $u =$ constant. If

$$E = \pi a^2 \left(\frac{|A|}{f}\right)^2 \tag{32}$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

denotes the total energy incident on to the aperture in unit time, the required fraction of energy is given by

$$L(u, v_0) = \frac{1}{E}\int_0^{r_0}\int_0^{2\pi} I(u, v) r \, \mathrm{d}r \, \mathrm{d}\psi$$

$$= \frac{1}{2I_0}\int_0^{v_0} I(u, v) v \, \mathrm{d}v, \tag{33}$$

where

$$v_0 = \frac{2\pi}{\lambda}\left(\frac{a}{f}\right) r_0. \tag{34}$$

If Lommel's expressions (21) for the intensity are substituted into (33), the integral may be developed into series involving Bessel functions. The derivation is lengthy and we shall therefore only give the final results, due to Wolf.*

Again two formally different expressions are obtained, one of which is convenient for computations when the boundary of the small circle is in the geometrical shadow, the other when it is in the direct beam of light. Suppressing the suffix zero, i.e. writing v in place of v_0, one has in the first case ($|v/u| \geqslant 1$)

$$L(u, v) = 1 - \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1}\left(\frac{u}{v}\right)^{2s} Q_{2s}(v), \tag{35a}$$

where

$$Q_{2s}(v) = \sum_{p=0}^{2s}(-1)^p [J_p(v) J_{2s-p}(v) + J_{p+1}(v) J_{2s+1-p}(v)]. \tag{36}$$

In the second case ($|v/u| \leqslant 1$)

$$L(u, v) = \left(\frac{v}{u}\right)^2 \left[1 + \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1}\left(\frac{v}{u}\right)^{2s} Q_{2s}(v)\right]$$

$$- \frac{4}{u}\left[Y_1(u, v)\cos\tfrac{1}{2}\left(u + \frac{v^2}{u}\right) + Y_2(u, v)\sin\tfrac{1}{2}\left(u + \frac{v^2}{u}\right)\right], \tag{35b}$$

where the Q's are again given by (36) and Y_1 and Y_2 are two of the functions

$$Y_n(u, v) = \sum_{s=0}^{\infty}(-1)^s (n+2s)\left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v)$$

$$= \tfrac{1}{2}\left[\frac{v^2}{u} V_{n-1}(u, v) + u V_{n+1}(u, v)\right]. \tag{37}$$

In Fig. 8.43, the contour lines of $L(u, v)$, computed from these formulae are displayed. They may be regarded as analogous in a certain sense to the rays of geometrical optics.

* E. Wolf, *Proc. Roy. Soc.*, A, **204** (1951), 533. Asymptotic approximations for $L(u, v)$ were derived by J. Focke, *Optica Acta*, **3** (1956), 161.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

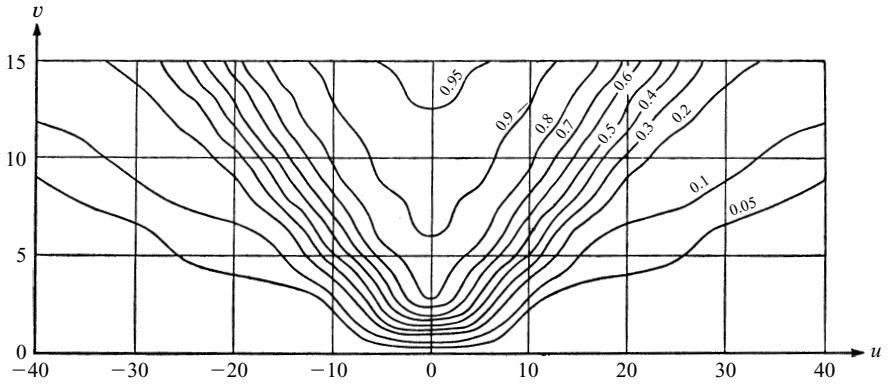


Fig. 8.43 Contour lines of $L(u, v)$, giving the fraction of the total energy which falls within small circles centred on the axis in selected receiving planes $u =$ constant. (After E. Wolf, *Proc. Roy. Soc.*, A, **204** (1951), 542.)

We note that in the special case when the receiving plane coincides with the focal plane ($u = 0$), (35a) reduces to

$$L(0, v) = 1 - Q_0(v)$$
$$= 1 - J_0^2(v) - J_1^2(v), \tag{38}$$

in agreement with Rayleigh's formula §8.5 (18).

Of special interest is the case when the circle over which the integral (33) is extended coincides with the cross-section of the geometrical cone of rays. Then $|v/u| = 1$ and the series in (35a) and (35b) can be summed, and give*

$$L(u, u) = 1 - J_0(u)\cos u - J_1(u)\sin u. \tag{39}$$

Hence the expression

$$\varepsilon(u) = J_0(u)\cos u + J_1(u)\sin u \tag{40}$$

gives, for the receiving plane $u =$ constant, the fraction of the total energy in the geometrical shadow. The function $\varepsilon(u)$ is plotted in Fig. 8.44; it is not monotonically decreasing but has maxima (apart from $u = 0$) when $J_1(u) = 0$ and minima when $\sin u = 0$ ($u \neq 0$).

8.8.4 The phase behaviour

Finally we consider the behaviour of the phase of the disturbance in the neighbourhood of the focus. According to (12) and (13) the phase is, apart from the additive term $(-\omega t)$ given by†

* E. Wolf, *loc. cit.*, p. 539.

† The symbol mod 2π on the right of an equation denotes that the two sides of the equation are indeterminate to the extent of an additive constant $2m\pi$ where m is any integer.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

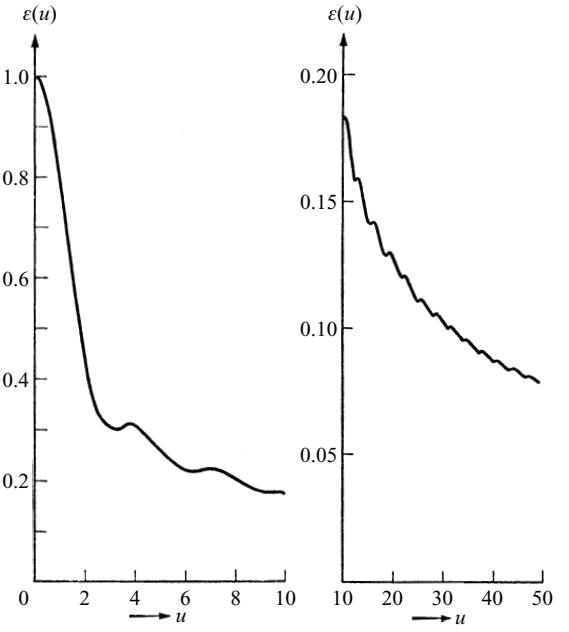


Fig. 8.44 The fraction $\varepsilon(u)$ of the total energy within the geometrical shadow. (After E. Wolf, *Proc. Roy. Soc.*, A, **204** (1951), 544.)

$$\phi(u, v) = \left(\frac{f}{a}\right)^2 u - \chi(u, v) - \frac{\pi}{2} \qquad (\mathrm{mod}\, 2\pi), \tag{41}$$

where

$$\cos\chi = \frac{C}{\sqrt{C^2 + S^2}}, \qquad \sin\chi = \frac{S}{\sqrt{C^2 + S^2}}, \tag{42}$$

the positive square root being taken in (42).

We note that, unlike the intensity distribution, the phase distribution cannot be expressed in terms of u and v alone, but has a structure which depends on the angular aperture of the geometrical pencil of rays. It also follows that each 'branch' of the multivalued function $\phi(u, v)$ is continuous in u and v at all points where the intensity does not vanish, and that at points of zero intensity it is indeterminate. At the focus $u = v = 0$, one of its values is $-\pi/2$.

The cophasal surfaces (surfaces $\phi =$ constant) are, of course, surfaces of revolution about the u-axis. We shall now show that they possess a further symmetry, expressed by the relation

$$\phi(-u, v) + \phi(u, v) = -\pi \qquad (\mathrm{mod}\, 2\pi). \tag{43}$$

From (14),

$$C(-u, v) = C(u, v), \qquad S(-u, v) = -S(u, v). \tag{44}$$

From (42) it then follows that

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$\cos\chi(-u, v) = \cos\chi(u, v), \qquad \sin\chi(-u, v) = -\sin\chi(u, v), \tag{45}$$

so that

$$\chi(-u, v) = -\chi(u, v). \tag{46}$$

The relation (43) now follows from (46) and (41). Hence the reflection in the plane $u = 0$ of any cophasal surface $\phi = \phi_0$ is a cophasal surface $\phi = -\pi - \phi_0$.

