IV

Geometrical theory of optical imaging

4.1 The characteristic functions of Hamilton

In §3.1 it was shown that, within the approximations of geometrical optics, the field may be characterized by a single scalar function $\mathcal{S}(\mathbf{r})$. Since $\mathcal{S}(\mathbf{r})$ satisfies the eikonal equation §3.1 (15), this function is fully specified by the refractive index function $n(\mathbf{r})$ alone, together with the appropriate boundary conditions.

Instead of the function $\mathcal{S}(\mathbf{r})$, closely related functions known as *characteristic functions* of the medium are often used. They were introduced into optics by W. R. Hamilton, in a series of classical papers.* Although on account of algebraic complexity it is impossible to determine the characteristic functions explicitly for all but the simplest media, Hamilton's methods nevertheless form a very powerful tool for systematic analytical investigations of the general properties of optical systems.

In discussing the properties of these functions and their applications, an isotropic but generally heterogeneous medium will be assumed.

4.1.1 The point characteristic

Let (x_0, y_0, z_0) and (x_1, y_1, z_1) be respectively the coordinates of two points P_0 and P_1 each referred to a different set of mutually parallel, rectangular axes† (Fig. 4.1). If the two points are imagined to be joined by all possible curves, there will, in general, be some amongst them, the optical rays, which satisfy Fermat's principle. Assume for the present that not more than one ray joins any two arbitrary points. The characteristic function V, or the *point characteristic*, is then defined as *the optical length* $[P_0P_1]$ *of the ray between the two points*, considered as a function of their coordinates,

* Sir W. R. Hamilton, *Trans. Roy. Irish Acad.*, **15** (1828), 69; *ibid*., **16** (1830), 1; *ibid*., **16** (1831), 93; *ibid*., **17** (1837), 1. Reprinted in *The Mathematical Papers of Sir W. R. Hamilton*, Vol. I (*Geometrical Optics*), eds. A. W. Conway and J. L. Synge (Cambridge, Cambridge University Press, 1931).

Many years later Bruns independently considered similar functions which he called *eikonals* (H. Bruns, *Abh. Kgl. Sächs. Ges. Wiss., math-phys. Kl.*, **21** (1895), 323). As already mentioned on p. 119, this term has come to be used in a wider sense. The characteristic functions of Hamilton are themselves often referred to as eikonals.

A useful introduction to Hamilton's methods is a monograph by J. L. Synge, *Geometrical Optics* (Cambridge, Cambridge University Press, 1937). The relationship between the work of Hamilton and Bruns was discussed by F. Klein in *Z. Math. Phys*., **46** (1901), 376, and *Ges. Math. Abh*., **2** (1922), 603, C. Carathéodory, *Geometrische Optik* (Berlin, Springer, 1937), p. 4, and in a polemic between M. Herzberger and J. L. Synge, *J. Opt. Soc. Amer.*, **27** (1937), 75, 133, 138.

† The use of two reference systems has some advantages, since P_0 and P_1 are often situated in different regions, namely, the object- and image-spaces of an optical system.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

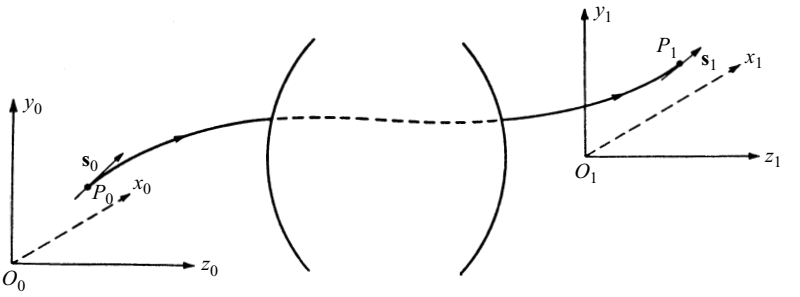

Fig. 4.1 Illustrating the definition of the point characteristic function.

$$V(x_0,\ y_0,\ z_0;\ x_1,\ y_1,\ z_1) = \int_{P_0}^{P_1} n\,\mathrm{d}s. \tag{1}$$

It is important to note that this function is defined *by the medium*.

