IX

The diffraction theory of aberrations

In Chapter V we studied the effects of aberrations on the basis of geometrical optics. In that treatment the image was identified with the blurred figure formed by the points of intersection of the geometrical rays with the image plane. Since geometrical optics gives an approximate model valid in the limit of very short wavelengths, it is to be expected that the geometrical theory gradually loses its validity as the aberrations become small. For example, in the limiting case of a perfectly spherical convergent wave issuing from a circular aperture, geometrical optics predicts for the focal plane an infinite intensity at the focus and zero intensity elsewhere, whereas, as has been shown in §8.5.2, the real image consists of a bright central area surrounded by dark and bright rings (the Airy pattern). In the neighbourhood of the focal plane the light distribution has also been seen to be of a much more complicated nature (see Fig. 8.41) than geometrical optics suggests. We are thus led to the study of the effects of aberrations on the basis of diffraction theory.

The first investigations in this field are due to Rayleigh.* His main contribution was the formulation of a criterion (discussed in §9.3) which, in an extended form, has come to be widely used for determining the maximum amounts of aberrations that may be tolerated in optical instruments. The subject was carried further by the researches of many writers who investigated the effects of various aberrations,† and we may mention, in particular, the more extensive treatments by Steward, Picht, and Born.‡

A very extensive diffraction treatment of image formation in the presence of aberrations is due to Nijboer, § carried out partly in collaboration with Zernike. It is

* Lord Rayleigh, *Phil. Mag.*, (5), **8** (1879), 403. Reprinted in his *Scientific Papers*, Vol. 1 (Cambridge, Cambridge University Press, 1899), p. 428.

† A historical survey of diffraction theory of aberrations was given by E. Wolf in *Rep. Progr. Phys.* (London, Physical Society), **14** (1951), 95.

‡ G. C. Steward, *Phil. Trans. Roy. Soc.*, A, **255** (1925), 131; also his book *The Symmetrical Optical System* (Cambridge, Cambridge University Press, 1928). J. Picht, *Ann. d. Physik*, (4), **77** (1925), 685. *ibid.*, **80** (1926), 491; also his *Optische Abbildung* (Braunschweig, Vieweg, 1931). M. Born, *Naturwissenschaften*, **20** (1932), 921; and his *Optik* (Berlin, Springer, 1933, reprinted 1965), p. 202.

§ B. R. A. Nijboer, Thesis, University of Groningen, 1942. The main part was also published in *Physica*, **10** (1943), 679; *ibid.*, **13** (1947), 605; F. Zernike and B. R. A. Nijboer, contribution to *La Théorie des Images Optiques* (Paris, Revue d'Optique, 1949), p. 227. An extension of the theory to somewhat larger aberrations was discussed by K. Nienhuis and B. R. A. Nijboer, *Physica*, **14** (1948), 590. Mention must also be made of a thesis by K. Nienhuis (University of Groningen, 1948), which is mainly concerned with the experimental study of the effects of aberrations. Some of the beautiful photographs obtained by Nienhuis are reproduced in §9.4.

Experimental investigations at microwave frequencies, concerning the structure of the image region in the presence of aberrations were described by M. P. Bachynski and G. Bekefi in the paper referred to on p. 489, and also in *Trans. Inst. Radio Eng.*, AP-4 (1956), 412.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

concerned with influence of small aberrations, when the departures of the wave-fronts from spherical form are only a fraction of the wavelength. The effects of large aberrations were studied on the basis of diffraction theory by van Kampen* with the help of asymptotic approximations; this treatment is based on a formal extension to functions of two variables of the principle of stationary phase, which has since been formulated rigorously, first by J. Focke (see Appendix III, p. 889).

In the main part of this chapter we shall give an account of the Nijboer–Zernike theory and examine the structure of diffraction images affected by primary aberrations. In the final section (§9.5) we shall go over from point objects to extended objects and investigate imaging with coherent and incoherent illumination. Imaging with partially coherent illumination will be considered in Chapter X.

9.1 The diffraction integral in the presence of aberrations

9.1.1 The diffraction integral

Consider a centred optical system with a point source of monochromatic light P_0 (Fig. 9.1). We take a Cartesian system of axes, with origin at the Gaussian image point $P_1^\star$ of P_0, with the z-axis along $CP_1^\star$, where C in the centre of the exit pupil. The y-axis is taken in the meridional plane (the plane containing P_0 and the axis of the system). The off-axis distances of P_0 and $P_1^\star$ will be denoted by Y_0 and $Y_1^\star$ respectively.

As in Chapter V the deformation of the wave-fronts in the region of the exit pupil will be described by the aberration function Φ. Let $\overline{Q}$ and Q be the points in which a ray in the image space intersects the wave-front through C and the Gaussian reference sphere respectively. Assuming that the refractive index of the image space is unity, Φ (taken as positive in Fig. 9.1) represents the distance $\overline{Q}Q$ measured along the ray.

Let R denote the radius $CP_1^\star$, of the Gaussian reference sphere and let s be the distance between Q and an arbitrary point P in the region of the image. The disturbance at Q is represented by $A\mathrm{e}^{\mathrm{i}k(\Phi-R)}/R$, where A/R is the amplitude at Q. According to the Huygens–Fresnel principle the disturbance at P is given by

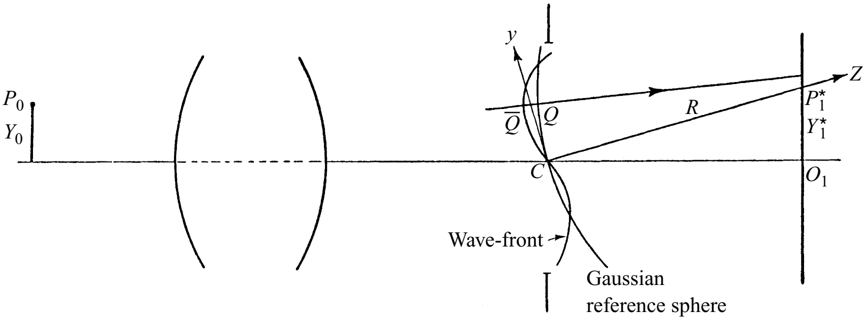


Fig. 9.1 Choice of reference system and notation.

[*] N. G. van Kampen, *Physica*, **14** (1949), 575; *ibid.*, **16** (1950), 817; *ibid.* **25** (1958), 437.

$$U(P) = -\frac{\mathrm{i}}{\lambda}\frac{A\mathrm{e}^{-\mathrm{i}kR}}{R}\iint \frac{\mathrm{e}^{\mathrm{i}k[\Phi+s]}}{s}\,\mathrm{d}S, \tag{1}$$

where the integration extends over the portion of the reference sphere that approximately fills the exit pupil. In (1) we have assumed that the angles involved are small, so that the variation of the inclination factor over the reference sphere may be neglected; we have also assumed that the amplitude of the wave is substantially constant over the wave-front, so that A can be taken outside the integral.

Let (ξ, η, ζ) be the coordinates of Q and (x, y, z) the coordinates of P, and let a be the radius of the exit pupil. As in §8.8, which was concerned with the special case of an aberration-free wave ($\Phi \equiv 0$), we set

$$\left.\begin{aligned} \xi &= a\rho\sin\theta, \qquad & x &= r\sin\psi, \\ \eta &= a\rho\cos\theta, \qquad & y &= r\cos\psi, \end{aligned}\right\} \tag{2}$$

and we have, as in §8.8 (2) and §8.8 (9)*

$$k(s-R) = -v\rho\cos(\theta-\psi) - \tfrac{1}{2}u\rho^2 + \left(\frac{R}{a}\right)^2 u, \tag{3}$$

where u and v are the two 'optical coordinates' of P,

$$u = \frac{2\pi}{\lambda}\left(\frac{a}{R}\right)^2 z, \qquad v = \frac{2\pi}{\lambda}\left(\frac{a}{R}\right)\sqrt{x^2+y^2}. \tag{4}$$

It will now be convenient to regard Φ as a function of $Y_1^\star$, ρ and θ,

$$\Phi = \Phi(Y_1^\star, \rho, \theta). \tag{5}$$

The element of the Gaussian reference sphere is $\mathrm{d}S = a^2\rho\,\mathrm{d}\rho\,\mathrm{d}\theta$, and if the angle which $CP_1^\star$ makes with the axis of the system is small, the range of integration may be taken as $0 \leqslant \rho \leqslant 1$, $0 \leqslant \theta < 2\pi$. Moreover, for points of observations in the region of the image, s may be replaced by R in the denominator of the integrand. Hence (1) becomes, on substitution from (3),

$$U(P) = U(u, v, \psi) = -\frac{\mathrm{i}}{\lambda}\frac{Aa^2}{R^2}\mathrm{e}^{\mathrm{i}(\frac{R}{a})^2 u}\int_0^1\int_0^{2\pi}\mathrm{e}^{\mathrm{i}[k\Phi(Y_1^\star,\rho,\theta)-v\rho\cos(\theta-\psi)-\frac{1}{2}u\rho^2]}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta, \tag{6}$$

so that the intensity at P is

$$I(P) = |U(P)|^2 = \left(\frac{Aa^2}{\lambda R^2}\right)^2\left|\int_0^1\int_0^{2\pi}\mathrm{e}^{\mathrm{i}[k\Phi(Y_1^\star,\rho,\theta)-v\rho\cos(\theta-\psi)-\frac{1}{2}u\rho^2]}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2. \tag{7}$$

It is convenient to express the intensity $I(P)$ as fraction of the intensity $I^\star$ which would be obtained at the Gaussian image point $P_1^\star$ if no aberrations were present. According to (7),

$$I^\star = \pi^2\left(\frac{Aa^2}{\lambda R^2}\right)^2, \tag{8}$$

* *R* now corresponds to *f* of §8.8. It is of interest to note that if §8.8 (2) and §8.8 (3) are used, the diffraction integral may again be expressed in the form of an angular spectrum of plane waves (see J. Focke, *Optica Acta*, **3** (1956), 110).

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

so that the *normalized intensity* is[*]

$$i(P) = \frac{I(P)}{I^\star} = \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi} \mathrm{e}^{\mathrm{i}[k\Phi(Y_1^\star,\rho,\theta)-v\rho\cos(\theta-\psi)-\frac{1}{2}u\rho^2]}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2. \tag{9}$$

In the absence of aberrations the intensity is a maximum at the Gaussian image point. When aberrations are present, this will in general no longer be the case, and we may call the point of maximum intensity the *diffraction focus*.[†] Often one is interested only in the maximum intensity in a particular plane of observation; this value, when normalized as in (9), is called the *Strehl intensity*.[‡]

From (9) some simple results, which will be needed later, may immediately be deduced.

9.1.2 The displacement theorem. Change of reference sphere

Let Φ and Φ' be two aberration functions such that

$$\Phi' = \Phi + H\rho^2 + K\rho\sin\theta + L\rho\cos\theta + M, \tag{10}$$

where H, K, L and M are constants of order λ. Further let $i(u, v, \psi)$ and $i'(u, v, \psi)$ be the corresponding normalized intensities. Then by (9)

$$i(u, v, \psi) = \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi} \mathrm{e}^{\mathrm{i}f(u,v,\psi;\rho,\theta)}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2, \tag{11}$$

where

$$f(u, v, \psi; \rho, \theta) = k\Phi - v\rho\cos(\theta-\psi) - \tfrac{1}{2}u\rho^2, \tag{12}$$

with a similar expression for i'. Now according to (10) the last expression may also be written in the form

$$\begin{aligned} f(u, v, \psi; \rho, \theta) &= k\Phi' - k[H\rho^2 + K\rho\sin\theta + L\rho\cos\theta + M] - v\rho\cos(\theta-\psi) - \tfrac{1}{2}u\rho^2 \\ &= k\Phi' - v'\rho\cos(\theta-\psi') - \tfrac{1}{2}u'\rho^2 - kM \\ &= f'(u', v', \psi'; \rho, \theta) - kM, \end{aligned} \tag{13}$$

where

$$u' = u + 2kH, \qquad v'\sin\psi' = v\sin\psi + kK, \qquad v'\cos\psi' = v\cos\psi + kL. \tag{14}$$

According to (2) and (4), (14) represent the transformation

[*] No confusion should arise between the symbol i used for the normalized intensity, and the symbol i used for $\sqrt{-1}$, since the former always occurs with arguments, e.g. $i(P)$, $i(u, v, \psi)$, etc.

[†] In general there may be, of course, more than one diffraction focus, but the diffraction focus is unique if the aberrations are sufficiently small.

[‡] This concept is due to K. Strehl, *Z. f. Instrumkde.*, **22** (1902), 213, who called it ‘Definitionshelligkeit’. In the English literature the less appropriate term ‘definition’ is often used. More recently the term ‘Strehl ratio’ is also employed. See, for instance, V. N. Mahajan, *J. Opt. Soc. Amer.*, **71** (1981), 75 and *ibid.*, **72** (1982), 1258.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$z' = z + 2\left(\frac{R}{a}\right)^2 H, \qquad x' = x + \left(\frac{R}{a}\right) K, \qquad y' = y + \left(\frac{R}{a}\right) L. \tag{15}$$

From (11) and (13) it follows that

$$i(u, v, \psi) = i'(u', v', \psi'). \tag{16}$$

We have thus established the following *displacement theorem*: *The addition to an aberration function of a term* $H\rho^2 + K\rho \sin\theta + L\rho\cos\theta + M$, *where* H, K, L *and* M *are constants of order* λ, *results in no change in the three-dimensional intensity distribution near focus apart from a displacement of the distribution as a whole in accordance with the transformation* (15); a shift of amount $2(R/a)^2 H$ occurs along the principal direction $CP_1^\star$ away from the exit pupil and shifts of amounts $(R/a)K$ and $(R/a)L$ occur in the positive x and y directions respectively.