In Fig. 8.45 the profiles of the cophasal surfaces of an $f/2$ homocentric pencil are shown. Far away from the focus these surfaces coincide with the spherical wave-fronts of geometrical optics, but they become more and more deformed as the focal region is approached. The profiles in the immediate neighbourhood of the geometrical focal plane are shown in Fig. 8.46. It is seen that close to the focus, the surfaces are substantially plane; however, it may be shown that the spacing between them is different from that appropriate to a parallel beam of light of the same wavelength. Moreover (see Fig. 8.41), the intensity is not uniform over each cophasal surface. In the immediate neighbourhood of the Airy dark rings (such as R_1 and R_2 in Fig. 8.46) the cophasal surfaces show a more complicated behaviour.* Proceeding from the geometrical focus in the v direction, the phase is seen to be constant between any two successive dark rings, but changes abruptly by π on crossing each ring. That this must be so may be seen from the expression (derived easily with the help of (24), (41) and (46)) for the complex disturbance in the geometrical focal plane:

$$U(0, v) = \frac{1}{i}\frac{2J_1(v)}{v}\sqrt{I_0}. \tag{47}$$

Since this expression is purely imaginary for all values of v, the phase of $U(0, v)$ can only be equal to $-\pi/2$ or $+\pi/2$ (mod 2π), and since it changes sign on crossing each

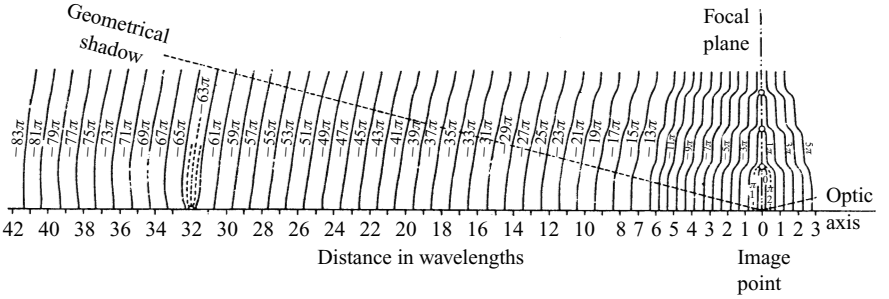


Fig. 8.45 Profiles of the cophasal surfaces $\phi(u, v) =$ constant near the geometrical focus, calculated with $\lambda = 5 \times 10^{-5}$ cm, $a = 2.5$ cm, $f = 10$ cm. (After G. W. Farnell, *Canad. J. Phys.*, **35** (1957), 780.)†

* It is an example of so-called wave-front dislocations which occur in the neighbourhood of points where the field amplitude is zero and consequently the phase is singular. For a discussion of such phenomena see J. F. Nye and M. V. Berry, *Proc. Roy. Soc. Lond.* A, **336** (1974), 165.

† The values associated with the surfaces of Fig. 8.45 differ by π from those of Farnell's paper. This is so because, for the sake of consistency with our analysis, the phase at the focus is taken as $-\pi/2$ in accordance with (51), whereas in the paper by Farnell it is taken as $+\pi/2$.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

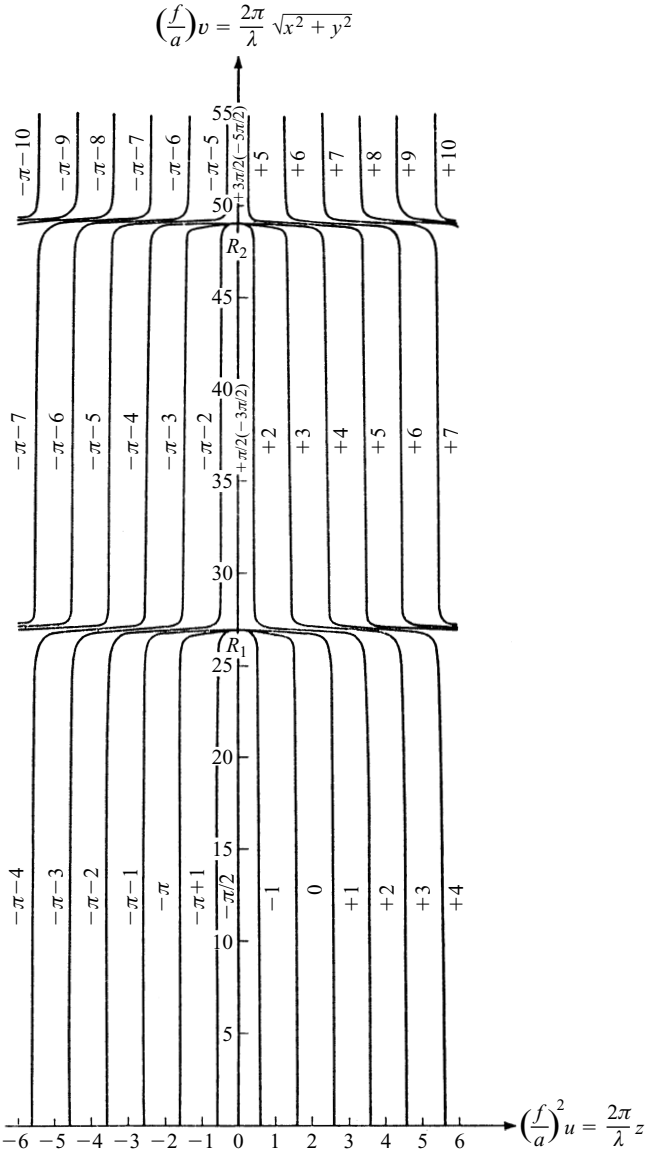


Fig. 8.46 Profiles of the cophasal surfaces in the immediate neighbourhood of the geometrical focal plane of an $f/3.5$ homocentric pencil. OR_1 and OR_2 are the radii of the first and second Airy dark rings. [After E. H. Linfoot and E. Wolf, *Proc. Phys. Soc.*, B, **69** (1956), 827.]

dark ring, the phase must then undergo a sudden jump by π. A discontinuity of this amount also occurs on crossing each axial point of zero intensity.

It is also of interest to study how the phase of the field changes as the point of observation moves along each ray through the focus. It is convenient to compare this variation with that of a spherical wave converging to the focus in the half-space $z < 0$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

and diverging from it in the half-space $z > 0$. If the phase $\tilde{\phi}$ of this comparison wave is taken as zero at the focus we have

$$\left.\begin{aligned} \tilde{\phi}(u, v) = -kR \quad &\text{when} \quad u < 0 \\ +\, kR \quad &\text{when} \quad u > 0 \end{aligned}\right\} \tag{48}$$

where as before $R = \sqrt{x^2 + y^2 + z^2} > 0$ is the distance of the point of observation from the focus. The difference

$$\delta(u, v) = \phi(u, v) - \tilde{\phi}(u, v) \tag{49}$$

is called the *phase anomaly.*

From (43), (48) and (49) it follows that

$$\delta(-u, v) + \delta(u, v) = -\pi \qquad (\mathrm{mod}\, 2\pi), \tag{50}$$

whilst at the focus itself we have

$$\delta(0, 0) = \phi(0, 0) = -\frac{\pi}{2} \qquad (\mathrm{mod}\, 2\pi). \tag{51}$$

The behaviour of the phase anomaly along selected rays through the focus of an $f/3.5$ homocentric pencil is shown in Fig. 8.47.

The figures show that on passing through the focus along any ray except the axial one, δ undergoes a rapid but continuous change in phase of π. This effect was observed a long time ago by Gouy* and has been the subject of many investigations.† Along the axis, however, the phase anomaly has a singular behaviour: it fluctuates periodically between the values 0 and $-\pi$.

It can be shown by considering the asymptotic approximation to the Huygens–Fresnel integral for large values of k (small wavelengths) that as light advances along a ray the phase changes suddenly by $\pi/2$ on passing through either of the two principal centres of curvature of the associated wave-fronts.‡ The case which we have just considered corresponds to the special situation when the two centres of curvature coincide. Thus the change by half a period is associated even with the geometrical optics solution.§ Fig. 8.47 shows how the discontinuity of geometrical optics goes over into a continuous transition when the finiteness of the wavelength is taken into account. Finally a reference may be made to a paper by Farnell,‖ which describes an experimental investigation into the structure of the phase distribution in the focal region of a microwave lens; very good agreement with the predictions of the theory was found.

* L. G. Gouy, *Compt. Rend. Acad. Sci. Paris*, **110** (1890), 1251; *Ann. Chim. (Phys.)* (6), **24** (1891), 145.

† For a survey of the literature see F. Reiche, *Ann. d. Physik*, (4), **29** (1909), 56, *ibid*., 401. References to some later investigations are given in E. H. Linfoot and E. Wolf, *Proc. Phys. Soc.*, B, **69** (1956), 827.

‡ H. Poincaré, *Théorie mathématique de la lumière*, Vol. II (Paris, Georges Carré, 1892), pp. 168–174. See also J. Walker, *The Analytical Theory of Light* (Cambridge, Cambridge University Press, 1904), pp. 91–93.

§ For discussions of this point see also A. Rubinowicz, *Phys. Rev.*, **54** (1938), 931 and C. J. Bouwkamp, *Physica*, **7** (1940), 485.

‖ G. W. Farnell, *Canad. J. Phys.*, **36** (1958), 935.