From (1) and §3.1 (26) it follows that

$$V(x_0,\ y_0,\ z_0;\ x_1,\ y_1,\ z_1) = \mathcal{S}(x_1,\ y_1,\ z_1) - \mathcal{S}(x_0,\ y_0,\ z_0), \tag{2}$$

where the function $\mathcal{S}$ is now associated with any pencil of rays to which the natural ray through P_0 and P_1 belongs (for example, a pencil produced by a point source at P_0).* Then by §3.1 (24), the unit vectors $\mathbf{s}_0$ and $\mathbf{s}_1$ at P_0 and P_1, in the direction of the ray, are given by

$$\left.\begin{aligned} \operatorname{grad}^0 V &= -n_0\mathbf{s}_0, \\ \operatorname{grad}^1 V &= \ \ n_1\mathbf{s}_1, \end{aligned}\right\} \tag{3}$$

the superscripts 0 and 1 implying that the operator grad is taken with respect to the coordinates $(x_0,\ y_0,\ z_0)$ and $(x_1,\ y_1,\ z_1)$ respectively.

The vector

$$\mathbf{g} = n\mathbf{s} \tag{4}$$

is sometimes called the *ray vector*. If α, β and γ are the angles which the ray vector makes with the coordinate axes, its projections†

$$p = n\cos\alpha, \qquad q = n\cos\beta, \qquad m = n\cos\gamma, \tag{5}$$

onto the axes are called the *ray components*. On account of the identity

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1,$$

they satisfy the relation

$$p^2 + q^2 + m^2 = n^2. \tag{6}$$

* In the language of the calculus of variations, $\mathcal{S}$ represents the solution which involves a two-parameter family (∞^2) of extremals, of the Hamilton–Jacobi equation, associated with Fermat's variational problem. The point characteristic function V represents the general solution, which involves all (∞^4) extremals (see Appendix I).

† This 'asymmetrical' notation is purposely chosen here to remind us that on account of the identity (6), only two of the ray components are independent.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

From (3) it follows that the ray components at P_0 and P_1 are given by

$$p_0 = -\frac{\partial V}{\partial x_0}, \qquad p_1 = \frac{\partial V}{\partial x_1}, \tag{7}$$

with similar expressions for q_0, q_1, and m_0, m_1. These relations show that, from the knowledge of the point characteristic, the components of the ray which joins any two points in the medium can immediately be determined. Further it follows from (6) and (7) that the point characteristic satisfies the eikonal equation in both sets of variables:

$$\left(\frac{\partial V}{\partial x_0}\right)^2 + \left(\frac{\partial V}{\partial y_0}\right)^2 + \left(\frac{\partial V}{\partial z_0}\right)^2 = n_0^2, \tag{8}$$

$$\left(\frac{\partial V}{\partial x_1}\right)^2 + \left(\frac{\partial V}{\partial y_1}\right)^2 + \left(\frac{\partial V}{\partial z_1}\right)^2 = n_1^2. \tag{9}$$

Instead of the point characteristic it is often convenient to use certain related functions (also introduced by Hamilton), known as the *mixed characteristic* and the *angle characteristic*. They may be derived from the point characteristic by means of Legendre transformations,* and are particularly useful when either P_0 or P_1 or both these points are at infinity.

4.1.2 The mixed characteristic

The mixed characteristic function W is defined by the equation†

$$W = V - \Sigma p_1 x_1, \tag{10}$$

where Σ denotes summation over the three similar terms with suffix one. Eq. (10) expresses W as a function of nine variables, but in general only six (and in a homogeneous medium only five) are independent. To show this, consider the effect of small displacements of the points P_0 and P_1. The corresponding change in W is then given by

$$\delta W = \delta V - \Sigma p_1 \delta x_1 - \Sigma x_1 \delta p_1. \tag{11}$$

Now by (7),

$$\delta V = \Sigma p_1 \delta x_1 - \Sigma p_0 \delta x_0. \tag{12}$$

From (11) and (12),

$$\delta W = -\Sigma p_0 \delta x_0 - \Sigma x_1 \delta p_1. \tag{13}$$

* A Legendre transformation transforms in general a function $f(x, y)$ into a function $g(x, z)$, where $z = \partial f/\partial y$, in such a way that the derivative of g with respect to the new variable z is equal to the old variable y.

† To bring out more clearly its physical meaning, we follow Synge, *loc. cit.*, in defining the mixed characteristic with opposite sign to that used by Hamilton.