The additive terms on the right-hand side of (10) may be interpreted as representing a change of reference sphere. Suppose that we choose a new reference sphere, centred on a point $P'(x', y', z')$ in the image region, and of radius R' such that the new sphere is at most a few wavelengths away from the Gaussian sphere. Let a ray $\overline{Q}Q$ intersect the new reference sphere at a point N. Then the wave aberration Φ' referred to this new sphere is (see Fig. 9.2)

$$\Phi' = \overline{Q}N = \overline{Q}Q - NQ \sim \overline{Q}Q - NG, \tag{17}$$

where G is the point in which the line NP' intersects the Gaussian reference sphere, the refractive index of the image space being assumed to be unity as before. Now $\overline{Q}Q = \Phi$ is the wave aberration referred to the Gaussian sphere and $NG = NP' - GP' = R' - s$, where s denotes the distance from G to P'. Hence (17) may be written as

$$\Phi' \sim \Phi + s - R' = \Phi + \frac{\lambda}{2\pi}\left[-v\rho\cos(\theta - \psi) - \tfrac{1}{2}u\rho^2 + \left(\frac{R}{a}\right)^2 u\right] + (R - R'), \tag{18}$$

where (3) was used. Here u, v and ψ are given by (3) and (4) with x', y', z' written in place of x, y, z. Relation (18) may be written in the form (10) with

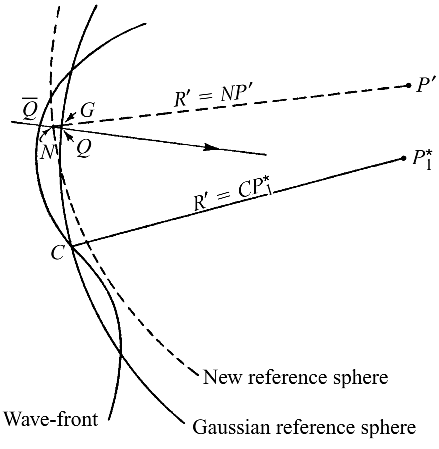


Fig. 9.2 Change of reference sphere.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$H = -\frac{1}{2}\left(\frac{a}{R}\right)^2 z', \quad K = -\left(\frac{a}{R}\right)x', \quad L = -\left(\frac{a}{R}\right)y', \quad M = z' + R - R'. \tag{19}$$

9.1.3 A relation between the intensity and the average deformation of wave-fronts

When the aberrations are small it is possible to express the intensity at the centre of the reference sphere in terms of the mean square value of the wave aberration. Let Φ_P be the wave aberration referred to a reference sphere centred on a point P in the image region. Then, according to (9) and (18), the normalized intensity at P may be expressed in the form

$$i(P) = \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi} \mathrm{e}^{\mathrm{i}k\Phi_P}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2 = \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi}[1 + \mathrm{i}k\Phi_P + \tfrac{1}{2}(\mathrm{i}k\Phi_P)^2 + \ldots]\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2. \tag{20}$$

Let $\overline{\Phi_P^n}$ denote the average value of the nth power of Φ_P, i.e.

$$\overline{\Phi_P^n} = \frac{\displaystyle\int_0^1\int_0^{2\pi}\Phi_P^n\rho\,\mathrm{d}\rho\,\mathrm{d}\theta}{\displaystyle\int_0^1\int_0^{2\pi}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta} = \frac{1}{\pi}\int_0^1\int_0^{2\pi}\Phi_P^n\rho\,\mathrm{d}\rho\,\mathrm{d}\theta. \tag{21}$$

If we assume that the aberrations are so small that we may neglect third and higher powers of $k\Phi_P$ in (20), the normalized intensity at P may be written in the form

$$i(P) \sim |1 + \mathrm{i}k\overline{\Phi_P} - \tfrac{1}{2}k^2\overline{\Phi_P^2}|^2 = 1 - \left(\frac{2\pi}{\lambda}\right)^2[\overline{\Phi_P^2} - (\overline{\Phi_P})^2]. \tag{22}$$

The quantity in the square brackets on the right is the ‘mean-square deformation’ $(\Delta\Phi)^2$ of the wave-front,

$$(\Delta\Phi_P)^2 = \frac{\displaystyle\int_0^1\int_0^{2\pi}(\Phi_P - \overline{\Phi_P})^2\rho\,\mathrm{d}\rho\,\mathrm{d}\theta}{\displaystyle\int_0^1\int_0^{2\pi}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta} = \overline{\Phi_P^2} - (\overline{\Phi_P})^2, \tag{23}$$

so that (22) may be written as

$$i(P) \sim 1 - \left(\frac{2\pi}{\lambda}\right)^2(\Delta\Phi_P)^2. \tag{24}$$

This formula implies that, when the aberrations are small, the normalized intensity at the centre of the Gaussian reference sphere in the region of focus is independent of the nature of the aberration and is smaller than the ideal value unity by an amount proportional to the mean-square deformation of the wave-front.*

* When the illumination of the pupil is nonuniform, (24) still applies, provided that $(\Delta\Phi_P)^2$ represents the amplitude-weighted mean square deformation of the wave-front (see, V. N. Mahajan, *loc. cit.*, (1981)).

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

9.2 Expansion of the aberration function

9.2.1 The circle polynomials of Zernike

In our discussion of the effects of aberrations on the basis of geometrical optics (Chapter V), we expanded the aberration function Φ in a power series. In the present treatment, where integrations over the unit circle must be carried out, it is more appropriate to expand Φ in terms of a complete set of polynomials that are orthogonal over the interior of the unit circle.* Many sets of polynomials with this property can be constructed; there is, however, one such set, introduced by Zernike,† which has certain simple properties of invariance. In Appendix VII it is shown how these *circle polynomials* of Zernike may be derived and some of their properties are discussed; here we shall only summarize the formulae needed in the present chapter.

The circle polynomials of Zernike are polynomials $V_n^l(X, Y)$ in two real variables X, Y, which, when expressed in polar coordinates ($X = \rho\sin\theta$, $Y = \rho\cos\theta$), are of the form

$$V_n^l(\rho\sin\theta, \rho\cos\theta) = R_n^l(\rho)e^{il\theta}, \tag{1}$$

where $l \gtreqless 0$, and $n \geqslant 0$ are integers, $n \geqslant |l|$, and $n - |l|$ is even. The orthogonality and normalizing properties are expressed by the formulae

$$\iint_{X^2+Y^2\leqslant 1} V_n^{l\star}(X, Y)V_{n'}^{l'}(X, Y)\mathrm{d}X\,\mathrm{d}Y = \frac{\pi}{n+1}\delta_{ll'}\delta_{nn'}, \tag{2}$$

where δ_{ij} is the Kronecker symbol and where the asterisks denote the complex conjugate. The radial functions $R_n^l(\rho)$ are polynomials in ρ, containing the powers ρ^n, $\rho^{n-2}, \ldots, \rho^{|l|}$ and, as is shown in Appendix VII, are closely related to Jacobi's polynomials (terminating hypergeometric series). As is seen from (1) and (2), the radial polynomials satisfy the relations

$$\int_0^1 R_n^l(\rho)R_{n'}^l(\rho)\rho\,\mathrm{d}\rho = \frac{1}{2(n+1)}\delta_{nn'}. \tag{3}$$

They are given by the formulae ($m = |l|$)

$$R_n^{\pm m}(\rho) = \frac{1}{\left(\dfrac{n-m}{2}\right)!\rho^m}\left[\frac{\mathrm{d}}{\mathrm{d}(\rho^2)}\right]^{\frac{n-m}{2}}\left[(\rho^2)^{\frac{n+m}{2}}(\rho^2-1)^{\frac{n-m}{2}}\right] \tag{4}$$

$$= \sum_{s=0}^{\frac{n-m}{2}}(-1)^s\frac{(n-s)!}{s!\left(\dfrac{n+m}{2}-s\right)!\left(\dfrac{n-m}{2}-s\right)!}\rho^{n-2s}. \tag{5}$$

The normalization has been chosen so that for all permissible values of n and m,

$$R_n^{\pm m}(1) = 1. \tag{6}$$

* The term 'complete' implies that any reasonably well-behaved function can be expanded as a series of functions of the set. For a more precise definition of the term see, for example, R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I (New York, Interscience Publishers, 1st English edition, 1953), pp. 51–54.

† F. Zernike, *Physica*, **1** (1934), 689.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

Table 9.1. *The radial polynomials* $R_n^m(\rho)$ *for* $m \leqslant 8$, $n \leqslant 8$.

$m\backslash^{n}$	0	1	2	3	4	5	6	7	8
0	1		$2\rho^2 - 1$		$6\rho^4 - 6\rho^2 + 1$		$20\rho^6 - 30\rho^4 + 12\rho^2 - 1$		$70\rho^8 - 140\rho^6 + 90\rho^4 - 20\rho^2 + 1$
1		ρ		$3\rho^3 - 2\rho$		$10\rho^5 - 12\rho^3 + 3\rho$		$35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho$	
2			ρ^2		$4\rho^4 - 3\rho^2$		$15\rho^6 - 20\rho^4 + 6\rho^2$		$56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho^2$
3				ρ^3		$5\rho^5 - 4\rho^3$		$21\rho^7 - 30\rho^5 + 10\rho^3$	
4					ρ^4		$6\rho^6 - 5\rho^4$		$28\rho^8 - 42\rho^6 + 15\rho^4$
5						ρ^5		$7\rho^7 - 6\rho^5$	
6							ρ^6		$8\rho^8 - 7\rho^6$
7								ρ^7	
8									ρ^8

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

The radial polynomials have the generating function

$$\frac{\left[1+z-\sqrt{1-2z(1-2\rho^2)+z^2}\right]^m}{(2z\rho)^m\sqrt{1-2z(1-2\rho^2)+z^2}} = \sum_{s=0}^{\infty} z^s R_{m+2s}^{\pm m}(\rho). \tag{7}$$

When $m = 0$, the left-hand side reduces to the generating function for the Legendre polynomials* of argument $2\rho^2 - 1$, so that

$$R_{2n}^0(\rho) = P_n(2\rho^2 - 1). \tag{8}$$

In Table 9.1 the explicit form of the polynomials for the first few values of the indices is given.

The following relation (also proved in the appendix) is of great importance in the Nijboer–Zernike theory:

$$\int_0^1 R_n^m(\rho) J_m(v\rho)\rho \, \mathrm{d}\rho = (-1)^{\frac{n-m}{2}} \frac{J_{n+1}(v)}{v} \tag{9}$$

where J is a Bessel function of the first kind.

Instead of the complex polynomials V, one may use the real polynomials†

$$\left.\begin{aligned} U_n^m &= \tfrac{1}{2}(V_n^m + V_n^{-m}) = R_n^m(\rho)\cos m\theta, \\ U_n^{-m} &= \tfrac{1}{2\mathrm{i}}(V_n^m - V_n^{-m}) = R_n^m(\rho)\sin m\theta. \end{aligned}\right\} \tag{10}$$

In our applications we shall need the polynomials $U_n^m = R_n^m(\rho)\cos m\theta$ only. This is so because the wave distortions are symmetrical about the meridional plane $\theta = 0$ and consequently the aberration function is an even function of θ.

9.2.2 Expansion of the aberration function

Following Nijboer, we expand the aberration function Φ in terms of Zernike's circle polynomials. As in §5.1 it follows from symmetry that the variables enter the expansion only in the combination $Y_1^{\star 2}$, ρ^2 and $Y_1^\star \rho\cos\theta$, so that the expansion must be of the form

$$\Phi(Y_1^\star, \rho, \theta) = \sum_l \sum_n \sum_m a_{lnm} Y_1^{\star\, 2l+m} R_n^m(\rho)\cos m\theta, \tag{11}$$

where l, n and m are nonnegative integers, $n \geqslant m$, $n - m$ is even and the a's are constants.

As we shall mainly be concerned with the diffraction image of a fixed object point ($Y_1^\star$ constant), it is convenient to suppress the explicit dependence of Φ on $Y_1^\star$ and re-write (11) in the form

$$\Phi = A_{00} + \frac{1}{\sqrt{2}}\sum_{n=2}^{\infty} A_{n0} R_n^0(\rho) + \sum_{n=1}^{\infty}\sum_{m=1}^{n} A_{nm} R_n^m(\rho)\cos m\theta. \tag{12}$$

[*] See, for example, R. Courant and D. Hilbert, *loc. cit.*, p. 85.

[†] A pictorial representation of some of these polynomials is given in an article by J. Y. Wang and D. E. Silva, *Appl. Opt.*, **19** (1980), 1510.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

The coefficients A_{nm} are functions of $Y_1^\star$, and the factor $1/\sqrt{2}$ has been introduced in the second term to simplify the final formulae.

If the aberrations are sufficiently small, the normalized intensity at the Gaussian focus may be expressed in a simple form in terms of the coefficients A. We have, on substituting from (21) into §9.1 (21) and using the orthogonality relations (3),

$$\left.\begin{aligned} \overline{\Phi} &= A_{00}, \\ \overline{\Phi^2} &= A_{00}^2 + \tfrac{1}{2}\sum_{n=1}^{\infty}\sum_{m=0}^{n}\frac{A_{nm}^2}{n+1}. \end{aligned}\right\} \tag{13}$$

The first relation implies that A_{00} represents the mean retardation of the wave-front behind the Gaussian reference sphere. The second relation is 'Parseval's formula' for the orthogonal set of functions $R_n^m(\rho)\cos m\theta$. On substituting from (13) into §9.1 (22), it follows that the normalized intensity at the Gaussian focus is

$$i(P_1^\star) = 1 - \frac{2\pi^2}{\lambda^2}\sum_{n=1}^{\infty}\sum_{m=0}^{n}\frac{A_{nm}^2}{n+1}. \tag{14}$$

A marked advantage of the expansion in terms of the circle polynomials arises in connection with the important problem of the 'balancing' of aberrations of different orders against each other in such a way as to obtain maximum intensity. Suppose that the aberration is represented by a single 'power series' term

$$\Phi = A'_{nm}\rho^n\cos^m\theta, \tag{15}$$

where A' is a small constant, of the order of a wavelength or less. We enquire whether it is possible to increase the intensity $i(P_1^\star)$ by introducing aberrations of lower order. More precisely, we wish to choose constants A'_{pq} in the expression

$$\Phi' = A'_{nm}\rho^n\cos^m\theta + \sum_{p<n}\sum_{q\leqslant p}A'_{pq}\rho^p\cos^q\theta, \tag{16}$$

so as to make the intensity at the Gaussian focus as large as possible.