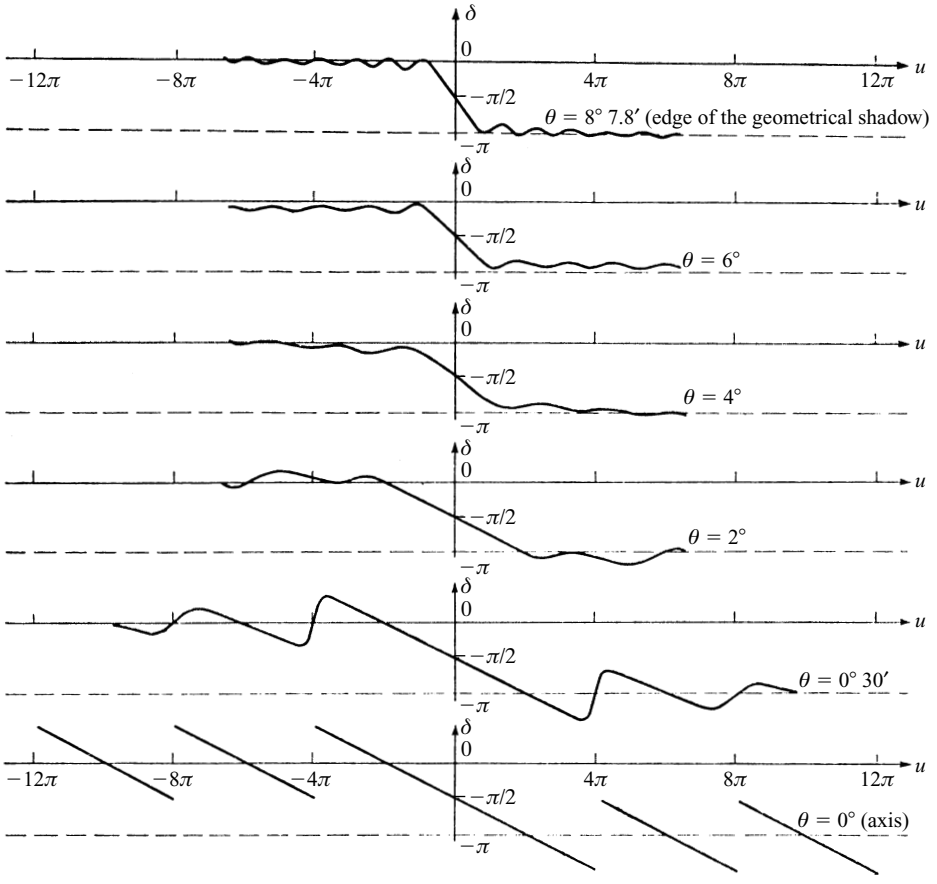


Fig. 8.47 Phase anomaly δ along geometrical rays through the focus of a homocentric $f/3.5$ pencil. The angle θ denotes the inclination of the ray to the axis. [After E. H. Linfoot and E. Wolf, *Proc. Phys. Soc.*, B, **69** (1956), 827.]

8.9 The boundary diffraction wave

If the edge of a diffracting aperture or obstacle is observed from points within the geometrical shadow, it appears luminous. This result was already known to Thomas Young* who attempted, prior to Fresnel, to explain diffraction on a wave-theoretical basis. Young believed that the incident light undergoes a kind of reflection at the edge of the diffracting body, and he regarded the diffraction pattern as arising from the interference of the incident wave and the reflected 'boundary wave'. However, Young's views were expressed in a qualitative manner only and did not gain much recognition.

That Young's theory contained an element of truth became evident after Sommerfeld in 1894 obtained a rigorous solution for the diffraction of plane waves by a plane, semiinfinite reflecting screen (see §11.5). This solution shows that in the geometrical shadow the light is propagated in the form of a cylindrical wave that appears to

* T. Young, *Phil. Trans. Roy. Soc.*, **92** (1802), 26.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

proceed from the edge of the screen, whilst in the illuminated region it is represented as superposition of the cylindrical wave and of the original incident wave.

The question arises whether also under more general conditions diffraction can be accounted for as the combined effect of an incident wave and a boundary wave. This problem had been investigated before the appearance of Sommerfeld's paper by Maggi,* but his results appear to have been forgotten. It was later investigated independently and much more fully by Rubinowicz.† The Maggi–Rubinowicz theory was developed further by Miyamoto and Wolf.‡

Consider a monochromatic light wave from a point source P_0 propagated through an aperture in a plane opaque screen. As before we assume that the linear dimensions of the aperture are large compared with the wavelength but small compared to the distance of P_0 and of the point of observation P from the screen, and that the angles of incidence and of diffraction are small. Then we have in the approximations of the Kirchhoff theory (§8.3.2)

$$U(P) = \frac{1}{4\pi}\iint_{\mathcal{A}} \left[\frac{e^{ikr}}{r}\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right) - \left(\frac{e^{iks}}{s}\right)\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right] dS, \tag{1}$$

where $\mathcal{A}$ denotes the diffracting aperture and the other symbols have the same meaning as before. We construct a closed surface bounded by (1) the opening $\mathcal{A}$, (2) the surface of a truncated cone $\mathcal{B}$ whose vertex is at P_0 and whose generators pass through the edge of the aperture, and (3) a portion $\mathcal{C}$ of a large sphere centred on P (Fig. 8.48). If R denotes the distance from P_0 to P, we have rigorously from Kirchhoff's integral theorem,

$$\frac{1}{4\pi}\iint_{\mathcal{A}+\mathcal{B}+\mathcal{C}} \left[\frac{e^{ikr}}{r}\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right) - \frac{e^{iks}}{s}\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right] dS = \frac{e^{ikR}}{R} \quad \text{or} \quad 0, \tag{2}$$

according to the point P lying inside or outside the surface. Now in the same way as in §8.3.2, the contribution from $\mathcal{C}$ can be made negligible by taking the radius of the sphere sufficiently large. We then obtain from (1) and (2),

$$U(P) = U^{(g)}(P) + U^{(d)}(P), \tag{3}$$

* G. A. Maggi, *Annali di Matem.* (2), **16** (1888), 21. Maggi's analysis is also discussed in a paper by F. Kottler, *Ann. d. Physik*, (4), **70** (1923), 413; and in B. B. Baker and E. T. Copson, *The Mathematical Theory of Huygens' Principle* (Oxford, Clarendon Press, 2nd edition, 1950), p. 74.

Experimental evidence for the 'existence' of the boundary wave was found by W. Wien, *Inaug. Diss., Berlin*, 1886; E. Maey, *Ann. d. Physik.* (9), **49** (1893), 69; and A. Kalaschnikow, *Journ. Russ. Phys. Chem. Ges.*, **44** (1912), *Phys. Teil*, 133. See also S. Banerji, *Phil. Mag.* (6), **37** (1919), 112; and S. K. Mitra, *ibid.* (6), **38** (1919), 289.

† A. Rubinowicz, *Ann. d. Physik.* (4), **53** (1917), 257; *ibid.* (4), **73** (1924); *ibid.*, **81** (1926), 153; *Acta Phys. Polonica*, **12** (1953), 225. See also G. N. Ramachandran, *Proc. Indian Acad. Sci. A*, **21** (1945), 165; L. C. Martin, *Proc. Phys. Soc.*, **55** (1943), 104; *ibid.*, **62**B (1949), 713. Y. V. Kathavate, *ibid. A*, **21** (1945), 177; R. S. Ingarden, *Acta Phys. Polon.*, **14** (1955), 77; O. Laporte and J. Meixner, *Z. Phys.*, **153** (1958), 129.

A very comprehensive account of researches relating to the boundary wave is given in A. Rubinowicz' book *Die Beugungswelle in der Kirchhoffschen Theorie der Beugung* (Berlin, Springer, 2nd ed., 1966).

‡ K. Miyamoto and E. Wolf, *J. Opt. Soc. Amer.*, **52** (1962) 615, 626; K. Miyamoto, *Proc. Phys. Soc.*, **79** (1962), 617. See also E. W. Marchand and E. Wolf, *J. Opt. Soc. Amer.*, **52** (1962), 761; A. Rubinowicz, *ibid.*, **52** (1962) 717; *Acta Phys. Polonica* **21** (1962), 61, 451; *Progress in Optics*, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 199.

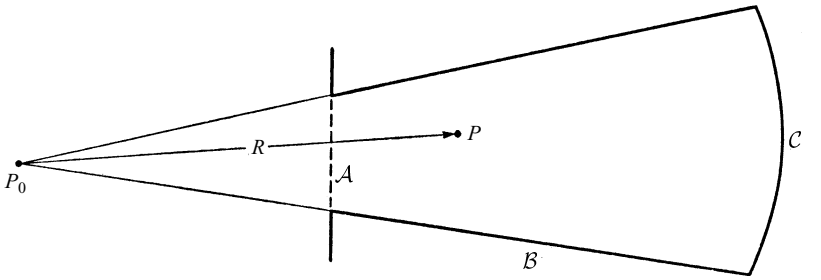


Fig. 8.48 Derivation of the boundary diffraction wave.

where

$$\left.\begin{aligned} U^{(g)}(P) &= \frac{e^{ikR}}{R} \quad \text{when } P \text{ is in the direct beam} \\ &= 0 \qquad \text{when } P \text{ is in the geometrical shadow} \end{aligned}\right\} \tag{4}$$

and

$$U^{(d)}(P) = -\frac{1}{4\pi}\iint_{\mathcal{B}}\left[\frac{e^{ikr}}{r}\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right) - \left(\frac{e^{iks}}{s}\right)\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right]dS. \tag{5}$$

$U^{(g)}$ represents the disturbance as predicted by geometrical optics, so that $U^{(d)}$ must represent the effect of diffraction. We shall now show that $U^{(d)}$ may be transformed into a line integral along the edge of the aperture.