One can also define a mixed characteristic

$$W' = V + \Sigma p_0 x_0,$$

the summation being taken over similar terms with suffix zero. The properties of W and W' are, of course, strictly analogous.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

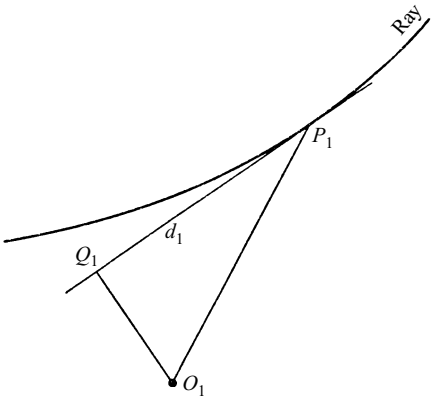

Fig. 4.2 Illustrating the meaning of the mixed characteristic function.

This relation shows that in general W can be expressed as a function of the six variables x_0, y_0, z_0, p_1, q_1 and m_1, and that, when it is so expressed,

$$p_0 = -\frac{\partial W}{\partial x_0}, \qquad x_1 = -\frac{\partial W}{\partial p_1}, \tag{14}$$

with similar expressions for q_0, y_1, m_0 and z_1. On account of (6), $W(x_0, y_0, z_0; p_1, q_1, m_1)$ satisfies the eikonal equation

$$\left(\frac{\partial W}{\partial x_0}\right)^2 + \left(\frac{\partial W}{\partial y_0}\right)^2 + \left(\frac{\partial W}{\partial z_0}\right)^2 = n_0^2. \tag{15}$$

It is to be observed (see Fig. 4.2) that the sum $\Sigma p_1 x_1$ has a simple geometrical interpretation:

$$\Sigma p_1 x_1 = n_1 d_1, \tag{16}$$

where $d_1 = Q_1 P_1$ is the projection of $O_1 P_1$ on to the tangent to the ray at P_1. If P_1 is situated in a homogeneous region, the portion of the ray near P_1 coincides with the line segment $Q_1 P_1$; according to (10) and (16) *W then represents the optical length of the ray from P_0 to the foot Q_1 of the perpendicular drawn from the origin O_1 on to the final portion of the ray* (Fig. 4.3):

$$W = [P_0 Q_1]. \tag{17}$$

Since in this case the refractive index of the medium around P_1 has a constant value it follows from (6) that

$$\delta m_1 = -\frac{p_1 \delta p_1 + q_1 \delta q_1}{m_1}, \tag{18}$$

and (13) becomes on substitution from (18)*

* If a function depends on variables which are connected by subsidiary relations, such as (6), some of the variables may be eliminated, or alternatively the relations may be used to express it as a *homogeneous* function in all the variables. The alternative procedure, which is somewhat more difficult to handle, was frequently employed by Hamilton.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

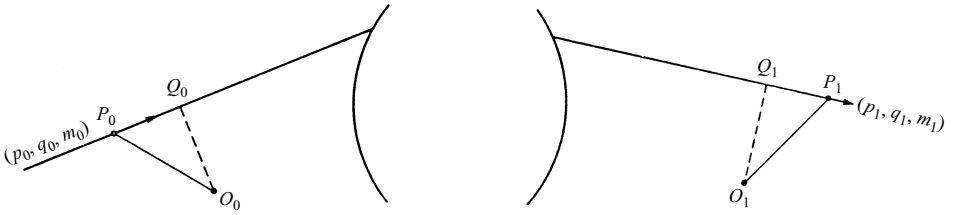

Fig. 4.3 Interpretation of Hamilton's characteristic functions when the initial and final media are homogeneous:

$$V(x_0, y_0, z_0; x_1, y_1, z_1) = [P_0P_1],$$
$$W(x_0, y_0, z_0; p_1, q_1) = [P_0Q_1],$$
$$T(p_0, q_0; p_1, q_1) = [Q_0Q_1].$$

$$\delta W = -\Sigma p_0 \delta x_0 - \left(x_1 - \frac{p_1}{m_1} z_1\right)\delta p_1 - \left(y_1 - \frac{q_1}{m_1} z_1\right)\delta q_1. \tag{19}$$

Hence *when the final medium is homogeneous, the mixed characteristic is expressible as a function of five variables:*

$$W = W(x_0, y_0, z_0; p_1, q_1),$$

and its derivatives then satisfy the relations

$$p_0 = -\frac{\partial W}{\partial x_0}, \qquad q_0 = -\frac{\partial W}{\partial y_0}, \qquad m_0 = -\frac{\partial W}{\partial z_0}, \tag{21}$$

$$x_1 - \frac{p_1}{m_1} z_1 = -\frac{\partial W}{\partial p_1}, \qquad y_1 - \frac{q_1}{m_1} z_1 = -\frac{\partial W}{\partial q_1}. \tag{22}$$

Eqs. (21) and (22) show that if a point on the ray in the initial medium and the components of the ray in the final medium are given, the ray components in the initial medium and points on the ray in the final medium may immediately be determined from the knowledge of the mixed characteristic.