With any particular choice of the constants A'_{pq}, the aberration function (16) may also be expressed in terms of the circle polynomials, in the form

$$\Phi' = \varepsilon_{nm}A_{nm}R_n^m(\rho)\cos m\theta + \sum_{p<n}\sum_{q\leqslant p}\varepsilon_{pq}A_{pq}R_p^q(\rho)\cos q\theta, \tag{17}$$

where

$$\left.\begin{aligned} \varepsilon_{nm} &= \frac{1}{\sqrt{2}} \quad \text{when} \qquad m=0,\ n\neq 0, \\ &= 1 \qquad \text{otherwise.} \end{aligned}\right\} \tag{18}$$

According to (5) the highest power of ρ in $R_n^m(\rho)$ is n and appears in the radial polynomial $R_n^m(\rho)$ with a coefficient $n!/[\frac{1}{2}(n+m)]![\frac{1}{2}(n-m)]!$. Hence, on comparing the coefficients of $\rho^n\cos^m\theta$ (or of $\rho^n\cos m\theta$) in (16) and (17), it follows that

$$A_{nm} = \varepsilon_{nm}\frac{[\frac{1}{2}(n+m)]![\frac{1}{2}(n-m)]!}{n!2^{m-1}}A'_{nm} \qquad \text{when } n\neq 0. \tag{19a}$$

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

It is also evident that

$$A_{00} = A'_{00}. \tag{19b}$$

Now with A'_{nm} and consequently A_{nm} fixed, it follows from (14) that the maximum intensity at $P_1^\star$ is obtained by making all the coefficients under the summation sign in (17) identically zero. The aberration function then becomes

$$\Phi' = \varepsilon_{nm} A_{nm} R_n^m(\rho) \cos m\theta, \tag{20}$$

and the normalized intensity at $P_1^\star$ now is

$$i(P_1^\star) = 1 - \frac{2\pi^2}{\lambda^2} \frac{A_{nm}^2}{n+1}, \tag{21}$$

A_{nm} being given, in terms of the coefficient A'_{nm}, by (19). It is now evident that, *in the single aberration terms* $A_{nm} R_n^m(\rho) \cos m\theta$ *of the expansion* (12)*, a number of terms of the form* $A'_{pq}\rho^p \cos^q\theta$ *have been combined, with* $p = n, n-2, \ldots, m$; $q = m, m-2, \ldots, 1$ *or* 0 *in such a way that, for a given (sufficiently small) value of the coefficient of* $\rho^n \cos^m \theta$, *the normalized intensity at the Gaussian focus is a maximum.*

We illustrate this result by a simple example. Suppose that a system suffers from a small amount of sixth-order spherical aberration ($\Phi = A'_{60}\rho^6$), and that we are in a position to introduce a controlled amount of fourth-order spherical aberration ($A'_{40}\rho^4$) and defocusing ($A'_{20}\rho^2$). We seek the values of the coefficients A'_{40} and A'_{20} which make the intensity at the diffraction focus as large as possible. Problems of this type were first studied by Richter,* who showed that the maximum is obtained when the two coefficients are chosen so that

$$\frac{A'_{40}}{A'_{60}} = -\frac{3}{2}, \qquad \frac{A'_{20}}{A'_{60}} = \frac{3}{5}. \tag{22}$$

A glance at Table 9.1, on p. 524, shows that this is precisely the ratio of the corresponding coefficients in the polynomial $R_6^0(\rho)$:

$$R_6^0(\rho) = 20\rho^6 - 30\rho^4 + 12\rho^2 - 1. \tag{23}$$

We see that, provided the aberrations are sufficiently small, the introduction of Zernike's circle polynomials automatically solves the problem of balancing of aberrations, in the sense explained; moreover, with the help of the displacement theorem, it also enables the determination of the position of the diffraction focus.

9.3 Tolerance conditions for primary aberrations

Before considering the difficult problem of determining the intensity distribution in the diffraction image in the presence of aberrations, we consider the much simpler problem of estimating the maximum amounts of aberrations that may be tolerated in an optical system.

It is evident from the discussion of the preceding section that, when aberrations are present, the maximum intensity in the diffraction image is smaller than the intensity at the Gaussian focus (centre of the Airy pattern) in an aberration-free system of the

* R. Richter, *Z. f. Instrumkde.*, **45** (1925), 1.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

same aperture and focal length. It was shown first by Rayleigh[*] that when a system suffers from primary spherical aberration of such an amount that the wave-front in the exit pupil departs from the Gaussian reference sphere by less than a quarter wavelength, the intensity at the Gaussian focus is diminished by less than 20 per cent – a loss of light that can usually be tolerated. Later workers found that in the presence of other commonly occurring aberrations the quality of the image is likewise not seriously affected when the wave-front deformation is less than a quarter of a wavelength. This result has become known as *Rayleigh's quarter wavelength rule* and is a useful criterion for the amount of aberration that can be tolerated in an image-forming system. This rule is, of course, only a rough guide as to the desirable state of correction of a system, since the light distribution in the image depends not only on the maximum deformation but also on the shape of the wave-fronts (type of aberration). Moreover, the loss of light that may be tolerated depends naturally on the particular use to which the instrument is put, and more stringent tolerances have to be imposed in certain cases.

When the conditions $|\Phi_{\text{max}}| = \lambda/4$ is applied to aberrations of different types, somewhat different values for the intensity at the diffraction focus are obtained.[†] It seems more appropriate to formulate tolerance criteria which correspond to a prescribed value of the intensity at the diffraction focus. Criteria of this type were considered by Maréchal,[‡] who used the relation that exists between the intensity at the centre of the reference sphere and the root-mean-square deviation of the wave-front from spherical form.

When the aberrations are sufficiently small, the intensity at a point P in the region of the image may according to §9.1 (24) be expressed in the form

$$i(P) \sim 1 - \left(\frac{2\pi}{\lambda}\right)^2 (\Delta\Phi_P)^2. \tag{1}$$

Following Maréchal *we shall regard a system to be well corrected when the normalized intensity at the diffraction focus F is greater than or equal to* 0.8. Now from (1), $i(F) \geqslant 0.8$ when $|\Delta\Phi_F| \lesssim \lambda/14$, so that this condition is equivalent to the requirement that *the root-mean-square departure of the wave-front from a reference sphere that is centred on the diffraction focus shall not exceed the value* $\lambda/14$.[§]

Let us now determine the position of the diffraction focus and the tolerances for

[*] Lord Rayleigh, *Phil. Mag.*, (5) **8** (1879), 403. Reprinted in his *Scientific Papers*, Vol 1. (Cambridge, Cambridge University Press, 1899), 432–435.

[†] V. N. Mahajan, *loc. cit.*, (1982).

[‡] A. Maréchal, *Rev. d'Opt.*, **26** (1947), 257.

[§] Maréchel's condition is actually based on a somewhat different inequality. Since $\tilde{\Phi}_P$ is constant it follows from §9.1 (20) that

$$\begin{aligned} i(F) &= \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi} \mathrm{e}^{\mathrm{i}k\tilde{\Phi}_F}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2 \\ &\geqslant \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi} \cos(k\tilde{\Phi}_F)\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\right|^2, \end{aligned}$$

where $\tilde{\Phi}_F = \Phi_F - \overline{\Phi}_F$. Now if $k|\tilde{\Phi}_F| < \pi/2$, i.e. if $|\tilde{\Phi}_F| < \lambda/4$, $\cos(k\tilde{\Phi}_F)$ may evidently be replaced by $1 - \frac{1}{2}(k\tilde{\Phi}_F)^2$ in this inequality. Hence for small aberrations, the above inequality becomes

Footnote continued on page 529.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

primary (Seidel) aberrations. In the notation of the present chapter, each primary aberration represents a wave-front deformation of the form*

$$\Phi = a'_{lnm} Y_1^{\star 2l+m} \rho^n \cos^m \theta, \tag{2}$$

where $2l + m + n = 4$. It is convenient to set

$$A'_{lnm} = a'_{lnm} Y_1^{\star 2l+m}, \tag{3}$$

and (2) becomes

$$\Phi = A'_{lnm} \rho^n \cos^m \theta. \tag{4}$$

The constant A' can easily be expressed in terms of the Seidel coefficients B, C, D, E and F, introduced in Chapter V. If we take the arbitrary constant λ_0 in §5.2 (7) and §5.2 (8) equal to unity, then λ_1 denotes the magnification between the pupil planes, and if we remember that we now have $n_1 = 1$, the variables ρ and y_0 of §5.3 (7) correspond to $a\rho/\lambda_1$ and $-\lambda_1 Y_1^\star/R$ of the present section; comparing (4) with §5.3 (7) we obtain

$$\left.\begin{aligned}
A'_{040} &= -\frac{1}{4}\left(\frac{a}{\lambda_1}\right)^4 B,\\
A'_{031} &= -\left(\frac{a}{\lambda_1}\right)^3 \left(\frac{\lambda_1 Y_1^\star}{R}\right) F,\\
A'_{022} &= -\left(\frac{a}{\lambda_1}\right)^2 \left(\frac{\lambda_1 Y_1^\star}{R}\right)^2 C,\\
A'_{120} &= -\frac{1}{2}\left(\frac{a}{\lambda_1}\right)^2 \left(\frac{\lambda_1 Y_1^\star}{R}\right)^2 D,\\
A'_{111} &= -\left(\frac{a}{\lambda_1}\right) \left(\frac{\lambda_1 Y_1^\star}{R}\right)^3 E.
\end{aligned}\right\} \tag{5}$$

In the expansion in terms of the circle polynomials, a typical term represents the aberration†

$$\Phi = \varepsilon_{nm} A_{lnm} R_n^m(\rho) \cos m\theta. \tag{6}$$

The terms with indices l, m and n such that $2l + m + n = 4$ (primary aberrations) are shown in the last column of Table 9.2. It is seen that some of the Seidel terms are now accompanied by terms of lower degrees, and these, according to the displacement theorem (§9.1.2), give rise to a bodily displacement of the intensity distribution. Now according to the theorem of §9.2.2, the image affected by an aberration represented by

$$i(F) \gtrsim \left[1 - \frac{2\pi^2}{\lambda^2}(\Delta\Phi_F)^2\right]^2. \tag{1a}$$

Maréchal's criterion is based on the inequality (1a) but the difference is evidently of no great practical consequence if the aberrations are small. For our purposes, the use of the relation (1) rather than (1a) has the advantage that it is more directly related to the extremal properties of Zernike's circle polynomials.

* We attach a prime to the coefficients in the power series representation, whilst unprimed coefficients refer to the representation in terms of the Zernike circle polynomials.

† The factor ε_{nm} which is equal to unity except for $m = 0$, $n \neq 0$ when it is equal to $1/\sqrt{2}$, is retained here for the sake of uniformity with the formulae of §9.2.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

Table 9.2. *Representation of the primary aberrations.*

Type of aberration	l	n	m	Representation in form (4)	Representation in form (6)
Spherical aberration	0	4	0	$A'_{040}\rho^4$	$\dfrac{1}{\sqrt{2}}A_{040}R_4^0(\rho) = \dfrac{1}{\sqrt{2}}A_{040}(6\rho^4 - 6\rho^2 + 1)$
Coma	0	3	1	$A'_{031}\rho^3\cos\theta$	$\mathrm{A}_{031}R_3^1\cos\theta = A_{031}(3\rho^3 - 2\rho)\cos\theta$
Astigmatism	0	2	2	$A'_{022}\rho^2\cos^2\theta$	$A_{022}R_2^2\cos 2\theta = A_{022}\rho^2(2\cos^2\theta - 1)$
Curvature of field	1	2	0	$A'_{120}\rho^2$	$\dfrac{1}{\sqrt{2}}A_{120}R_2^0(\rho) = \dfrac{1}{\sqrt{2}}A_{120}(2\rho^2 - 1)$
Distortion	1	1	1	$A'_{111}\rho\cos\theta$	$A_{111}R_1^1(\rho)\cos\theta = A_{111}\rho\cos\theta$

(6) has maximum intensity at the Gaussian focus. Hence from a comparison of corresponding terms in the last two columns in Table 9.2, we may immediately determine the coordinates of the diffraction focus of an image affected by a primary aberration. We illustrate this by considering primary spherical aberration in detail. This aberration is represented by

$$\Phi = A'_{040}\rho^4. \tag{7}$$

The corresponding expression (6) is

$$\Phi = \frac{1}{\sqrt{2}}A_{040}R_4^0(\rho) = \frac{1}{\sqrt{2}}A_{040}(6\rho^4 - 6\rho^2 + 1). \tag{8}$$

Now if

$$A'_{040} = \frac{6}{\sqrt{2}}A_{040}, \tag{9}$$

then, according to the displacement theorem, the intensity distribution will be the same in both cases; but the distribution corresponding to (7) will be displaced relative to that corresponding to (8) in accordance with §9.1 (15), with

$$H = 6A_{040}/\sqrt{2} = A'_{040}, \qquad K = L = 0, \qquad M = -A_{040}/\sqrt{2} = A'_{040}/6,$$

i.e. according to the transformation

$$x' = x, \qquad y' = y, \qquad z' = z + 2\left(\frac{R}{a}\right)^2 A'_{040}. \tag{10}$$

Since the diffraction image associated with (8) has maximum intensity at the origin $x = y = z = 0$, the diffraction focus F for primary spherical aberration represented by (7) is at the point

$$x_F = y_F = 0, \qquad z_F = 2\left(\frac{R}{a}\right)^2 A'_{040}. \tag{11}$$

The point F specified by (11) has a simple geometrical interpretation. Let ΔY and ΔZ be the lateral and longitudinal spherical aberrations, measured as positive when the ray intersects the axes on the positive side of the Gaussian focus. By §5.1 (17), with $\Phi = A'_{040}\rho^4$, $\rho = Y/a$, $D_1 \sim -R \sim -R'$, $X_0 = X = 0$, $n_1 = 1$,

$$\Delta Y = Y_1 - Y_1^\star = 4\left(\frac{R}{a}\right)\left(\frac{Y}{a}\right)^3 A'_{040}, \tag{12a}$$

and hence by elementary geometry and the use of the preceding relation

$$\Delta Z = Z - Z_1^\star \sim \frac{R}{Y}\Delta Y = 4\left(\frac{R}{a}\right)^2\left(\frac{Y}{a}\right)^2 A'_{040}. \tag{12b}$$

For the marginal ray ($Y = a$) this gives $(\Delta Z)_{\max} = 4(R/a)^2 A'_{040}$, so that (11) is seen to imply that* *the diffraction focus in the presence of a small amount of primary spherical aberration is situated midway between the paraxial and marginal foci.*

Next let us determine the tolerance for primary spherical aberration. For any aberration characterized by (6) the normalized intensity at the Gaussian focus will be according to §9.2 (14), greater than or equal to 0.8 if

$$1 - \frac{2\pi^2}{\lambda^2}\frac{A_{lnm}^2}{n+1} \geqslant 0.8,$$

i.e. provided that

$$|A_{lnm}| \lesssim \frac{\lambda\sqrt{n+1}}{10}. \tag{13}$$

In particular, for primary spherical aberration this gives

$$|A_{040}| \lesssim 0.22\lambda,$$

or, by (9),

$$|A'_{040}| \lesssim 0.94\lambda. \tag{14}$$

This is the required tolerance condition for primary spherical aberration and implies that the maximum deviation of the wave-front from the Gaussian reference sphere must be less than 0.94 of the wavelength.