We note first that the spheres $r =$ constant cut orthogonally the truncated cone $\mathcal{B}$. Hence on $\mathcal{B}$,

$$\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) = 0. \tag{6}$$

Also

$$\frac{d}{dn}\left(\frac{e^{iks}}{s}\right) = \frac{d}{ds}\left(\frac{e^{iks}}{s}\right)\cos(n, s) = \left(\frac{ik}{s} - \frac{1}{s^2}\right)e^{iks}\cos(n, s). \tag{7}$$

Hence (5) reduces to

$$U^{(d)}(P) = -\frac{1}{4\pi}\iint_{\mathcal{B}}\frac{e^{ik(r+s)}}{rs}\left(ik - \frac{1}{s}\right)\cos(n, s)dS. \tag{8}$$

We can take as the element dS the area $ABB'A'$ bounded by segments of two neighbouring generators and by the arcs of circles in which the spheres $r =$ constant and $r + dr =$ constant intersect the cone (Fig. 8.49). If $d\phi$ is the angle between the two generators, then

$$dS = r\,dr\,d\phi. \tag{9}$$

Let Q and Q' be the points of intersection of the two generators with the edge Γ of the aperture and let dl be the length of the element of Γ between these two points. If

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

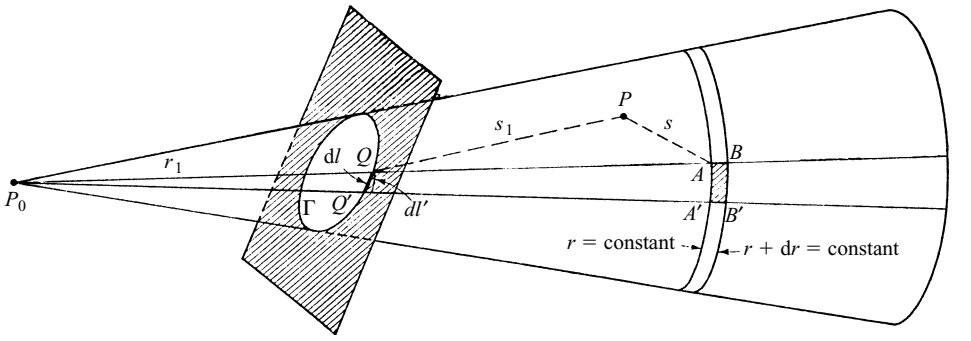


Fig. 8.49 Derivation of the boundary diffraction wave:

$$P_0A = r, \qquad P_0Q = r_1, \qquad PA = s, \qquad PQ = s_1.$$

$\mathrm{d}l'$ denotes the corresponding element of arc of the circle in which the sphere with radius $r_1 = P_0Q$ intersects the cone then

$$\mathrm{d}l' = r_1 \mathrm{d}\phi = \mathrm{d}l \cos(\mathrm{d}l, \mathrm{d}l') = \mathrm{d}l \sin(r_1, \mathrm{d}l). \tag{10}$$

From (9) and (10)

$$\mathrm{d}S = \frac{r}{r_1} \sin(r_1, \mathrm{d}l)\mathrm{d}r\,\mathrm{d}l. \tag{11}$$

Also, by projection on the normal to the cone at A and Q (the normals at these points being parallel),

$$s \cos(n, s) = s_1 \cos(n, s_1). \tag{12}$$

On substituting from (11) and (12) into (8) it follows that

$$\begin{aligned} U^{(d)}(P) &= -\frac{1}{4\pi}\int\!\!\int_{\mathcal{B}} \frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{rs}\left(\mathrm{i}k - \frac{1}{s}\right)\frac{s_1}{s}\cos(n, s_1)\frac{r}{r_1}\sin(r_1, \mathrm{d}l)\mathrm{d}r\,\mathrm{d}l \\ &= -\frac{1}{4\pi}\int_{\Gamma} \mathrm{d}l \frac{s_1}{r_1}\cos(n, s_1)\sin(r_1, \mathrm{d}l)\int_{r_1}^{\infty} \mathrm{e}^{\mathrm{i}k(r+s)}\left(\frac{\mathrm{i}k}{s^2} - \frac{1}{s^3}\right)\mathrm{d}r. \end{aligned} \tag{13}$$

Next we shall show that the integrand of the second integral in (13) is a total differential, namely that

$$\mathrm{e}^{\mathrm{i}k(r+s)}\left[\frac{\mathrm{i}k}{s^2} - \frac{1}{s^3}\right] = \frac{\mathrm{d}}{\mathrm{d}r}\left\{\frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{s[s + r - r_1 + s_1\cos(s_1, r_1)]}\right\}. \tag{14}$$

We have, on carrying out the differentiation on the right of (14),

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$\frac{\mathrm{d}}{\mathrm{d}r}\left\{\frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{s[s+r-r_1+s_1\cos(s_1,\,r_1)]}\right\}$$
$$=\frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{s[s+r-r_1+s_1\cos(s_1,\,r_1)]}\left\{\mathrm{i}k\left(1+\frac{\mathrm{d}s}{\mathrm{d}r}\right)-\frac{1}{s}\frac{\mathrm{d}s}{\mathrm{d}r}\right.$$
$$\left.-\frac{1}{s+r-r_1+s_1\cos(s_1,\,r_1)}\left(1+\frac{\mathrm{d}s}{\mathrm{d}r}\right)\right\}. \tag{15}$$

Now from the triangle APQ

$$s^2=s_1{}^2+(r-r_1)^2+2s_1(r-r_1)\cos(s_1,\,r_1), \tag{16}$$

whence, differentiating with respect to r whilst keeping r_1 and s_1 fixed,

$$s\frac{\mathrm{d}s}{\mathrm{d}r}=r-r_1+s_1\cos(s_1,\,r_1). \tag{17}$$

On substituting from this equation into (15), the identity (14) follows. Hence

$$\int_{r_1}^{\infty}\mathrm{e}^{\mathrm{i}k(r+s)}\left(\frac{\mathrm{i}k}{s^2}-\frac{1}{s^3}\right)\mathrm{d}r=\left[\frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{s[s+r-r_1+s_1\cos(s_1,\,r_1)]}\right]_r^{\infty}$$
$$=-\frac{\mathrm{e}^{\mathrm{i}k(r_1+s_1)}}{s_1{}^2[1+\cos(s_1,\,r_1)]}, \tag{18}$$

and (13) finally becomes

$$U^{(d)}(P)=\frac{1}{4\pi}\oint_{\Gamma}\frac{\mathrm{e}^{\mathrm{i}k(r_1+s_1)}}{r_1s_1}\cdot\frac{\cos(n,\,s_1)}{[1+\cos(s_1,\,r_1)]}\sin(r_1,\,\mathrm{d}l)\mathrm{d}l. \tag{19}$$

This formula, together with (3) and (4), is the *Rubinowicz representation* of the Kirchhoff diffraction integral and may be regarded as the mathematical formulation of Young's theory. It expresses the effect of diffraction in terms of the incident wave that is propagated in accordance with the laws of geometrical optics and a boundary diffraction wave that may be thought of as arising from the scattering of the incident radiation by the boundary of the aperture.

Since U is a continuous function of position, it follows from (4) that the boundary wave $U^{(d)}$ is discontinuous across the edge of the geometrical shadow so as to compensate for the discontinuity in the 'geometrical wave' $U^{(g)}$. The discontinuity in $U^{(d)}$ arises from the factor $[1+\cos(s_1,\,r_1)]$ in the denominator in (19).

By the same argument which was given in connection with the classification of diffraction phenomena on pp. 429–430 and which illustrates the physical significance of the principle of the stationary phase (see Appendix III), it follows that only those points of the domain of integration contribute substantially to $U^{(d)}$ for which the phase of the integrand is stationary, i.e. for which

$$\frac{\mathrm{d}}{\mathrm{d}l}[k(r_1+s_1)]=0. \tag{20}$$

This relation may also be written in the form of a 'reflection law',

$$\cos(r_1,\,\mathrm{d}l)=-\cos(s_1,\,\mathrm{d}l). \tag{21}$$

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

8.10 Gabor's method of imaging by reconstructed wave-fronts (holography)

In an attempt to improve the resolving power of the electron microscope, Gabor* proposed a two-step method of optical imagery. In the first step an object is illuminated with a coherent electron wave or a coherent light wave. The object is assumed to be such that a considerable part of the wave penetrates undisturbed through it.† A diffraction pattern, called a *hologram*, which is formed by the interference of the secondary wave arising from the presence of the object with the strong background wave, is recorded on a photographic plate. If the plate, suitably processed, is replaced in the original position and is illuminated by the background wave alone, the wave that is transmitted by the plate contains information about the original object, and this can be extracted from the photograph by optical processes. In order to 'reconstruct' the object from this 'substitute' wave, it is only necessary to send it through a suitable image-forming system, and an image will appear in the plane conjugate to the plane in which the object was situated. We shall only be concerned with the optical principle involved‡.

8.10.1 Producing the positive hologram

Consider a monochromatic wave from a small source S, impinging on a semitransparent object σ (Fig. 8.50(a)). Let $\mathcal{H}$ be a screen some distance behind the object and let $U = Ae^{i\psi}$ represent the complex disturbance at a typical piont of $\mathcal{H}$, A being the (real) amplitude, and ψ the phase of the disturbance. We may regard U as the sum of two terms,

$$U = U^{(i)} + U^{(s)} = e^{i\psi_i}(A^{(i)} + A^{(s)}e^{i(\psi_s - \psi_i)}). \tag{1}$$

Here $U^{(i)} = A^{(i)}e^{i\psi_i}$ denotes the *incident wave* (or *coherent background*); it is the field which would be produced at $\mathcal{H}$ in the absence of the object. The other term, $U^{(s)} = A^{(s)}e^{i\psi_s}$ represents the *secondary*, or *diffracted*, *wave* and it is this wave that contains information about the object. In terms of the amplitudes and phases, the amplitude of the total disturbance U may according to (1) be written as

$$A = \sqrt{UU^\star} = \sqrt{A^{(i)2} + A^{(s)2} + 2A^{(i)}A^{(s)}\cos(\psi_s - \psi_i)}. \tag{2}$$

As usual, we have suppressed the time harmonic factor $e^{-i\omega t}$, and thus have implicitly assumed that the secondary wave issuing from the object is of the same frequency as the incident wave. An object of this type is called a *Rayleigh scatterer*; to a good approximation, practically all nonfluorescent objects are of this type.