4.1.3 The angle characteristic

The angle characteristic T may be defined by means of the relation

$$T = V + \Sigma p_0 x_0 - \Sigma p_1 x_1. \tag{23}$$

If P_0 and P_1 are slightly displaced, the corresponding change in T is given by

$$\delta T = \Sigma x_0 \delta p_0 - \Sigma x_1 \delta p_1, \tag{24}$$

where (12) was used. Hence T is expressible as a function of the six ray components, and when expressed in this way,

$$x_0 = \frac{\partial T}{\partial p_0}, \qquad x_1 = -\frac{\partial T}{\partial p_1}, \tag{25}$$

with similar relations for the other coordinates.

It is seen from (23) that *if the regions in which P_0 and P_1 are situated are both*

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

homogeneous, T represents the optical length of the ray between the feet Q_0 and Q_1 of perpendiculars drawn from O_0 and O_1 on to the initial and final portions of the ray (see Fig. 4.3),

$$T = [Q_0 Q_1]. \tag{26}$$

In this case the angle characteristic may be expressed as a function of four variables only. For if we substitute for δm_1 from (18) and for δm_0 from a similar relation, (24) becomes

$$\begin{aligned}\delta T = &\left(x_0 - \frac{p_0}{m_0} z_0\right)\delta p_0 + \left(y_0 - \frac{q_0}{m_0} z_0\right)\delta q_0 \\ &- \left(x_1 - \frac{p_1}{m_1} z_1\right)\delta p_1 - \left(y_1 - \frac{q_1}{m_1} z_1\right)\delta q_1.\end{aligned} \tag{27}$$

This relation shows that *when the initial and final media are homogeneous, the angle characteristic is expressible as a function of the four variables p_0, q_0, p_1 and q_1*:

$$T = T(p_0, q_0; p_1, q_1), \tag{28}$$

and its derivatives then satisfy the relations

$$\left.\begin{aligned} x_0 - \frac{p_0}{m_0} z_0 &= \frac{\partial T}{\partial p_0}, \qquad y_0 - \frac{q_0}{m_0} z_0 = \frac{\partial T}{\partial q_0}, \\ x_1 - \frac{p_1}{m_1} z_1 &= -\frac{\partial T}{\partial p_1}, \qquad y_1 - \frac{q_1}{m_1} z_1 = -\frac{\partial T}{\partial q_1}. \end{aligned}\right\} \tag{29}$$

If the components of the initial and the final portion of a ray are given, the coordinates of points on these portions may, according to (29), be determined immediately from the knowledge of the angle characteristic.

4.1.4 Approximate form of the angle characteristic of a refracting surface of revolution

Let

$$z = c_2(x^2 + y^2) + c_4(x^2 + y^2)^2 + \cdots, \tag{30}$$

where c_2, c_4, ... are constants, be the equation of refracting surface of revolution, referred to Cartesian axes, whose origin O coincides with the axial point (called pole) of the surface, and whose z direction is along the axis of symmetry. If r denotes the radius of curvature at the pole of the surface (measured as positive when the surface is convex towards light incident from the negative z direction), then

$$c_2 = \frac{1}{2r}. \tag{31}$$

For a spherical surface of radius r, $c_4 = 1/8r^3$. For a general surface of revolution we may write

$$c_4 = \frac{1}{8r^3}(1 + b); \tag{32}$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

the constant b (sometimes called the *deformation coefficient*) is a rough measure of the departure of the surface from spherical form. In terms of r and b,

$$z = \frac{x^2 + y^2}{2r} + \frac{(x^2 + y^2)^2}{8r^3}(1 + b) + \cdots . \tag{33}$$

It will be assumed that the regions on either side of the surface are homogeneous and of refractive indices n_0 and n_1 respectively. The angle characteristic will be referred to systems of axes parallel to those at O and with origins at axial points $O_0(0, 0, a_0)$ and $O_1(0, 0, a_1)$ ($a_0 < 0$, $a_1 > 0$, $r > 0$ in Fig. 4.4).