In a strictly similar manner we may find the position of the diffraction focus and the tolerance for the other primary aberrations. In particular, the diffraction focus in the presence of a small amount of primary astigmatism is found to be at the point whose coordinates are

$$x_F = y_F = 0, \qquad z_F = \left(\frac{R}{a}\right)^2 A'_{022}. \tag{15}$$

This result has again a simple physical interpretation. According to (5) and §5.3 (18), the radii R_t and R_s of the tangential and sagittal focal surfaces are given by (assuming $n_1 = 1$ again)

$$\frac{1}{R_t} = -\frac{4}{a^2}\left(\frac{R}{Y_1^\star}\right)^2 A'_{022}, \qquad \frac{1}{R_s} = 0, \tag{16}$$

[*] We neglect here the small effect arising from the different choice of the z directions in §5.1 and in the present discussion.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

Table 9.3. *Position of the diffraction focus and tolerance conditions for primary aberrations.*

Types of aberration	Coordinates of diffraction focus F			Tolerance condition $[i(F) \geqslant 0.8]$
	x_F	y_F	z_F	
Spherical aberration	0	0	$2\left(\frac{R}{a}\right)^2 A'_{040}$	$\lvert A'_{040}\rvert \lesssim 0.94\lambda$
Coma	0	$\frac{2}{3}\left(\frac{R}{a}\right) A'_{031}$	0	$\lvert A'_{031}\rvert \lesssim 0.60\lambda$
Astigmatism	0	0	$\left(\frac{R}{a}\right)^2 A'_{022}$	$\lvert A'_{022}\rvert \lesssim 0.35\lambda$
Curvature of field	0	0	$2\left(\frac{R}{a}\right)^2 A'_{120}$	–
Distortion	0	$\left(\frac{R}{a}\right) A'_{111}$	0	–

so that the abscissae Z_t and Z_s of the two focal lines are

$$Z_t = -\frac{Y_1^{\star 2}}{2R_t} = 2\left(\frac{R}{a}\right)^2 A'_{022}, \qquad Z_s = 0. \tag{17}$$

Hence (15) implies that *the diffraction focus in the presence of a small amount of primary astigmatism is situated midway between the tangential and sagittal focal lines.*

Since primary curvature of field and primary distortion are represented by terms of the second and first degree in ρ respectively, it follows, in accordance with the displacement theorem, that the only effect of these aberrations is a ‘bodily shift’ of the three-dimensional distribution associated with an aberration-free image. Thus in the presence of a small amount of primary curvature of field or primary distortion, the normalized intensity i at the diffraction focus is unity, but the diffraction focus does not coincide with the Gaussian image point.

In Table 9.3 the results relating to primary aberrations are summarized.

9.4 The diffraction pattern associated with a single aberration

We now consider the diffraction image in the presence of an aberration represented by a single term of the expansion §9.2 (11),

$$\Phi = a_{lnm} Y_1^{\star 2l+m} R_n^m(\rho) \cos m\theta. \tag{1}$$

As before, we suppress the explicit dependence on $Y_1^\star$, and set

$$\alpha_{lnm} = \frac{2\pi}{\lambda} a_{lnm} Y_1^{\star 2l+m} = \frac{2\pi}{\lambda} \varepsilon_{nm} A_{lnm}. \tag{2}$$

We also get

$$C = -\frac{i\pi A}{\lambda}\left(\frac{a}{R}\right)^2 \mathrm{e}^{\mathrm{i}(R/a)^2 u}. \tag{3}$$

The diffraction integral §9.1 (6) then becomes

$$U(u,\, v,\, \psi) = \frac{C}{\pi}\int_0^1\int_0^{2\pi} \mathrm{e}^{\mathrm{i}[-v\rho\,\cos(\theta-\psi)-\frac{1}{2}u\rho^2+\alpha_{lnm}R_n^m(\rho)\cos\, m\theta]}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta. \tag{4}$$

The integral (4) may be developed in an infinite series, by expanding both the terms $\mathrm{e}^{-\mathrm{i}v\rho\,\cos(\theta-\psi)}$ and $\mathrm{e}^{\mathrm{i}\alpha_{lnm}R_n^m}(\rho)\cos m\theta$ with the help of the Jacobi identity [§8.8 (29)]

$$\mathrm{e}^{\mathrm{i}z\,\cos\,\phi} = J_0(z) + 2\sum_{s=1}^{\infty}\mathrm{i}^s J_s(z)\cos s\phi. \tag{5}$$

Multiplying the two expansions together we find that

$$\mathrm{e}^{\mathrm{i}[-v\rho\,\cos(\theta-\psi)+\alpha_{lnm}R_n^m(\rho)\cos\, m\theta]}$$
$$= 4\sum_{s=0}^{\infty}{}'\sum_{s'=0}^{\infty}{}'\mathrm{i}^s(-\mathrm{i})^{s'}J_s[\alpha_{lnm}R_n^m(\rho)]J_{s'}(v\rho)\cos ms\theta\cos[s'(\theta-\psi)], \tag{6}$$

where the prime on the summation sign implies that the terms in $s = 0$ and $s' = 0$ are each to be taken with a factor $\frac{1}{2}$. We substitute this double series into (4) and integrate with respect to θ term by term. This gives

$$U(u,\, v,\, \psi) = 4C\sum_{s=0}^{\infty}{}'(-\mathrm{i})^{(m-1)s}\cos ms\psi\int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}J_s[\alpha_{lnm}R_n^m(\rho)]J_{ms}(v\rho)\rho\,\mathrm{d}\rho, \tag{7}$$

the term $s = 0$ being again taken with the factor $\frac{1}{2}$.

As we are interested in small aberrations (α small), we develop the term $J_s[\alpha_{lnm}R_n^m(\rho)]$ under the integral sign in a power series, and re-arrange the resulting expression according to powers of α_{lnm}. This gives

$$U(u,\, v,\, \psi) = C[U_0 + \mathrm{i}\alpha_{lnm}U_1 + (\mathrm{i}\alpha_{lnm})^2U_2 + (\mathrm{i}\alpha_{lnm})^3U_3 + (\mathrm{i}\alpha_{lnm})^4U_4 + \ldots], \tag{8a}$$

where

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$\left.\begin{aligned}
U_0 &= 2\int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2} J_0(v\rho)\rho\,\mathrm{d}\rho,\\
U_1 &= 2(-\mathrm{i})^m \cos m\psi \int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2} R_n^m(\rho) J_m(v\rho)\rho\,\mathrm{d}\rho,\\
U_2 &= \frac{1}{2!}\left\{\int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}\{R_n^m(\rho)\}^2 J_0(v\rho)\rho\,\mathrm{d}\rho\right.\\
&\qquad\left. + \mathrm{i}^{2m}\cos 2m\psi \int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}\{R_n^m(\rho)\}^2 J_{2m}(v\rho)\rho\,\mathrm{d}\rho\right\},\\
U_3 &= \frac{1}{2\times 3!}\left\{3(-\mathrm{i})^m \cos m\psi \int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}[R_n^m(\rho)]^3 J_m(v\rho)\rho\,\mathrm{d}\rho\right.\\
&\qquad\left. + (-\mathrm{i})^{3m}\cos 3m\psi \int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}[R_n^m(v\rho)]^3 J_{3m}(v\rho)\rho\,\mathrm{d}\rho\right\},\\
U_4 &= \frac{1}{2^2\times 4!}\left\{3\int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}[R_n^m(\rho)]^4 J_0(v\rho)\rho\,\mathrm{d}\rho\right.\\
&\qquad + 4\mathrm{i}^{2m}\cos 2m\psi \int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}[R_n^m(\rho)]^4 J_{2m}(v\rho)\rho\,\mathrm{d}\rho\\
&\qquad\left. + \mathrm{i}^{4m}\cos 4m\psi \int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}[R_n^m(\rho)]^4 J_{4m}(v\rho)\rho\,\mathrm{d}\rho\right\}.
\end{aligned}\right\}\qquad (8b)$$

Nijboer found that where both u and α_{lnm} are of the order of unity, about four terms in the expansion (8a) suffice to give the intensity to within a few per cent.

To evaluate the integrals in (8b) we may proceed as follows. We express the factor $\mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}$ in terms of the radial polynomials, with the help of the following well-known formula due to Bauer*

$$\mathrm{e}^{\mathrm{i}z\cos\phi} = \left(\frac{\pi}{2z}\right)^{\frac{1}{2}}\sum_{s=0}^{\infty}\mathrm{i}^s(2s+1)J_{s+\frac{1}{2}}(z)P_s(\cos\phi), \qquad (9)$$

where the P's are the Legendre polynomials. If we set $\cos\phi = 2\rho^2 - 1$ and use the relation $P_s(2\rho^2-1) = R_{2s}^0(\rho)$ [§9.2 (8)], it follows that

$$\mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2} = \mathrm{e}^{-\frac{1}{4}\mathrm{i}u}\mathrm{e}^{-\frac{1}{4}\mathrm{i}u(2\rho^2-1)} = \mathrm{e}^{-\frac{1}{4}\mathrm{i}u}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(-\mathrm{i})^s(2s+1)J_{s+\frac{1}{2}}(\tfrac{1}{4}u)R_{2s}^0(\rho). \qquad (10)$$

On substitution from (10) into (8b) integrals are obtained each of which consists of a Bessel function multiplied by a product of the radial polynomials. Now these integrals can be evaluated by the use of the formula §9.2 (9)

* See, for example, G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge, Cambridge University Press, 2nd edition, 1944), p. 368.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$\int_0^1 R_n^m(\rho) J_m(v\rho)\rho\,\mathrm{d}\rho = (-1)^{\frac{n-m}{2}} \frac{J_{n+1}(v)}{v}, \tag{11}$$

provided each of the products of the radial polynomials is expressed as a linear combination of the form $\sum_p A_p R_p^m(\rho)$, with m equal to the order of the Bessel function by which the product is multiplied. It is not easy to obtain a general expression for the coefficients A_p, but if m and n are not too large such linear relations may easily be established with the help of Table 9.1, p. 524, as will be seen later from some examples. For the discussion of methods relating to more general cases we refer the reader to Nijboer's thesis.

In the special case of an aberration-free wave we obtain, on substituting from (10) into (7),

$$\begin{aligned} U(u, v, \psi) &= 2C\mathrm{e}^{-\frac{1}{4}\mathrm{i}u}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(-\mathrm{i})^s(2s+1)J_{s+\frac{1}{2}}(\tfrac{1}{4}u)\int_0^1 R_{2s}^0(\rho)J_0(v\rho)\rho\,\mathrm{d}\rho \\ &= 2C\mathrm{e}^{-\frac{1}{4}\mathrm{i}u}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(\mathrm{i})^s(2s+1)J_{s+\frac{1}{2}}(\tfrac{1}{4}u)\frac{J_{2s+1}(v)}{v}, \end{aligned} \tag{12}$$

where (11) was used. Though formally different, this expansion is equivalent to the series developments of Lommel, given in §8.8.1.

From the expansion (8) some general properties of the diffraction image may immediately be deduced. It is seen that U remains unchanged when ψ is replaced by $\psi + 2\pi\mu/m$ ($\mu = 1, 2, \ldots, m$); hence *the z-axis is an m-fold axis of symmetry.* Moreover, *the planes through the z-axis which makes angles $\pi\mu/m$ with the plane $x = 0$ are planes of symmetry.* When $m = 0$, there is, of course, rotational symmetry.

Next consider the symmetry with respect to the plane $z = 0$. We observe that when u is replaced by $-u$, all integrals in (8) change into their conjugates. Now, if m is odd, the coefficients by which these integrals are multiplied are all real. In this case, therefore, $U(u, v, \psi)$ is changed into its conjugate, and, in consequence, the intensity remains unaltered. Hence, if *m is odd, the intensity distribution is symmetrical with respect to the plane $z = 0$.* If, on the other hand, m is even (but $m \neq 0$), those coefficients which involve a factor $\cos 2\mu m\psi$ (μ being an integer) are real, whereas the others, which involve a factor $\cos(2\mu + 1)m\psi$, are pure imaginary. Therefore, if u is replaced by $-u$ and at the same time ψ is replaced by $\psi + \pi/m$, U now changes into its conjugate. Hence *if m is even but different from zero, the intensity at any point in the plane $z =$ constant is the same as the intensity at the point resulting from reflection in the plane $z = 0$ and additional rotation through an angle π/m* about the *z-axis.* It follows that whereas, as has already been pointed out, the z-axis is in general an m-fold axis of symmetry, *it is, when m is even, a $2m$-fold axis of symmetry with respect to the pattern in the plane $z = 0$.* When $m = 0$ (spherical aberration), the diffraction image is not symmetrical with respect to the plane $z = 0$.

Finally, we observe that *when the sign of the aberration constant α_{lnm} is changed, the intensity distribution is not altered if u is replaced by $-u$ when m is even, and if ψ is replaced by $\psi + \pi$ when m is odd.*

We now briefly consider the structure of the image when the system suffers from a primary aberration of a small amount.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

9.4.1 Primary spherical aberration

In this case $l = m = 0$ and $n = 4$. The aberration function is independent of θ and the three-dimensional diffraction image has rotational symmetry about the principal direction $v = 0$.