* D. Gabor, *Nature*, **161** (1948), 777; *Proc. Roy. Soc.*, A, **197** (1949), 454; *Proc. Phys. Soc.*, B, **64** (1951), 449.

† This means that, if the wave is expressed as the sum of the incident wave and a diffracted secondary wave, the scattering of the secondary wave is neglected. This neglect represents what is usually known as Born's first-order approximation, and finds many applications in the theory of scattering of light, X-rays and electrons. It is discussed in §13.1.2. For the limits of validity of this approximation in electron microscopy and interferometry see D. Gabor, *Rev. Mod. Phys.*, **28** (1956), 260.

‡ For a detailed treatment of this subject and for an account of various modifications and applications of this technique, see for example, R. J. Collier, C. B. Burckhardt and L. H. Lin, *Optical Holography* (San Diego, CA, Academic Press, 1971) or P. Hariharan, *Optical Holography* (Cambridge, Cambridge University Press, 2nd edition, 1996).

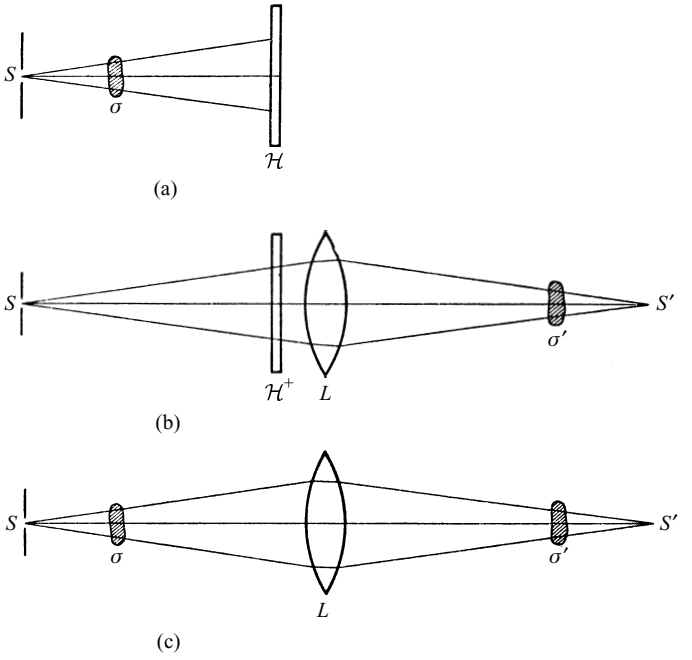


Fig. 8.50 Illustrating Gabor's method of imaging by reconstructed wave-fronts: (a) formation of the hologram; (b) reconstruction; (c) equivalent one-stage imaging.

Suppose now that a photographic plate is placed in the $\mathcal{H}$-plane. Let α be the *transmission factor* of the plate, defined by analogy with the transmission function F of §8.6.1 as the ratio of the complex amplitude of the wave emerging from the plate to that of the wave incident on the plate. The corresponding transmission factor for the intensity is $\tau = \alpha\alpha^\star$; and the quantity

$$D = -\log_{10}\tau = -\log_{10}\alpha\alpha^\star \tag{3}$$

is called the *density* of the plate. The product E of the intensity $I = A^2$ of light that reaches the plate and the time t of exposure,

$$E = It, \tag{4}$$

is called simply *exposure* (or light sum) and the curve giving D against $\log_{10} E$ is known as the *Hurter–Driffield curve*; a typical form of this curve is shown in Fig. 8.51. In the range between the points P and Q the curve is practically a straight line, and if Γ is its slope, the density of the negative is evidently given by

$$D = D_0 + \Gamma \log_{10}\frac{E}{E_0}, \tag{5}$$

where D_0 and E_0 are constants. Using (3) it follows that

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

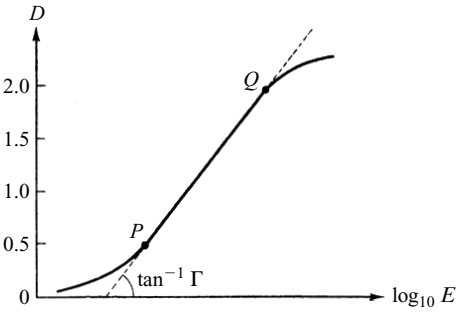


Fig. 8.51 The Hurter–Driffield curve (photographic response).

$$\tau = \tau_0 \left(\frac{E}{E_0} \right)^{-\Gamma}. \tag{6}$$

For pure absorption, without phase change, α is a real number, the square root $\sqrt{\tau}$ of the intensity transmission. Thus, in this case, the amplitude transmission factor α_n of the 'negative hologram' is given by a relation of the form

$$\alpha_n = (K_n A)^{-\Gamma_n}, \tag{7}$$

where K_n is proportional to the square root of the time of the exposure.

Suppose now that we take a positive print of the negative hologram. The amplitude transmission factor α_p of the positive is

$$\alpha_p = [K_p (K_n A)^{-\Gamma_n}]^{-\Gamma_p} = K A^{\Gamma}, \tag{8}$$

where $\Gamma = \Gamma_n \Gamma_p$ is the 'overall gamma' of the negative–positive process and $K = K_p^{-\Gamma_p} K_n^{\Gamma}$.

8.10.2 The reconstruction

In the reconstruction process (Fig. 8.50(b)) the positive hologram ($\mathcal{H}^+$), whose amplitude transmission factor is given by (8), is illuminated by the coherent background $U^{(i)}$ alone. The background is obtained simply by removing the object, otherwise preserving the geometry of the original arrangement. A substitute wave U' is transmitted by the plate, and this, according to (2) and (8), is represented by

$$U' = \alpha_p U^{(i)} = K A^{(i)} \mathrm{e}^{\mathrm{i}\psi_i} [A^{(i)2} + A^{(s)2} + 2A^{(i)} A^{(s)} \cos(\psi_s - \psi_i)]^{\frac{1}{2}\Gamma}. \tag{9}$$

If we choose $\Gamma = 2$, then

$$U' = K A^{(i)2} \mathrm{e}^{\mathrm{i}\psi_i} \left[A^{(i)} + \frac{A^{(s)2}}{A^{(i)}} + A^{(s)} \mathrm{e}^{\mathrm{i}(\psi_s - \psi_i)} + A^{(s)} \mathrm{e}^{-\mathrm{i}(\psi_s - \psi_i)} \right]. \tag{10}$$

On comparing (10) and (1) it is seen that if $A^{(i)}$ is constant, i.e. if the background is uniform, the substitute wave U' contains a component, called the *reconstructed wave*, proportional to U (the first and third term in (10)). The remainder of (10) consists of two terms. One has the same phase as the background and an amplitude $A^{(s)2}/A^{(i)}$ times that of the background. This term can be made very small by making the

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

background sufficiently strong.* The other term has the same amplitude ($KA^{(i)2}A^{(s)}$) as the reconstructed wave, but has a phase shift of opposite sign relative to the background. We say that it represents a *conjugate wave* and we shall show that it may be regarded as being due to a fictitious object of a similar nature as the true object, but situated in a different plane.

To show this we return for the moment to the arrangement of Fig. 8.50(a) and denote by O_1 any point in the object σ, and by P any point in the $\mathcal{H}$-plane (Fig. 8.52). When the object is illuminated from the point source S, the point P may be assumed to receive light along two rays, namely along the direct ray SP (associated with the coherent background) and along the 'diffracted secondary ray' O_1P (corresponding to the secondary wave). Suppose first that there is no change of phase on diffraction at O_1. Then the diffracted light at P is delayed with respect to the direct light by the path difference $O_1P - O_1A$, where A is the point on the intersection with SO_1 of the sphere centered on S and passing through P. Now the conjugate wave is advanced relative to the direct wave by the same amount as the diffracted secondary wave is retarded with respect to it, so that the conjugate wave catches up with the direct wave at that point O_2 on the line SO_1, where

$$O_1P - O_1A = PO_2 - AO_2. \tag{11}$$

If r_1, R and r_2 are the distances of the points O_1, A and O_2 from S, and α is the angle between SP and the line SO_1AO_2 (see Fig. 8.52), (11) may be written as

$$\sqrt{(R\cos\alpha - r_1)^2 + R^2\sin^2\alpha} - (R - r_1) = \sqrt{(r_2 - R\cos\alpha)^2 + R^2\sin^2\alpha} - (r_2 - R). \tag{12}$$

Expanding both sides in powers of α and retaining leading terms only, it follows that

$$\frac{1}{r_1} + \frac{1}{r_2} = \frac{2}{R}. \tag{13}$$

If the hologram were curved to a sphere of radius R, (13) would hold with respect to any point on it, and it follows that the conjugate wave could then be regarded as being due to a fictitious object which is the image of the original object in a spherical mirror, of radius R and centre S (see §4.4 (16)). This result evidently also holds, as a good approximation, in the case of a plane hologram, provided that the off-axis angles of the rays are small enough. If, moreover, r_1/R is small compared to unity, then according

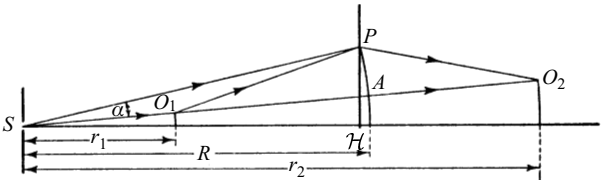


Fig. 8.52 Illustrating the position of the conjugate object.

* A strong coherent background is also utilized in a method due to F. Zernike (see *Proc. Phys. Soc.*, **61** (1948), 158) for the display of weak secondary interference fringes.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

to (13) $r_2 \sim -r_1$, so that the conjugate object and the true object are situated symmetrically about the source S.