According to (8) the expansion of the diffraction integral in powers of α is

$$U(u, v, \psi) = C[U_0 + \mathrm{i}(\alpha_{040})U_1 + (\mathrm{i}\alpha_{040})^2 U_2 + \cdots], \tag{13}$$

where U_0 represents the disturbance in the aberration-free image and $U_1, U_2, \ldots$ are given by the other expressions in (8b), with $m = 0$ and $n = 4$. In particular,

$$U_1 = 2\int_0^1 \mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2} R_4^0(\rho) J_0(v\rho)\rho\, \mathrm{d}\rho. \tag{14}$$

A substitution for $\mathrm{e}^{-\frac{1}{2}\mathrm{i}u\rho^2}$ from (10) gives

$$U_1 = 2\mathrm{e}^{-\frac{1}{2}\mathrm{i}u}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(-\mathrm{i})^s(2s+1)J_{s+\frac{1}{2}}(\tfrac{1}{4}u)\int_0^1 R_{2s}^0(\rho)R_4^0(\rho)J_0(v\rho)\rho\, \mathrm{d}\rho. \tag{15}$$

To evaluate the integral on the right, the product of the radial polynomials will be replaced, as already explained, by a linear combination of radial polynomials whose upper index is equal to the order of the Bessel function (zero in this case). The linear combination is easily found by using the relation §9.2 (8)

$$R_{2s}^0(\rho) = P_s(2\rho^2 - 1), \tag{16}$$

and certain well-known relations involving the Legendre polynomials. We have

$$R_{2s}^0(\rho)R_4^0(\rho) = P_s(2\rho^2 - 1)P_2(2\rho^2 - 1). \tag{17}$$

Now $P_2(t) = \frac{1}{2}(3t^2 - 1)$ so that the right-hand side of (17) may be written as

$$P_s(t)P_2(t) = \tfrac{3}{2}t^2 P_s(t) - \tfrac{1}{2}P_s(t). \tag{18}$$

Applying several times the recurrence relation*

$$tP_s(t) = \frac{1}{2s+1}[(s+1)P_{s+1}(t) + sP_{s-1}(t)], \tag{19}$$

it follows that

$$P_2(t)P_s(t) = a_s P_{s+2}(t) + b_s P_s(t) + c_s P_{s-2}(t), \tag{20}$$

where

$$a_s = \frac{3}{2}\frac{(s+2)(s+1)}{(2s+3)(2s+1)}, \qquad b_s = \frac{(s+1)s}{(2s+3)(2s-1)}, \qquad c_s = \frac{3}{2}\frac{s(s-1)}{(2s+1)(2s-1)}. \tag{21}$$

Hence (17) becomes

$$R_{2s}^0(\rho)R_4^0(\rho) = a_s R_{2s+4}^0(\rho) + b_s R_{2s}^0(\rho) + c_s R_{2s-4}^0(\rho). \tag{22}$$

* See, for example, E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge, Cambridge University Press, fourth edition, 1940), p. 308.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

Finally, substituting from (22) into (15) and using (11), it follows that

$$U_1 = 2\mathrm{e}^{-\frac{1}{2}\mathrm{i}u}\sqrt{\frac{2\pi}{u}}\frac{1}{v}\sum_{s=0}^{\infty}(\mathrm{i})^s(2s+1)J_{s+\frac{1}{2}}(\tfrac{1}{4}u)[a_sJ_{2s+5}(v)+b_sJ_{2s+1}(v)+c_sJ_{2s-3}(v)]. \tag{23}$$

In a similar way* one may obtain series developments for U_2, U_3, ... Using these expansions one may then calculate the intensity $I = |U|^2$ at a number of points in the region of the image and construct the isophotes (lines of equal intensity). In any plane at right angle to the principal direction $v = 0$ the isophotes are, of course, circles.

Fig. 9.3 shows the isophotes in a meridional plane for primary spherical aberration $\Phi = -0.48\lambda\rho^4$, whilst in Figs 9.4 and 9.5 photographs are reproduced which show the appearance of images in various receiving planes in the presence of spherical aberration of somewhat larger amounts.†

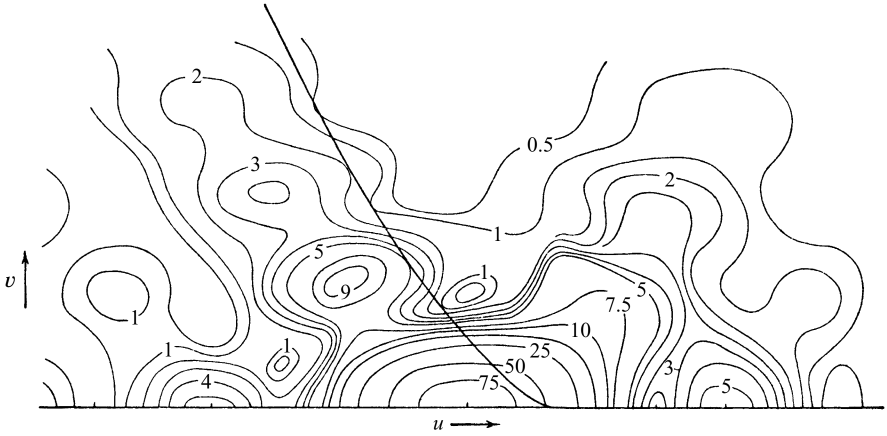


Fig. 9.3 Isophotes in a meridional plane, in presence of primary spherical aberration $\Phi = 0.48\lambda\rho^4$. The thick line indicates the geometrical caustic. The intensity is normalized to 100 at the centre of the aberration-free image. Strehl intensity in best receiving plane: 0.95. (After F. Zernike and B. R. A. Nijboer, contribution to *La Théorie des Images Optiques* (Paris, Revue d'Optique, 1949), p. 232.)

* For details see F. Zernike and B. R. A. Nijboer, contribution to *La Théorie des Images Optiques* (Paris, Revue d'Optique, 1949), p. 227.

† The expansion (13) is unsuitable for computing the intensity when the aberrations are not small compared to the wavelength. Isophotes in a meridional plane for primary spherical aberration of several wavelengths were determined with the help of a mechanical integrator by A. Maréchal, and are published in E. H. Linfoot, *Recent Advances in Optics* (Oxford, Claredon Press, 1955), pp. 60–61, and in M. Françon's contribution to *Encyclopedia of Physics*, Vol, XXIV, (ed. S. Flügge, Berlin, Springer, 1956), pp. 321–322; and by J. Focke, *Optica Acta*, **3** (1956), 110, who used asymptotic approximations. See also J. Picht, *Ann. d. Physik*, (4), **77** (1925), 685.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

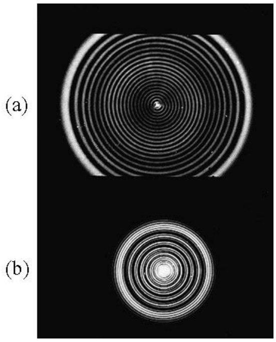


Fig. 9.4 Images in the marginal focal plane (a) and in the plane of the geometrical circle of least confusion (b) in the presence of primary spherical aberration $\Phi = 16\lambda\rho^4$. (After K. Nienhuis, Thesis (University of Groningen, 1948), p. 56.)

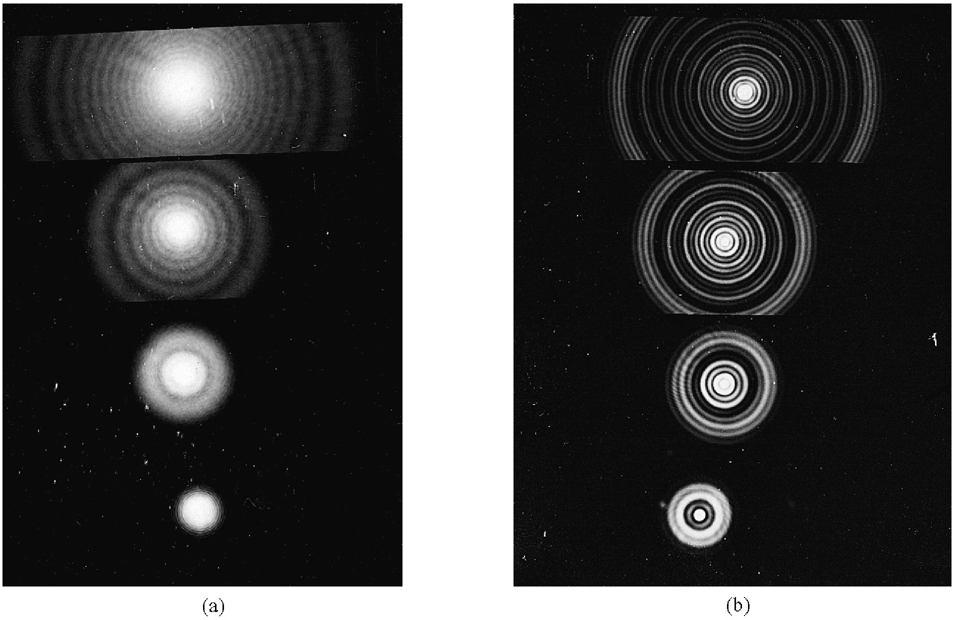


Fig. 9.5 Images in the Gaussian focal plane (a), and in the plane of the geometrical circle of least confusion (b), in the presence of primary spherical aberration $\Phi = 17.5\lambda\rho^4$, $8.4\lambda\rho^4$, $3.7\lambda\rho^4$, and $1.4\lambda\rho^4$. (Scale of (b) is three times that of (a).) (After K. Nienhuis, Thesis (University of Groningen, 1948), p. 56.)

9.4.2 Primary coma

We now have $l = 0$, $n = 3$, $m = 1$. According to Table 9.3, p. 532, the diffraction focus is in the plane $z = 0$ and the disturbance in this plane is given by

$$U(0, v, \psi) = C[U_0(0, v, \psi) + (\mathrm{i}\alpha_{031})U_1(0, v, \psi) + (\mathrm{i}\alpha_{031})^2U_2(0, v, \psi) + \cdots]. \tag{24}$$

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

With $u = 0$, the integral U_0, defined in (8b), represents the disturbance $2J_1(v)/v$ in the focal plane of the aberration-free system (Airy pattern), and U_1 can immediately be evaluated by the use of (11). To evaluate U_2, U_3, ... we must again express the products of the circle polynomials by the appropriate linear combination of such polynomials. In particular, it may be verified with the help of Table 9.1, p. 524, that

$$(R_3^1)^2 = \tfrac{1}{4}R_0^0 + \tfrac{1}{20}R_2^0 + \tfrac{1}{4}R_4^0 + \tfrac{9}{20}R_6^0 = \tfrac{2}{5}R_2^2 + \tfrac{3}{5}R_6^2. \tag{25}$$

Using these relations in the expression for U_2 in (8b) and applying (11) the integrals in U_2 are immediately evaluated and we have in all

$$\left.\begin{aligned} U_0(0, v, \psi) &= \frac{2J_1(v)}{v}, \\ U_1(0, v, \psi) &= \mathrm{i}\cos\psi\,\frac{2J_4(v)}{v}, \\ U_2(0, v, \psi) &= \frac{1}{2v}\{\tfrac{1}{4}J_1(v) - \tfrac{1}{20}J_3(v) + \tfrac{1}{4}J_5(v) - \tfrac{9}{20}J_7(v) \\ &\quad - \cos 2\psi[\tfrac{2}{5}J_3(v) + \tfrac{3}{5}J_7(v)]\}. \end{aligned}\right\} \tag{26}$$

The isophotes for primary coma of various amounts are shown in Figs 9.6 and 9.7. The data for Fig. 9.6 were computed from series expansions, those for Fig. 9.7 by numerical integrations. Photographs showing images affected by primary coma are given in Fig. 9.8. The figures show that when the aberration is of the order of a wavelength, the image resembles neither the Airy pattern nor the pattern predicted by geometrical optics. As the aberration is increased the true image soon becomes of the form specified by geometrical optics, but is broken up by a series of dark bands; these may be shown to arise from the interference of rays diffracted from diametrically opposite points of the aperture.

Fig. 9.6 also illustrates the general result established in §9.2, that when a small aberration is represented in terms of a circle polynomial, the intensity pattern is displaced so as to have its maximum at the origin of the coordinates.

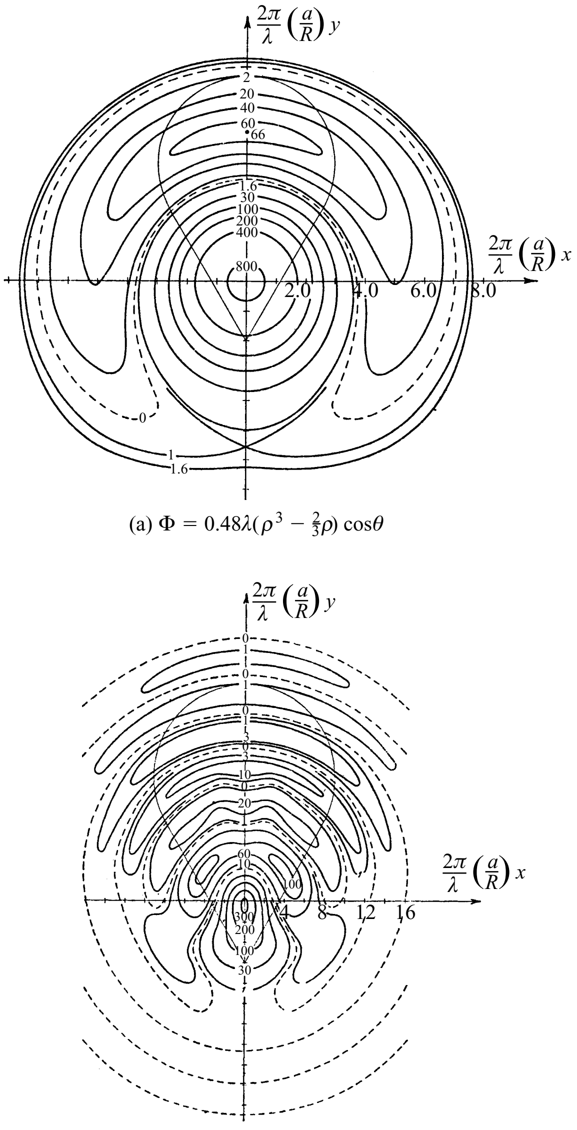
9.4.3 Primary astigmatism

The effect of a small amount of primary astigmatism may be investigated in a similar manner. We now have $l = 0$, $n = m = 2$, and as shown in §9.3 the diffraction focus is midway between the two focal lines. We consider the light distribution in the *central plane*, i.e. the plane through this point, at right angles to the principal direction. When the aberration is represented in terms of the appropriate circle polynomial ($A_{022}R_2^2(\rho)\cos 2\theta$) the central plane is the plane $u = 0$.

The disturbance in the central plane is given by

$$U(0, v, \psi) = C[U_0(0, v, \psi) + (\mathrm{i}\alpha_{022})U_1(0, v, \psi) + (\mathrm{i}\alpha_{022})^2 U_2(0, v, \psi) + \cdots], \tag{27}$$

where $U_0(0, v, \psi)$ represents as before the Airy pattern distribution and $U_1(0, v, \psi)$

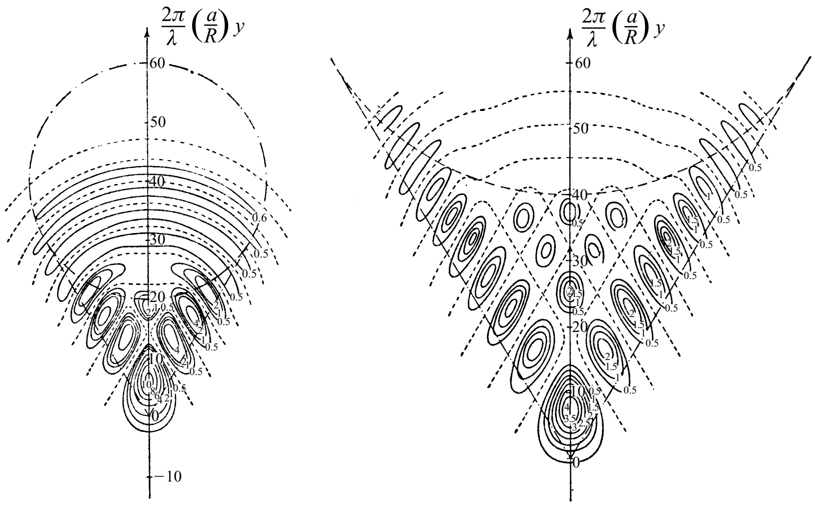


(a) $\Phi = 0.48\lambda(\rho^3 - \frac{2}{3}\rho)\cos\theta$

(b) $\Phi = 1.4\lambda(\rho^3 - \frac{2}{3}\rho)\cos\theta$

Fig. 9.6 Isophotes in the plane $z = 0$ in the presence of primary coma. The dashed curves represent lines of zero intensity. The boundary of the geometrical confusion figures is also shown. The intensity is normalized to 1000 at the centre of the aberration-free image. Strehl intensity: (a) 0.879; (b) 0.306. (Fig. (a) after B. R. A. Nijboer, Thesis (University of Groningen, 1942), p. 62; Fig. (b) after K. Nienhuis and B. R. A. Nijboer, *Physica*, **14** (1949), 599.)

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press



(a) $\Phi = 3.2\lambda\rho^3\cos\theta$ (b) $\Phi = 6.4\lambda\rho^3\cos\theta$

Fig. 9.7 Isophotes in the plane $z = 0$ in the presence of primary coma. The intensity is normalized to 100 at the centre of the aberration-free image. (After R. Kingslake, *Proc. Phys. Soc.*, **61** (1948), 147.)

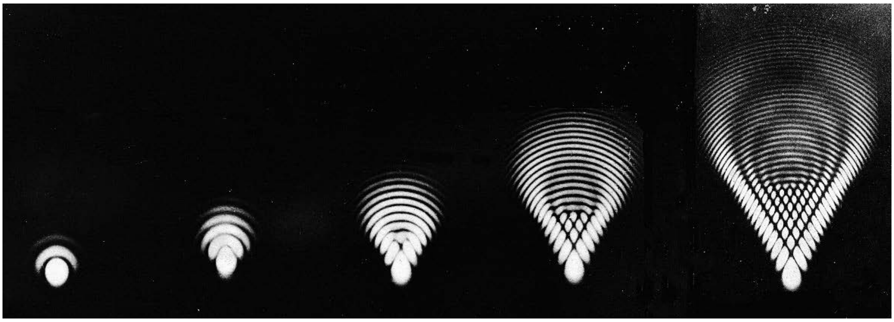


Fig. 9.8 Images in the Gaussian focal plane in the presence of coma $\Phi = 0.3\lambda\rho^3\cos\theta$, $\lambda\rho^3\cos\theta$, $2.4\lambda\rho^3\cos\theta$, $5\lambda\rho^3\cos\theta$, $10\lambda\rho^3\cos\theta$. (After K. Nienhuis, Thesis (University of Groningen, 1948), p. 40.)

may immediately be evaluated with the help of (11). To evaluate U_2 we use the identities

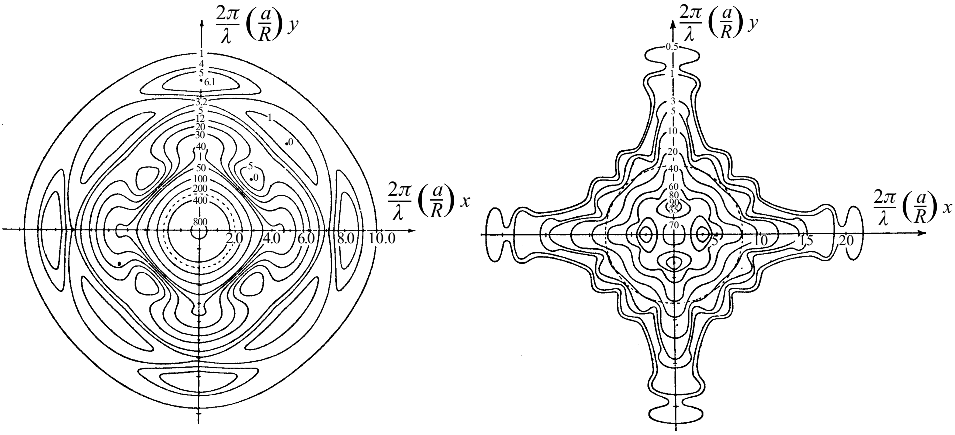
$$\left.\begin{aligned}(R_2^2)^2 &= \tfrac{1}{3}R_0^0 + \tfrac{1}{2}R_2^0 + \tfrac{1}{6}R_4^0,\\ &= R_4^4,\end{aligned}\right\}\qquad(28)$$

and proceed as before. We obtain in all

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$\left.\begin{aligned} U_0(0,\,v,\,\psi) &= \frac{2J_1(v)}{v}, \\ U_1(0,\,v,\,\psi) &= -2\cos 2\psi\,\frac{2J_3(v)}{v}, \\ U_2(0,\,v,\,\psi) &= \frac{1}{2v}\left[\tfrac{1}{3}J_1(v) - \tfrac{1}{2}J_3(v) + \tfrac{1}{6}J_5(v) + \cos 4\psi J_5(v)\right]. \end{aligned}\right\} \qquad (29)$$

In Fig. 9.9 isophote diagrams of astigmatic images are shown. Fig. 9.9(a) was computed from the expansion (27), with terms up to and including the fourth power of α taken into account. Photographs of astigmatic images are given in Figs 9.10 and 9.11. It is seen that when only a small amount of astigmatism is present, the isophotes in the central plane are circular near the centre but have a more complex form in the outer part of the image. When the astigmatism is increased, the image has a cushion-like appearance, and is crossed by interference fringes.



(a) $\Phi = 0.16\lambda\,\rho^2\cos2\theta$

(b) $\Phi = 0.64\lambda\,\rho^2\cos2\theta$

Fig. 9.9 Isophotes in the central plane in the presence of primary astigmatism. The dotted circles represent the boundary of the geometrical confusion figure. The intensity is normalized to 1000 at the centre of the aberration-free image. Strehl intensity: (a) 0.84; (b) 0.066. (Fig. (a) after B. R. A. Nijboer, Thesis (University of Groningen, 1942), p. 55; Fig. (b) after K. Nienhuis, Thesis (University of Groningen, 1948), p. 13.)

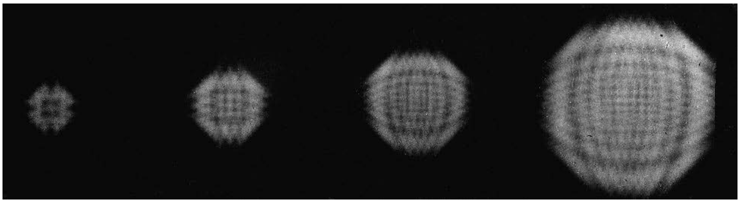


Fig. 9.10 Images in the central plane in the presence of primary astigmatism $\Phi = 1.4\lambda\rho^2\cos 2\theta,\ 2.7\lambda\rho^2\cos 2\theta,\ 3.5\lambda\rho^2\cos 2\theta,\ 6.5\lambda\rho^2\cos 2\theta$. (After K. Nienhuis, Thesis (University of Groningen, 1948), p. 32.)

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

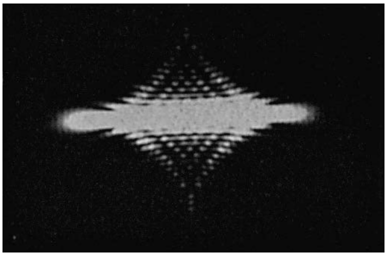


Fig. 9.11 Image in the plane containing a focal line, in the presence of primary astigmatism $\Phi = 2.7\lambda\rho^2 \cos 2\theta$. (After K. Nienhuis, Thesis (University of Groningen, 1948), p. 32.)

As regards the two remaining primary aberrations (curvature and distortion), we have already seen that they do not affect the structure of the three-dimensional image, but only the position of the diffraction focus. The isophote diagrams in the region of focus are therefore identical with that for an aberration-free image (Fig. 8.41), but are displaced relative to the Gaussian focus by the amounts indicated in Table 9.3, p. 532.

9.5 Imaging of extended objects

So far we have been concerned with images of point sources. We shall now describe some general methods, based on the techniques of Fourier transforms, relating to imaging of extended objects. These methods were developed chiefly by Duffieux,* partly in collaboration with Lansraux, and were later extended and applied to particular problems by many writers.†

We consider separately imaging with coherent and imaging with incoherent light.

9.5.1 Coherent illumination

We shall specify points in the Gaussian image plane and in the plane of the exit pupil by Cartesian coordinates (x_1, y_1) and (ξ, η) respectively, referred to parallel axes with their origins at the axial points. Points in the object plane will conveniently be specified by scale-normalized coordinates (x_0, y_0) which are such that if (X_0, Y_0) are the Cartesian coordinates of a typical object point and M the lateral magnification, then

$$x_0 = MX_0, \qquad y_0 = MY_0, \tag{1}$$

so that an object point and its Gaussian image point have now the same coordinate numbers.‡

* P. M. Duffieux, *L'Intégrale de Fourier et ses Applications à l'Optique* (Chez l'Auteur, Rennes, Société Anonyme des Imprimeries Oberthur, 1946), English translation P.-M. Duffieux, *The Fourier Transform and its Applications to Optics* (New York, Wiley, 2nd ed., 1983); P. M. Duffieux and G. Lansraux, *Rev. d'Optique*, **24** (1945), 65, 151, 215.

† See, for example, A. Blanc-Lapierre, *Ann. de l'Inst. Henri Poincaré*, **13** (1953), 245; H. H. Hopkins, *Proc. Roy. Soc.*, A, **217** (1953), 408; *ibid.*, A, **231** (1955), 91; *Proc. Phys. Soc.*, B, **69** (1956), 562; K. Miyamoto, *Progress in Optics*, Vol. I, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1961), p. 41.

‡ (x_0, y_0) and (x_1, y_1) may be regarded as the Seidel variables of §5.2 with the choice $l_1 = C = 1$ of the arbitrary constants.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

The imaging properties of the system may be characterized by means of a *transmission function* $K(x_0, y_0; x_1, y_1)$, defined as the complex amplitude, per unit area of the x_0, y_0-plane, at the point (x_1, y_1) in the Gaussian image plane, due to a disturbance of unit amplitude and zero phase at the object point (x_0, y_0). The transmission function depends, of course, also on the wavelength λ of the light, but as we shall only be concerned with monochromatic light we need not consider this dependence.

Let $U_0(x_0, y_0)$ represent the complex disturbance in the plane of the object. The element at (x_0, y_0) makes a contribution $\mathrm{d}U_1(x_1, y_1) = U_0(x_0, y_0) \times K(x_0, y_0; x_1, y_1)\mathrm{d}x_0\,\mathrm{d}y_0$ to the disturbance at the point (x_1, y_1) in the image plane. Hence the total disturbance at (x_1, y_1) is

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_0(x_0, y_0)K(x_0, y_0; x_1, y_1)\mathrm{d}x_0\,\mathrm{d}y_0. \tag{2}$$

The integral extends only formally over an infinite domain, since U_0K is zero outside the area which does not send light into the image space of the system.

Now when we were dealing with point sources, we specified the properties of the system by the complex disturbance in the exit pupil, and this was characterized by the aberration function and by an amplitude factor, the latter being assumed to be constant in systems of moderate aperture. It is not difficult to find an expression for the transmission function in terms of these quantities. For this purpose we consider first the limiting form of (2) when the source reduces to a point source of unit strength and zero phase at the point $x_0 = x_0'$, $y_0 = y_0'$, i.e. when

$$U_0(x_0, y_0) = \delta(x_0 - x_0')\delta(y_0 - y_0'), \tag{3}$$

where δ is the Dirac delta function (see Appendix IV). Then (2) gives

$$U_1(x_1, y_1) = K(x_0', y_0'; x_1, y_1), \tag{4}$$

i.e. the transmission function K represents the disturbance due to the point source (3). Let us take the Gaussian reference sphere with centre at the Gaussian image point $x_1' = x_0'$, $y_1' = y_0'$. Let R be the radius of this reference sphere and let

$$H(x_0', y_0'; \xi, \eta) = \frac{\mathrm{i}}{\lambda} G(x_0', y_0'; \xi, \eta)\frac{\mathrm{e}^{-\mathrm{i}kR}}{R} \tag{5}$$

be the disturbance at a typical point (ξ, η) on this sphere, due to the point source (3). Apart from an additive factor $\pi/2$, the phase of G is then the aberration function Φ of the system, whilst the amplitude of G is a measure of the nonuniformity in the amplitude of the image-forming wave. The factor i/λ has been introduced on the right-hand side of (5) to simplify later formulae. Now by the Huygens–Fresnel principle, the disturbance in the image plane is related to the disturbance on the Gaussian reference sphere by the formula (small angles of diffraction assumed)

$$U_1(x_1, y_1) = -\frac{\mathrm{i}}{\lambda}\iint H(x_0', y_0'; \xi, \eta)\frac{\mathrm{e}^{\mathrm{i}ks}}{\mathrm{s}}\mathrm{d}\xi\,\mathrm{d}\eta, \tag{6}$$

where s is the distance from the point (ξ, η) of this sphere to the point (x_1, y_1) in the Gaussian image plane, and the integral extends over the portion of the reference sphere that approximately fills the aperture. We also have according to §8.8 (2) and §8.8 (7), with $x = x_1 - x_0'$, $y = y_1 - y_0'$, $z = 0$ and with R in place of f,

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$s \sim R - \frac{(x_1 - x_0')\xi + (y_1 - y_0')\eta}{R}. \tag{7}$$

From the formulae (4)–(7) we obtain

$$K(x_0, y_0; x_1, y_1) = \frac{1}{(\lambda R)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} G(x_0, y_0; \xi, \eta)\mathrm{e}^{-\frac{2\pi\mathrm{i}}{\lambda R}[(x_1 - x_0)\xi + (y_1 - y_0)\eta]}\mathrm{d}\xi\,\mathrm{d}\eta, \tag{8}$$

G being taken to be zero at points (ξ, η) which are outside the opening. This is the required relation between the transmission function K and the *'pupil function'* G of the system.