It has been assumed so far that there is no phase change on diffraction by the object. If there is a change of phase, the preceding considerations regarding the position of the conjugate object still apply, provided that the conjugate object is assumed to produce a phase change of equal amount but of opposite sign to that produced by the true object.

Returning to (10), we see that, provided the background is uniform and strong in comparison with the scattered wave, the substitute wave U' is effectively the same as the original wave, apart from a contribution that may be regarded as arising from a conjugate object. Hence, *if a lens L is placed behind the positive hologram and the hologram is illuminated by the background wave alone* [Fig. 8.50(b)], *an image σ' of the original object will be formed in a plane conjugate to that of σ*, but this image will, in general, be perturbed by a contribution due to the conjugate object. Conditions may be found under which this perturbing effect is not serious. Roughly speaking, one may expect that this effect will be small if the separation between the image of the original object and its conjugate exceeds the focal tolerance of the image-forming pencil, this being given by §8.8 (27).

Fig. 8.50(c) illustrates the equivalent one-stage imaging and in Fig. 8.53 the enlarged object, the hologram, and the reconstructed image are shown relating to one of the first experiments of this type.

Later Leith and Upatnieks* considerably improved the wave-front reconstruction technique, by modifying the manner in which the coherent background is superposed onto the beam that is transmitted through the object. In their arrangement, which employs laser light, the background beam is incident on the plane of the hologram at an angle, with the help of a prism or a mirror system. The conjugate images are formed in different directions and consequently an image of high quality may be reconstructed without any disturbing effect of the other.†

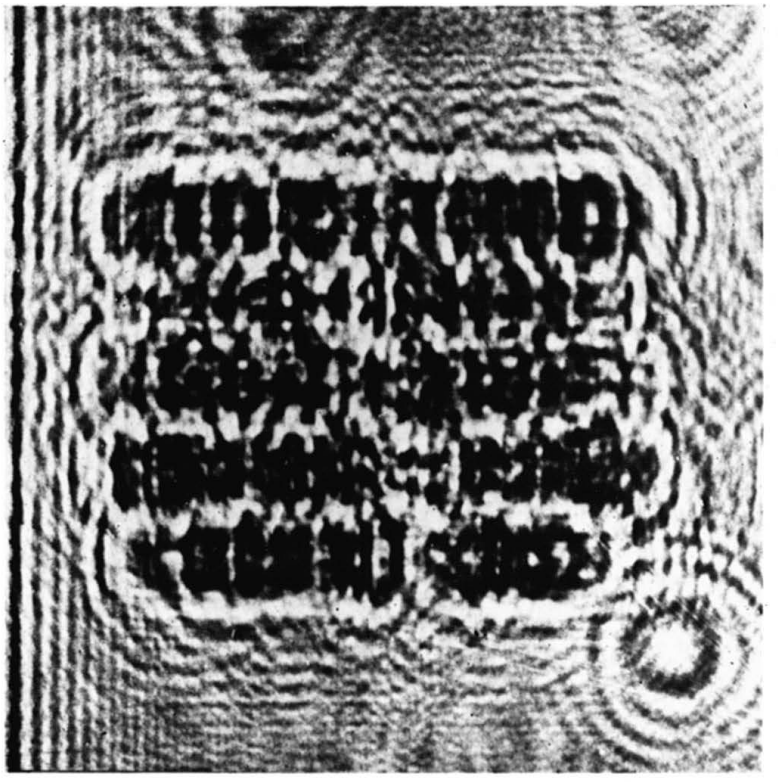
The hologram contains three-dimensional information. This fact was also further demonstrated by Leith and Upatnieks and is illustrated in Fig. 8.54.

The results confirm our conclusion that, with *coherent light*, it is possible to reconstruct an object to within a high degree of accuracy from a record of the *intensity distribution* alone, taken in any plane behind the object.

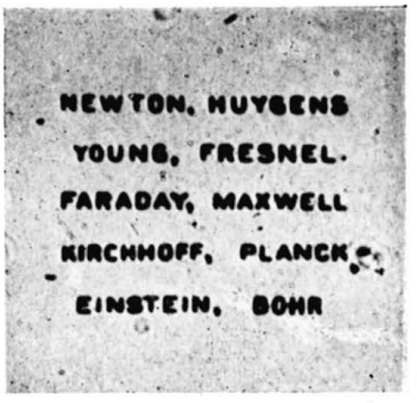
It is not essential for the success of the method that the wave-fronts from the source S should be strictly spherical. It is only necessary that the wave-fronts of the background wave in the arrangements illustrated in Figs. 8.50(a) and (b) should be of the same geometrical form. Nor is it necessary to employ the same source, or indeed radiation of the same wavelengths. It is in this connection that the potentialities of the method for electron microscopy are apparent. For one of the chief factors that limit the resolving power of the electron microscope is the spherical aberration of the objective, and as Gabor has pointed out, the effect of the spherical aberration can, in principle, be eliminated, or better expressed — compensated — by the reconstruction method. The hologram could be obtained with the electron beam, and the reconstruction with light. One would then have to *imitate* the spherical aberration of the electron objective

* E. N. Leith and J. Upatnieks, *J. Opt. Soc. Amer.*, **52** (1962), 1123; **53** (1963), 1377; **54** (1964), 1295.

† The diffraction theory of this modified holographic scheme was discussed by E. Wolf and J. R. Shewell, *J. Math. Phys.*, **11** (1970), 2254.



hologram

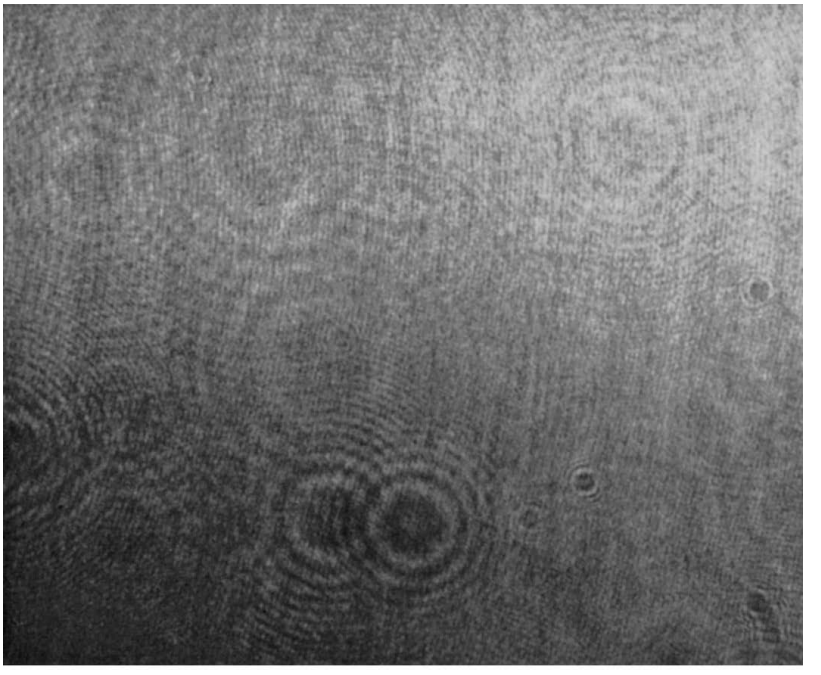


original

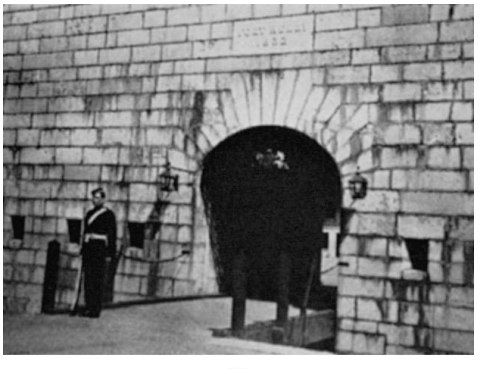
reconstruction

Fig. 8.53 Imaging by reconstructed wave-fronts. The object was a microphotograph of 1.5 mm diam., illuminated with light of wavelength $\lambda = 4358$ Å through a pinhole of diameter 0.2 mm, reduced by a microscope objective to 5 μm nominal diameter at 50 mm from the object. Geometrical magnification 12. Effective aperture of lens used in reconstruction was 0.025. Noisy background chiefly due to imperfections of illuminating objective. (After D. Gabor, *Proc. Roy. Soc.*, A, **197** (1949), 454.)

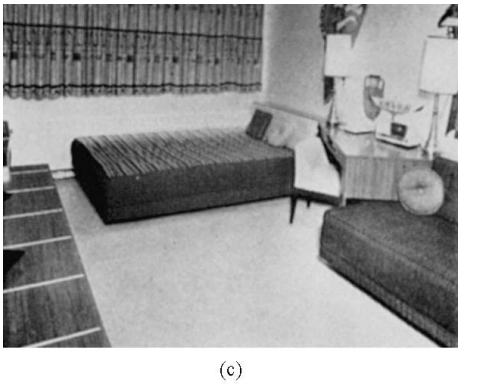
https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press



(a)



(b)



(c)

Fig. 8.54 Imaging by reconstructed wavefronts. Reconstruction of two objects from a single hologram. Hologram (a) and reconstructions (b) and (c) of two transparencies which were illuminated with diffused coherent light. The transparencies were placed 14 and 24 in from the hologram recording plate in such positions that neither obscured the other when viewed from the position where the hologram was recorded. The observable structures seen in the hologram are mainly diffraction patterns produced by dust particles. (After E. N. Leith and J. Upatnieks, *J. Opt. Soc. Amer.*, **54** (1964), 1297.)

by employing light waves that suffer from spherical aberration of the same amount (naturally scaled in the ratio $\lambda_{\text{light}}/\lambda_{\text{electron}}$).* Although, at the time of its invention, difficulties of a technical nature prevented the application of Gabor's method to electron microscopy, the correctness of the basic principles of his considerations was verified by experiments with light. In 1971 Dennis Gabor was awarded the Nobel Prize for Physics for his invention and development of the holographic method.