Since K may be regarded as the disturbance in the image of a point source, it has, when considered as a function of x_1, y_1, a fairly sharp maximum at or near the Gaussian image point $x_1 = x_0$, $y_1 = y_0$ and falls off rapidly, though as a rule not monotonically, with increasing distance from this point. In a well-corrected system K will only be appreciable in an area whose size is of the order of the first dark ring of the Airy pattern. Considered as a function of (x_0, y_0), the transmission function varies slowly as this point explores the object surface. More precisely, the working field may be divided up into regions, each large compared to the finest detail that the system can resolve, with the property that in each such region A, K is to a good approximation a function of the displacement vector from the Gaussian image point, but not of the position of the image point itself. For example, in a well-corrected system, $K(x_0, y_0; x_1, y_1)$ represents, apart from a constant factor, the Airy pattern, centred on the Gaussian image point of (x_0, y_0). In such cases we may write

$$K(x_0, y_0; x_1, y_1) = K_A(x_1 - x_0, y_1 - y_0). \tag{9}$$

A region A with this property is said to be an *isoplanatic region* of the system. We shall restrict out discussion to objects that are so small that they fall within such an isoplanatic region.* In this case the equations (2) and (8) may be replaced by

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_0(x_0, y_0)K(x_1 - x_0, y_1 - y_0)\mathrm{d}x_0\,\mathrm{d}y_0, \tag{2a}$$

and

$$K(x_1 - x_0, y_1 - y_0) = \frac{1}{(\lambda R)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} G(\xi, \eta)\mathrm{e}^{-\frac{2\pi\mathrm{i}}{\lambda R}[(x_1 - x_0)\xi + (y_1 - y_0)\eta]}\mathrm{d}\xi\,\mathrm{d}\eta, \tag{8a}$$

the function G now being independent of the object point.

Let us represent U_0 U_1, and K as Fourier integrals:

$$U_0(x_0, y_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathcal{U}_0(f, g)\mathrm{e}^{-2\pi\mathrm{i}[fx_0 + gy_0]}\mathrm{d}f\,\mathrm{d}g, \tag{10a}$$

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathcal{U}_1(f, g)\mathrm{e}^{-2\pi\mathrm{i}[fx_1 + gy_1]}\mathrm{d}f\,\mathrm{d}g, \tag{10b}$$

$$K(x, y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathcal{K}(f, g)\mathrm{e}^{-2\pi\mathrm{i}[fx + gy]}\mathrm{d}f\,\mathrm{d}g. \tag{10c}$$

* A thorough discussion of the conditions under which (9) holds has been published by P. Dumontet, *Optica Acta*, **2** (1955), 53.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

Then, by the Fourier inversion formula

$$\mathcal{U}_0(f,\, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_0(x_0,\, y_0)\mathrm{e}^{2\pi\mathrm{i}[fx_0+gy_0]}\mathrm{d}x_0\,\mathrm{d}y_0, \tag{11a}$$

$$\mathcal{U}_1(f,\, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_1(x_1,\, y_1)\mathrm{e}^{2\pi\mathrm{i}[fx_1+gy_1]}\mathrm{d}x_1\,\mathrm{d}y_1, \tag{11b}$$

$$\mathcal{K}(f,\, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K(x,\, y)\mathrm{e}^{2\pi\mathrm{i}[fx+gy]}\mathrm{d}x\,\mathrm{d}y. \tag{11c}$$

According to (2a) U_1 is a convolution (also called resultant or *Faltung*) of U_0 and K; and on Fourier inversion we obtain by the convolution theorem,* the simple relation

$$\mathcal{U}_1(f,\, g) = \mathcal{U}_0(f,\, g)\mathcal{K}(f,\, g). \tag{12}$$

This equation implies that if the disturbances in the object plane and in the image plane are each considered as a superposition of space-harmonic components of all possible ‘spatial frequencies’ f, g, then each component of the image depends only on the corresponding component of the object, and the ratio of the components is $\mathcal{K}$. Thus the transition from the object to the image is equivalent to the action of a *linear filter*. Moreover, comparison of (10c) and (8a) shows that

$$\mathcal{K}\left(\frac{\xi}{\lambda R}, \frac{\eta}{\lambda R}\right) = G(\xi,\, \eta), \tag{13}$$

so that *the frequency response function* (also called the *transmission factor)* $\mathcal{K}(f,\, g)$ *for coherent illumination is equal to the value of the pupil function G at the point*

$$\xi = \lambda Rf, \qquad \eta = \lambda Rg \tag{14}$$

of the Gaussian reference sphere.

Since G is zero at points in the ξ, η-plane which are outside the boundary of the aperture, the spectral amplitudes belonging to frequencies above a certain value are not transmitted by the system. If the aperture is circular, and of radius a, then evidently frequency pairs such that

$$f^2 + g^2 > \left(\frac{a}{\lambda R}\right)^2 \tag{15}$$

are not transmitted. To illustrate this result consider a one-dimensional object the properties of which do not vary in the x direction. Let Δy_0 be the period belonging to the frequency g. Then by (15) the system can only transmit information about spectral components for which

$$\Delta y_0 = \frac{1}{g} > \frac{\lambda}{\sin\theta_1}, \tag{16}$$

where $\sin\theta_1 \sim a/R$ is the angular semiaperture on the image side, assumed to be small. Now $y_0 = MY_0$, where M is the linear magnification, and if we assume that the

* See I. N. Sneddon, *Fourier Transforms* (New York, McGraw-Hill, 1951), p. 23.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

system obeys the sine condition then (§4.5.1) $n_0 \sin\theta_0/n_1 \sin\theta_1 = M$, and (16) may be written as

$$\Delta Y_0 > \frac{\lambda_0}{n_0 \sin\theta_0}, \tag{17}$$

where $\lambda_0 = n_1\lambda$ is the vacuum wavelength and $n_0 \sin\theta_0$ is the numerical aperture of the system. Thus if the disturbance across the object plane varies sinusoidally with displacement, information about it can only be obtained in the image plane if the period exceeds the value given by the right-hand side of (17).

9.5.2 Incoherent illumination

We now consider the case when the light that emanates from the different elements of the object plane is incoherent, e.g. when the object is a primary source. Using the same coordinates as before, let $I_0(x_0, y_0)$ be the intensity at a typical point in the object plane. The intensity of the light that reaches the point (z_1, y_1) in the plane of the image from the element $\mathrm{d}x_0\,\mathrm{d}y_0$ centred on the object point (x_0, y_0) is $\mathrm{d}I_1(x_1, y_1) = I_0(x_0, y_0)|K(x_0, y_0; x_1, y_1)|^2\,\mathrm{d}x_0\,\mathrm{d}y_0$ where K is again the transmission function of the system. Since the object is assumed to be incoherent, the intensities from the different elements of the object plane are additive, so that the total intensity at (x_1, y_1) is given by

$$I_1(x_1, y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_0(x_0, y_0)|K(x_0, y_0; x_1, y_1)|^2\,\mathrm{d}x_0\,\mathrm{d}y_0. \tag{18}$$

If we again restrict ourselves to sufficiently small objects, we may replace (18) by*

$$I_1(x_1, y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_0(x_0, y_0)|K(x_1 - x_0, y_1 - y_0)|^2\,\mathrm{d}x_0\,\mathrm{d}y_0. \tag{19}$$

Eq. (19) shows that for imaging with incoherent illumination the intensity distribution in the image is a convolution of the intensity distribution in the object with the squared modulus of the transmission function. We represent these functions as Fourier integrals of the form (10), and denote by $\mathcal{I}_0(f, g)$, $\mathcal{I}_1(f, g)$ and $\mathcal{L}(f, g)$ their 'spatial spectra'. Then, by the Fourier inversion theorem, we have, in place of (11),

* We note that to replace (18) by (19) it is not necessary that K should satisfy the full isoplanatic condition (9); it is sufficient that the modulus of K alone satisfies it, i.e. that throughout the region A occupied by the object, we have to a good approximation

$$|K(x_0, y_0; x_1, y_1)| = |K_A(x_1 - x_0, y_1 - y_0)|.$$

The analysis of Dumontet (*loc. cit.*) shows that as a rule this condition holds to a good approximation over a considerably larger region of the object plane than the relation (9). Hence the representation of optical imaging as a linear filter has a wider range of validity for incoherent than for coherent illumination. However, to be able to express the frequency response function in terms of the pupil function of the system by a relatively simple formula [(22) below], we restrict ourselves to objects that are so small that the full isoplanatic condition holds.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

$$\mathcal{I}_0(f, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_0(x_0, y_0)\mathrm{e}^{2\pi\mathrm{i}[fx_0+gy_0]}\mathrm{d}x_0\,\mathrm{d}y_0, \tag{20a}$$

$$\mathcal{I}_1(f, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_1(x_1, y_1)\mathrm{e}^{2\pi\mathrm{i}[fx_1+gy_1]}\mathrm{d}x_1\,\mathrm{d}y_1, \tag{20b}$$

$$\mathcal{L}(f, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |K(x, y)|^2\mathrm{e}^{2\pi\mathrm{i}[fx+gy]}\mathrm{d}x\,\mathrm{d}y. \tag{20c}$$

From (19) we obtain by the convolution theorem

$$\mathcal{I}_1(f, g) = \mathcal{I}_0(f, g)\mathcal{L}(f, g). \tag{21}$$

Thus the transformation from the object to the image is again a *linear filter*, but it is now the spatial spectrum of the intensity and not of the complex amplitude that is transformed in this way. The frequency response function is the function $\mathcal{L}(f, g)$, and this, by (20c), (10c) and the convolution theorem, may be expressed in the form

$$\mathcal{L}(f, g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathcal{K}(f'+f, g'+g)\mathcal{K}^{\star}(f'+g')\mathrm{d}f'\,\mathrm{d}g'. \tag{22}$$

The integral on the right-hand side of (22) is known as the *autocorrelation function* (of the function $\mathcal{K}$) and occurs in the analysis of many physical problems of statistical nature. We shall encounter it again later in connection with the theory of partial coherence.

We have already shown that $\mathcal{K}(f, g)$ is the value of the pupil function G at an appropriate point of the Gaussian reference sphere. If we substitute from (13) into (22) it follows that

$$\mathcal{L}\left(\frac{\xi}{\lambda R}, \frac{\eta}{\lambda R}\right) = \frac{1}{(\lambda R)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} G(\xi'+\xi, \eta'+\eta)G^{\star}(\xi', \eta')\mathrm{d}\xi'\,\mathrm{d}\eta', \tag{23}$$

and we have thus established the important result that *apart from a constant factor, the frequency response function* $\mathcal{L}(f, g)$ *for incoherent illumination is the autocorrelation function of the pupil function of the system.*

Let $\mathcal{A}$ be the area of the exit pupil. Since the pupil function $G(\xi', \eta')$ is zero at points outside the boundary of the aperture, the domain of the ξ', η'-plane over which the integrand of (23) does not vanish is the area common to the aperture $\mathcal{A}$ and to an identical aperture displaced relative to $\mathcal{A}$ by a translation of amounts ξ and η in the negative ξ' and η' directions respectively (see Fig. 9.12). When ξ and η are so large that the two areas do not overlap, the value of the response function is evidently zero; thus, as in the coherent case, spatial frequencies only up to certain maximum values are transmitted by the system. In particular, for a circular aperture of radius a, the areas will have no region in common when $\xi^2 + \eta^2 \geqslant (2a)^2$, i.e. when

$$f^2 + g^2 > \left(\frac{2a}{\lambda R}\right)^2. \tag{24}$$

By the same argument as that leading from (15) to (17), this implies that, with incoherent illumination, an aplanatic system can only transmit information about the spectral components whose period ΔY_0 is such that

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

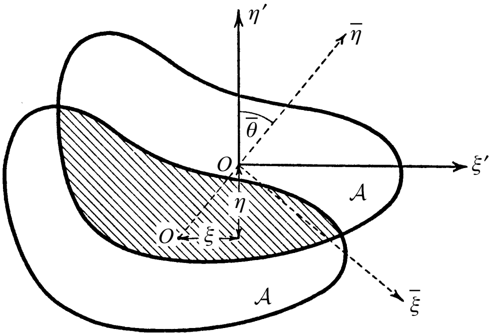


Fig. 9.12 Region of integration (shown shaded) relating to the evaluation of the response function $\mathcal{L}(f, g)$ for incoherent illumination, for the frequency pair $f = \xi/\lambda R$, $g = \eta/\lambda R$.

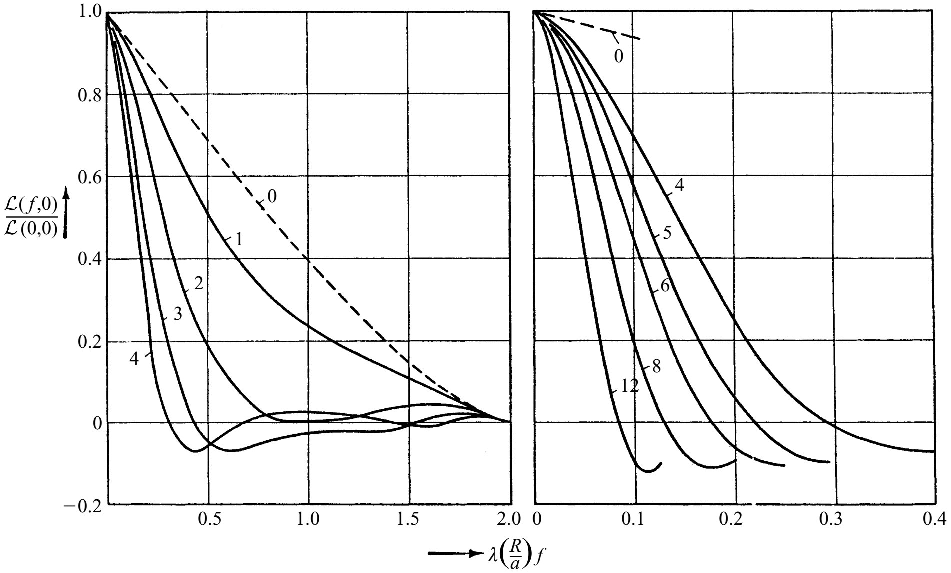


Fig. 9.13 The normalized frequency response curves for incoherent illumination of a system free of geometrical aberrations but suffering from defect of focus. $\Phi = (m\lambda/\pi)\rho^2$, $|G| = 1$. The number on each curve is the value of the parameter $m = (\pi/2\lambda)(a/R)^2 z$, z being the distance between the plane of observation and the Gaussian focal plane. (After H. H. Hopkins, *Proc, Roy. Soc.*, A, **231** (1955), 98.)

$$\Delta Y_0 > \frac{0.5\lambda_0}{n_0 \sin\theta_0}. \tag{25}$$

It is seen that the limiting value is precisely one-half of the value obtained for imaging with coherent light.

Although the response function $\mathcal{L}$ of a given system depends on two variables f and g, it is possible, in principle, to deduce all information about it from experiments involving one-dimensional test objects. To see this consider a frequency pair (f, g)

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

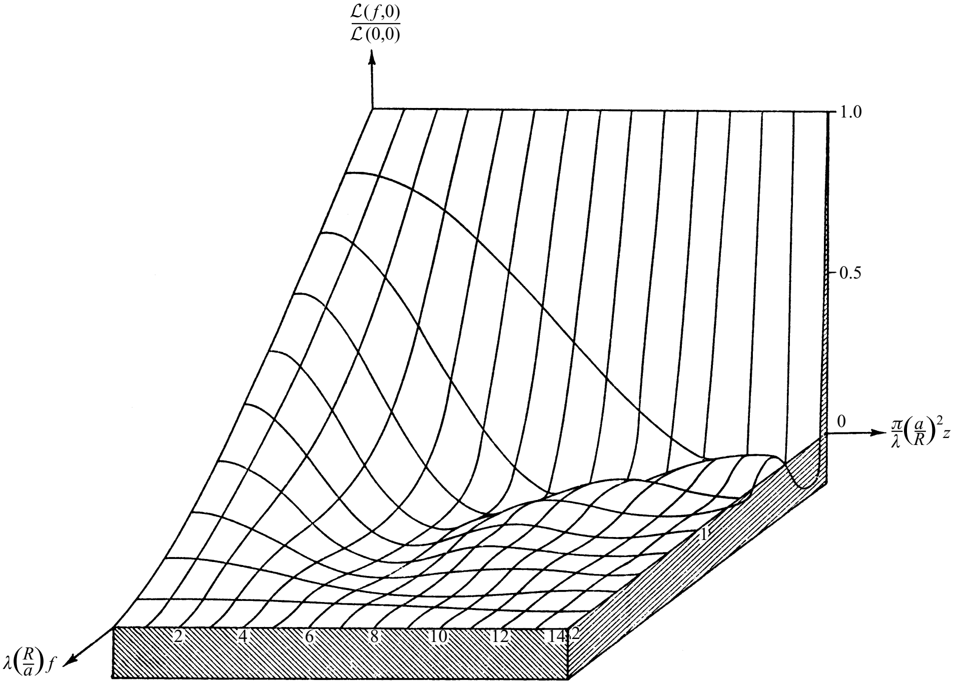


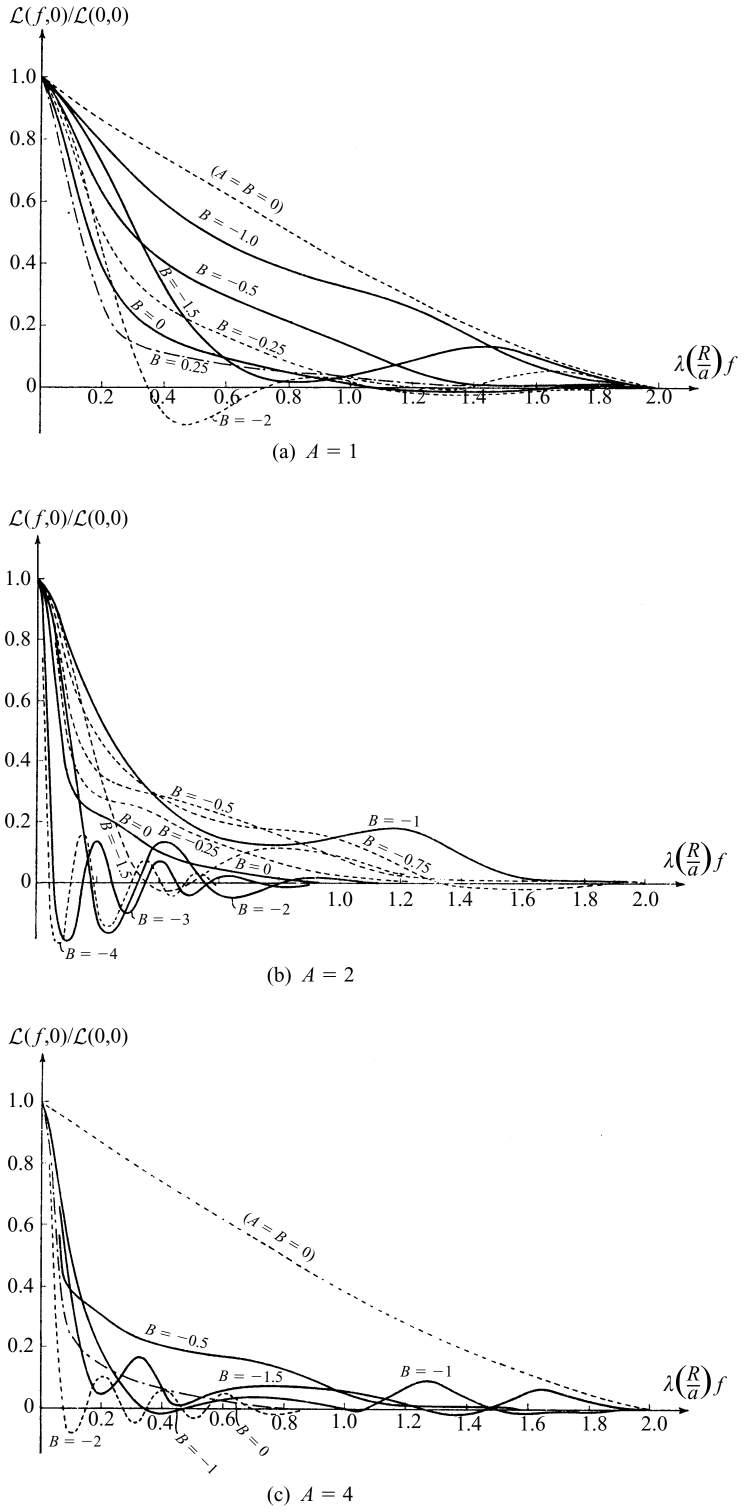
Fig. 9.14 The normalized frequency response curves for incoherent illumination of a system free of geometrical aberrations but suffering from defect of focus, as a function of the spatial frequency f and of defocusing z, $|G| = 1$. The curves shown in Fig. 9.13 are the sections of the surface by planes at right angles to the z-axis. (After W. H. Steel, *Optica Acta*, **3** (1956), 67.)

and introduce polar coordinates such that $f = h\sin\theta$, $g = h\cos\theta$. Suppose now that the axes are rotated in their own plane through an angle $\overline{\theta}$ in the positive θ direction. Then f and g transform into $\overline{f} = h\sin(\theta - \overline{\theta})$ and $\overline{g} = h\cos(\theta - \overline{\theta})$, but the value of $\mathcal{L}$ evidently remains unchanged. Now we may choose the angle of rotation $\overline{\theta}$ equal to $\tan^{-1} f/g$, which corresponds to taking the new η-axis ($O\overline{\eta}$) along the line OO (see Fig. 9.12). Then $\overline{f} = 0$, and $\overline{g} = \sqrt{f^2 + g^2}$, and it follows that *the value of the response function* $\mathcal{L}$ *of an optical system for the frequency pair* (f, g), *is equal to that for a one-dimensional structure of frequency* $\sqrt{f^2 + g^2}$ *with its direction of periodicity inclined at the angle* $\tan^{-1} f/g$ *to the meridional plane* $\theta = 0$. This result considerably simplifies the analytical evaluation of the response functions in any given case. A similar result evidently holds in connection with the response function $\mathcal{K}(f, g)$ of coherent objects but the result is of less immediate practical interest.

Let us now consider the frequency response of a centred system that is free of aberrations, but suffers from defect of focus. It follows from the discussion of §9.1.2 that the displacement of the receiving plane by a small amount z in the positive z direction is formally equivalent to the introduction of a wave aberration of amount

$$\Phi(\xi, \eta) = \frac{1}{2}\left(\frac{a}{R}\right)^2 z\rho^2 \qquad \left(\rho^2 = \frac{\xi^2 + \eta^2}{a^2} \leqslant 1\right), \tag{26}$$

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press



(a) $A = 1$

(b) $A = 2$

(c) $A = 4$

Fig. 9.15 The normalized frequency response curves for incoherent illumination, at selected focal settings of a system suffering from a small amount of primary spherical aberration $\Phi = A(\rho^4 + B\rho^2)\lambda$, $|G| = 1$. The value $B = 0$ corresponds to the paraxial focal plane and $B = -2$ corresponds to the receiving plane through the marginal focus. (After G. Black and E. H. Linfoot, *Proc. Roy. Soc.*, A, **239** (1957), 522.)

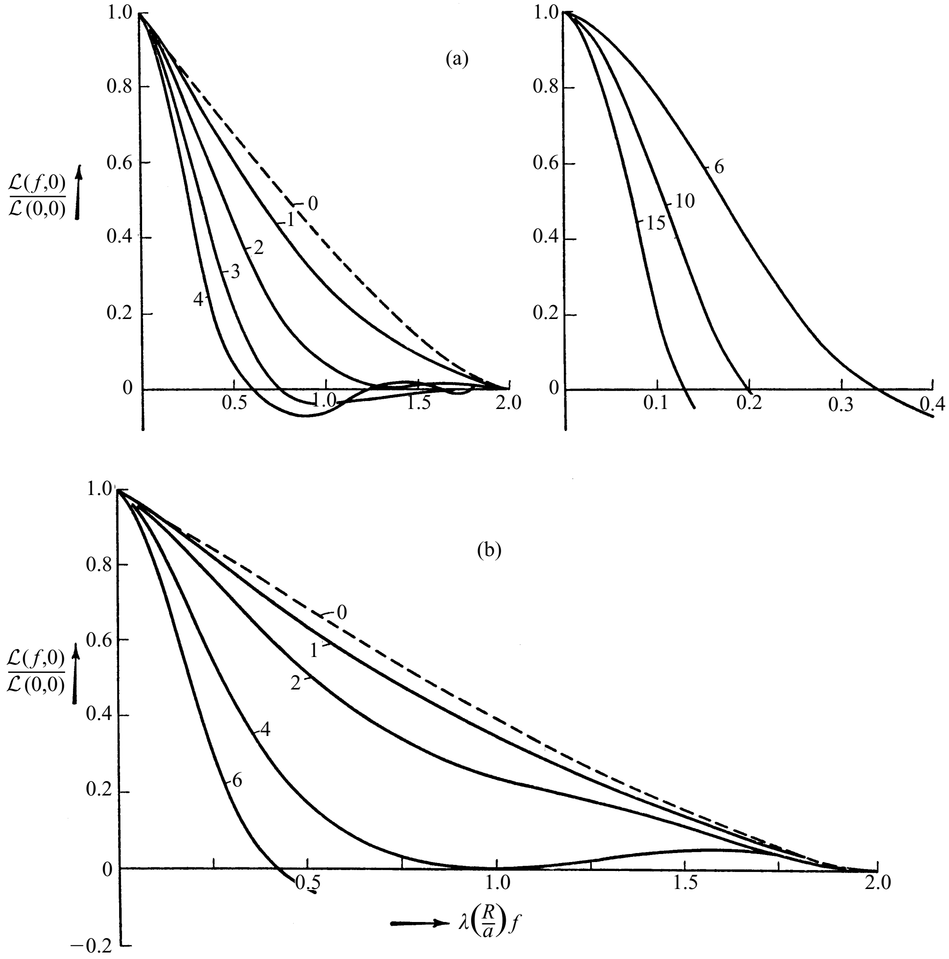


Fig. 9.16 The normalized frequency response curves for incoherent illumination, in the presence of primary astigmatism $\Phi = (m\lambda/\pi)\rho^2 \cos^2\theta$, $|G| = 1$, for a receiving plane midway between the tangential and sagittal focal lines. Line periodic along the meridian $\theta = \pi/4$[(a)], and along the meridians $\theta = 0$ or $\theta = \pi/2$[(b)]. The number on each curve is the value of m. (After M. De, *Proc. Roy. Soc.*, A, **233** (1955), 96.)

so that, if the amplitude of the wave is constant over the Gaussian reference sphere, the pupil function is, apart from a constant factor,

$$G(\xi, \eta) = \mathrm{e}^{\mathrm{i}k\Phi(\xi,\eta)} = \mathrm{e}^{\mathrm{i}\frac{1}{2}k\left(\frac{a}{R}\right)^2 z\rho^2}. \tag{27}$$

The response function may be determined from (23) and (27) and the results are shown in Figs 9.13 and 9.14. It is seen that there is a very rapid deterioration of the response of the system for higher frequencies, with the introduction of a small amount of defect of focus in excess of the value corresponding to $\Phi = (\lambda/\pi)\rho^2$, i.e. in excess of $z = (2\lambda/\pi)(R/a)^2$.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press

Figs 9.15 and 9.16 show the response curves for systems suffering from primary spherical aberration and primary astigmatism.

A survey of instruments for measuring frequency response functions has been given by K. Murata.*

* K. Murata, *Progress in Optics*, Vol. 5, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 199.

https://doi.org/10.1017/9781108769914.012 Published online by Cambridge University Press