As a method of two-step photography, wave-front reconstruction has an important precursor. This is a method of optical Fourier analysis for the reconstruction of crystal structures from their diffraction patterns, first proposed by Boersch,† and independently by Bragg‡ who called the device the 'X-ray microscope'. If a crystal is illuminated by a parallel beam of X-rays, the angular distribution of intensity of the diffracted light, caught on a photographic plate, is, at least for small angles, the absolute square of the Fourier transform of the electron density distribution in the crystal, projected on a plane at right angles to the original beam.§ The fact that only the intensities are recorded, while the phases in the experiment remain physically undefined, prevents, in general, a direct reconstruction. The necessary data must be collected from other photographs, from previous knowledge of the chemical structure, and by guesswork. The work is comparatively simple when the direction of illumination is a crystal axis; in this case it can be shown that the diffracted wavelets are all in phase or in antiphase if the illumination is strictly coherent, so that each observed diffraction maximum then corresponds to two possible phases only; this, at any rate, greatly reduces the number of combinations which must be tried.

There are, however, certain crystals in which each cell contains a heavy atom, and the amplitudes due to the diffraction by the lattice of heavy atoms are so overwhelmingly large that the resulting amplitudes can have only one sign. One can say that in this case the heavy atoms produce a coherent background. Thus, in the case of these (rather exceptional) crystals, one only needs to take the square root of the intensities of the diffracted light in order to obtain the Fourier transform of the density distribution projected at right angles to the crystal axis, parallel to the direction of illumination.

Now an amplitude distribution proportional to the square root of the intensity can be obtained from a print of the diffraction pattern taken with an overall gamma of unity [see (8)]. In the X-ray microscope this print is illuminated with a plane wave-front and the light transmitted and diffracted by the print is focused by a lens. According to the theory of Fraunhofer diffraction (§8.3) the amplitude pattern in the image plane is the Fourier transform of the distribution in the plane of the plate and hence results from the original density distribution by two successive Fourier transformations; from Fourier's theorem it follows that the distribution in the focal plane is a faithful image of this (in general complex) density distribution. It is of course assumed that the aperture of the lens is large enough to admit all the diffracted rays that carry

* Light waves with a prescribed amount of spherical aberration can be produced by means of appropriate aspherical surfaces.

† H. Boersch, *Z. techn. Phys.*, **19** (1938), 337.

‡ W. L. Bragg, *Nature*, **149** (1942), 470; see also, M. J. Buerger, *J. Appl. Phys.*, **21** (1950), 909.

§ See, for example, A. Guinier and D. L. Dexter, *X-ray Studies of Materials* (New York, Interscience Publishers, 1963), Chapter 4.

appreciable amounts of energy and that the requirement of all wavelets being in phase is satisfied. The latter condition is, in fact, only rarely fulfilled.

In the method of wave-front reconstruction, a background is artificially added. There are no conditions to be satisfied regarding symmetry or even periodicity. On the other hand this method is not suitable for X-ray analysis owing to the practical impossibility of producing a strong coherent background. However, the method may well have applications in future electron diffraction studies.

8.11 The Rayleigh–Sommerfeld diffraction integrals

Towards the end of §8.3.2 we noted that even though the Kirchhoff diffraction theory is entirely adequate for the treatment of the majority of problems encountered in instrumental optics, it contains some rather unsatisfactory features. They arise from the nature of the assumed boundary conditions. For this reason other diffraction formulae are sometimes used. Among them the so-called Rayleigh–Sommerfeld diffraction integrals are particularly popular. In this section we will derive them and we will briefly comment on the relative merits of the different formulations. We begin with deriving a formal solution to two important boundary value problems from which the Rayleigh–Sommerfeld diffraction integrals readily follow.

8.11.1 The Rayleigh diffraction integrals

We recall §8.3 (5), viz.,

$$\iint_S \left(U \frac{\partial U'}{\partial n} - U' \frac{\partial U}{\partial n} \right) \mathrm{d}S = 0, \tag{1}$$

where $U(x,y,z)\exp(-\mathrm{i}\omega t)$ and $U'(x,y,z)\exp(-\mathrm{i}\omega t)$ are two monochromatic wave functions defined throughout a domain $\mathcal{V}$ bounded by a closed surface S and $\partial/\partial n$ denotes differentiation along the inward normal to S. We again make the choice

$$U'(x,\ y,\ z) = \frac{\mathrm{e}^{\mathrm{i}ks}}{s}, \tag{2}$$

where s denotes the distance from an arbitrary point of observation $P(x,y,z)$ to a point of integration $P'(x',\ y',\ z')$ on S. Then, as shown in §8.3.1 if the point P is inside the volume $\mathcal{V}$, (1) implies that [§8.3 (7)]

$$\iint_S \left[U \frac{\partial}{\partial n} \left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s} \right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s} \frac{\partial U}{\partial n} \right] \mathrm{d}S = 4\pi U(P). \tag{3}$$

If, however, the point P is located outside the volume $\mathcal{V}$, (1) implies that

$$\iint_S \left[U \frac{\partial}{\partial n} \left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s} \right) - \frac{\mathrm{e}^{\mathrm{i}ks}}{s} \frac{\partial U}{\partial n} \right] \mathrm{d}S = 0. \tag{4}$$

Suppose that the domain $\mathcal{V}$ is the half-space $z \geqslant 0$. The surface S of integration then consists of the plane $z' = 0$ and a hemisphere in that half-space, of infinitely large radius, centred at the origin. We assume that U behaves as an outgoing spherical wave at infinity in $\mathcal{V}$, i.e. that at points sufficiently far away from the origin

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$U(x,\,y,\,z) \sim \frac{\mathrm{e}^{\mathrm{i}kr}}{r}, \tag{5}$$

with $r = \sqrt{x^2 + y^2 + z^2}$ and $z \geqslant 0$. Then, for the same reason as explained on p. 422 in deriving the Fresnel–Kirchhoff diffraction integral there are no contributions on the integral in (4) from the large sphere ($r \to \infty$) in the half-space $z > 0$, centred at the origin. If the point $P(x,\,y,\,z)$ is situated in the half-space $z > 0$, (3) gives

$$\frac{1}{4\pi}\iint_{(z'=0)}\left[U\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}kR^+}}{R^+}\right) - \frac{\mathrm{e}^{\mathrm{i}kR^+}}{R^+}\frac{\partial U}{\partial z'}\right]\mathrm{d}x'\,\mathrm{d}y' = U(x,\,y,\,z), \tag{6}$$

where (see Fig. 8.55)

$$R^+ = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}, \qquad (z > 0). \tag{7}$$

The point $P(x,\,y,\,-z)$ will then be situated in the half-space $z < 0$ and (4) gives

$$\frac{1}{4\pi}\iint_{(z'=0)}\left[U\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}kR^-}}{R^-}\right) - \frac{\mathrm{e}^{\mathrm{i}kR^-}}{R^-}\frac{\partial U}{\partial z'}\right]\mathrm{d}x'\,\mathrm{d}y' = 0, \tag{8}$$

where

$$R^- = \sqrt{(x-x')^2 + (y-y')^2 + (z+z')^2}. \tag{9}$$

It is to be noted that

$$\left.\frac{\mathrm{e}^{\mathrm{i}kR^+}}{R^+}\right|_{z'=0} = \left.\frac{\mathrm{e}^{\mathrm{i}kR^-}}{R^-}\right|_{z'=0}, \tag{10a}$$

whereas

$$\left.\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}kR^-}}{R^-}\right)\right|_{z'=0} = -\left.\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}kR^+}}{R^+}\right)\right|_{z'=0}. \tag{10b}$$

If we make use of (10a) and (10b) in (8) we obtain the formula

$$\frac{1}{4\pi}\iint_{(z'=0)}\left[-U\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}kR^+}}{R^+}\right) - \frac{\mathrm{e}^{\mathrm{i}kR^+}}{R^+}\frac{\partial U}{\partial z'}\right]\mathrm{d}S = 0. \tag{11}$$

If we subtract (11) from (6) and set $R^+|_{z'=0} = s$ we find at once that

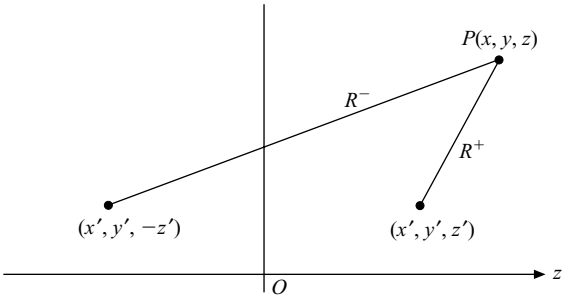


Fig. 8.55 Illustrating the geometrical significance of the distances R^+ and R^- defined by (7) and (9).

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$U(x,\ y,\ z) = \frac{1}{2\pi}\iint_{(z'=0)} U(x',\ y',\ 0)\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right)\mathrm{d}x'\,\mathrm{d}y'. \tag{12}$$

If we add (11) and (6) and again set $R^+|_{z'=0} = s$ we find that

$$U(x,\ y,\ z) = -\frac{1}{2\pi}\iint_{(z'=0)}\left[\frac{\partial U(x',\ y',\ z')}{\partial z'}\right]\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\,\mathrm{d}x'\,\mathrm{d}y'. \tag{13}$$

Formulas (12) and (13) were first derived by Lord Rayleigh* and are known as *the Rayleigh diffraction formulae* of the first and the second kind respectively. Evidently the Rayleigh diffraction formula of the first kind [(12)] provides a solution of the Helmholtz equation valid throughout the half-space $z > 0$ which takes on prescribed values on the plane $z = 0$ and is outgoing at infinity in that half-space. The Rayleigh diffraction formula of the second kind [(13)] provides a solution of the Helmholtz equation valid throughout the half-space $z > 0$, whose normal derivative $\partial U(x,\ y,\ z)/\partial z$ takes on prescribed values on the plane $z = 0$ and is outgoing at infinity in that half-space. Using the terminology of the theory of differential equations the Rayleigh diffraction formula of the first kind provides the solution to a Dirichlet boundary value problem,† whereas the Rayleigh diffraction formula of the second kind provides the solution to a Neumann boundary value problem.

8.11.2 The Rayleigh–Sommerfeld diffraction integrals

We now consider the same problem as we did in connection with Kirchhoff's diffraction theory in §8.3.2, namely diffraction of a monochromatic wave at an aperture $\mathcal{A}$ in an opaque screen. More specifically we wish to determine the wave disturbance at a point $P(x,\ y,\ z)$ in the half-space $z > 0$ into which the wave propagates. We again assume that the linear dimensions of the aperture are large compared with the wavelength but small compared with the distance P from the screen. As noted before, the values of U and of its normal derivative $\partial U/\partial z$ in the aperture and on the dark side $\mathcal{B}$ of the screen (see Fig. 8.3) are never known exactly. However, it seems reasonable to assume that under our assumptions $U \approx U^{(i)}$ on $\mathcal{A}$ and $U \approx 0$ on $\mathcal{B}$, with $U^{(i)}$ representing the incident field. With these assumptions the Rayleigh diffraction formula of the first kind [(12)] gives the following expression for the diffracted field in the half-space $z > 0$:

$$U(x,\ y,\ z) = \frac{1}{2\pi}\iint_{\mathcal{A}} U^{(i)}(x',\ y',\ 0)\frac{\partial}{\partial z'}\left(\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\right)\mathrm{d}x'\,\mathrm{d}y'. \tag{14}$$

It seems also reasonable to expect that under the stated assumptions $\partial U/\partial z \approx \partial U^{(i)}/\partial z$ in the aperture $\mathcal{A}$ and $\partial U/\partial z \approx 0$ on the dark side $\mathcal{B}$ of the screen. The Rayleigh diffraction integral of the second kind [(13)] then gives the following expression for the diffracted field in the half-space $z > 0$:

* Lord Rayleigh, *Phil. Mag.*, **43** (1897), 259; also his *Scientific Papers*, Vol. 4 (Cambridge, Cambridge University Press, 1903), p. 283.

† Precise conditions which will ensure a unique solution of the Dirichlet problem were discussed by R. K. Luneburg, *Mathematical Theory of Optics* (Berkeley and Los Angeles, CA, University of California Press, 1964), §45.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

$$U(x,\ y,\ z) = -\frac{1}{2\pi}\iint_{\mathcal{A}} \frac{\partial U^{(i)}(x',\ y',\ z')}{\partial z'} \frac{\mathrm{e}^{\mathrm{i}ks}}{s}\,\mathrm{d}x'\,\mathrm{d}y'. \tag{15}$$

Formulae (14) and (15) are known as the *Rayleigh–Sommerfeld diffraction integrals of the first and the second kind respectively.* They seem to have been first presented by Sommerfeld for the case when the incident field is a divergent spherical wave.* In that case

$$U^{(i)} = \frac{A\mathrm{e}^{\mathrm{i}kr}}{r}, \qquad \frac{\partial U^{(i)}}{\partial z} = \frac{A\mathrm{e}^{\mathrm{i}kr}}{r}\left(\mathrm{i}k - \frac{1}{r}\right)\cos(n,\ r), \tag{16}$$

where $\cos(n,\ r)$ denotes the angle between the normal to the aperture plane (the positive z direction and the direction of diffraction [see Fig. 8.3(b))]. The Rayleigh–Sommerfeld diffraction formula of the first kind [(14), with superscript I] then gives

$$U^{(\mathrm{I})}_{RS}(x,\ y,\ z) = \frac{1}{2\pi}\iint_{\mathcal{A}} \frac{A\mathrm{e}^{\mathrm{i}kr}}{r}\left(\mathrm{i}k + \frac{1}{s}\right)\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\cos(n,\ s)\mathrm{d}S. \tag{17}$$

Similarly, the Rayleigh–Sommerfeld diffraction formula of the second kind [(15), with superscript II] gives,

$$U^{(\mathrm{II})}_{RS}(x,\ y,\ z) = -\frac{1}{2\pi}\iint_{\mathcal{A}} \frac{A\mathrm{e}^{\mathrm{i}kr}}{r}\left(\mathrm{i}k - \frac{1}{r}\right)\frac{\mathrm{e}^{\mathrm{i}ks}}{s}\cos(n,\ r)\mathrm{d}S. \tag{18}$$

Under usual circumstances $r \gg \lambda,\ s \gg \lambda$ and consequently $k \gg 1/s,\ k \gg 1/r$ and (17) and (18) acquire the simpler forms

$$U^{(\mathrm{I})}_{RS}(x,\ y,\ z) \approx \frac{\mathrm{i}A}{\lambda}\iint_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{rs}\cos(n,\ s)\mathrm{d}S, \tag{19}$$

and

$$U^{(\mathrm{II})}_{RS}(x,\ y,\ z) \approx -\frac{\mathrm{i}A}{\lambda}\iint_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{rs}\cos(n,\ r)\mathrm{d}S. \tag{20}$$

We note that both these expressions have the same form as the Fresnel–Kirchhoff diffraction formula §8.3 (17), viz.,

$$U_{FK}(x,\ y,\ z) = -\frac{\mathrm{i}A}{2\lambda}\iint_{\mathcal{A}} \frac{\mathrm{e}^{\mathrm{i}k(r+s)}}{rs}[\cos(n,\ r) - \cos(n,\ s)]\mathrm{d}S. \tag{21}$$

For small angles of incidence and for small angles of diffraction $\cos(n,\ r) \approx 1$ $\cos(n,\ s) \approx -1$ and we see that under these circumstances

$$U^{(\mathrm{I})}_{RS}(x,\ y,\ z) \approx U^{(\mathrm{II})}_{RS}(x,\ y,\ z) \approx U_{FK}(x,\ y,\ z), \tag{22}$$

i.e., under these circumstances all the three formulae predict essentially the same values for the diffracted field.

Irrespective of whether or not the incident field is a diverging spherical wave, the following relation holds:

$$U^{(K)} = \tfrac{1}{2}(U^{(\mathrm{I})} + U^{(\mathrm{II})}). \tag{23}$$

* A. Sommerfeld, *Optics* (New York, Academic Press, 1954), p. 199 *et seq*.

https://doi.org/10.1017/9781108769914.011 Published online by Cambridge University Press

Here $U^{(K)}$ is the field predicted by Kirchhoff's theory (§8.3 (14), with the boundary conditions §8.3 (15) and with contribution from the large hemisphere $\mathcal{C}$ neglected) and $U^{(\mathrm{I})}$ and $U^{(\mathrm{II})}$ are the fields given by the Rayleigh–Sommerfeld integrals of the first kind [(14)] and the second kind [(15)] respectively.

Eq. (22) shows that far away from the aperture and when the angles of incidence and diffraction are small, the three diffraction formulae give essentially the same results. No simple comparison can be made under other circumstances. Nevertheless there is a widely held belief that the Rayleigh–Sommerfeld integrals are superior to Kirchhoff's. This belief is undoubtedly due to the fact that the Rayleigh–Sommerfeld integrals are 'manifestly consistent' in the sense that they reproduce the assumed boundary values as the field point approaches the plane of the aperture.* However, the assumed values do not take into account diffraction at the edge of the aperture and, consequently, do not correctly represent the true physical field in the aperture plane. Moreover, with a black screen the exact boundary values are not known. The same criticism applies, of course, to the Kirchhoff diffraction integral which is not 'manifestly consistent', because it does not recover the assumed boundary values as the field point approaches the aperture.† Although neither of these theories predicts accurately the field in the aperture and in its immediate vicinity there is some evidence that the Kirchhoff theory gives results that are in closer agreement with observation.‡

Finally we note that unlike Kirchhoff's diffraction theory, the Rayleigh–Sommerfeld formulation is restricted to diffraction at apertures in plane screens, which limits its practical usefulness.

* N. Mukunda, *J. Opt. Soc. Amer.*, **52** (1962), 336.

† As mentioned on p. 424 Kirchhoff's diffraction integral may be interpreted in a consistent way which differs, however, from Kirchhoff's original formulation.

‡ See E. W. Marchand and E. Wolf, *J. Opt. Soc. Amer.* **56** (1966), 1712, Sec. 4. The Kirchhoff theory was tested with microwaves diffracted at a black (i.e. essentially completely absorbing) half-plane. It was found to be in good agreement with experiment on the side of the screen away from the source. (J. F. Nye, J. H. Hannay and W. Liang, *Proc. Roy. Soc. Lond.* A. **449** (1995), 515). It should be noted that the Sommerfeld theory discussed in that reference is not the Rayleigh–Sommerfeld formulation outlined in this section.