If P is the point of intersection of the incident ray with the refracting surface, and if Q_0 and Q_1 are the feet of the perpendiculars drawn from O_0 and O_1 onto the incident and the refracted ray, the angle characteristic T is then, according to (26),

$$\begin{aligned} T &= [Q_0P] + [PQ_1] \\ &= \{xp_0 + yq_0 + (z - a_0)m_0\} - \{xp_1 + yq_1 + (z - a_1)m_1\}, \end{aligned} \tag{34}$$

where (x, y, z) are the coordinates of P with respect to the axes at O, and (p_0, q_0, m_0), (p_1, q_1, m_1) are the components of the ray incident and refracted at P.

The coordinates (x, y, z) may be eliminated from (34) with the help of the law of refraction. According to §3.2.2, the law of refraction is equivalent to the assertion that the vector $\mathbf{N}(p_0 - p_1, q_0 - q_1, m_0 - m_1)$ is normal to the surface at P. Hence if (33) is written in the form

$$F(x, y, z) \equiv z - \frac{x^2 + y^2}{2r} - \frac{(x^2 + y^2)^2}{8r^3}(1 + b) - \cdots = 0, \tag{35}$$

then

$$\left.\begin{aligned} \lambda\frac{\partial F}{\partial x} &= -\lambda\left[\frac{x}{r} + \cdots\right] = p_0 - p_1, \\ \lambda\frac{\partial F}{\partial y} &= -\lambda\left[\frac{y}{r} + \cdots\right] = q_0 - q_1, \\ \lambda\frac{\partial F}{\partial z} &= \quad \lambda \qquad\qquad = m_0 - m_1. \end{aligned}\right\} \tag{36}$$

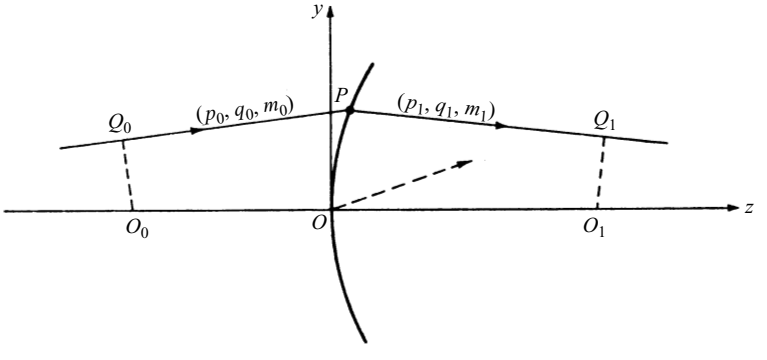

Fig. 4.4 The angle characteristic of a refracting surface of revolution. (The points O_0, O, O_1, Q_1, P, Q_0 are not necessarily coplanar.)

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

These equations imply that

$$\left.\begin{aligned} x &= -r\frac{p_0 - p_1}{m_0 - m_1} + \Delta x, \\ y &= -r\frac{q_0 - q_1}{m_0 - m_1} + \Delta y, \end{aligned}\right\} \tag{37}$$

where Δx and Δy are quantities of the third order in p, q, x/r, y/r. To express z in terms of the ray components, we substitute from (37) into (35), and obtain

$$\begin{aligned} z = {} & \frac{r}{2(m_1 - m_0)^2}[(p_0 - p_1)^2 + (q_0 - q_1)^2] + \frac{1}{m_1 - m_0}[\Delta x(p_0 - p_1) + \Delta y(q_0 - q_1)] \\ & + \frac{1}{8}\frac{r(1+b)}{(m_1 - m_0)^4}[(p_0 - p_1)^2 + (q_0 - q_1)^2]^2 + \cdots . \end{aligned} \tag{38}$$

To find the expansion of T up to and including the fourth-order terms it is not necessary to evaluate Δx and Δy; for when we substitute from (37) and (38) into (34), the contributions involving Δx and Δy are seen to be of order higher than the fourth and may therefore be neglected. Eq. (34) then becomes

$$\begin{aligned} T(p_0, q_0, m_0; p_1, q_1, m_1) = {} & -m_0 a_0 + m_1 a_1 + \frac{r}{2(m_1 - m_0)}[(p_0 - p_1)^2 + (q_0 - q_1)^2] \\ & - \frac{1}{8}\frac{r(1+b)}{(m_1 - m_0)^3}[(p_0 - p_1)^2 + (q_0 - q_1)^2]^2. \end{aligned} \tag{39}$$

Eq. (39) is the expansion of the angle characteristic up to the fourth order, the angle characteristic being considered as a function of all the six ray components. Two of the components may be eliminated by using the identity (6). We have from (6)

$$\left.\begin{aligned} m_0 &= n_0 - \frac{1}{2n_0}(p_0^2 + q_0^2) - \frac{1}{8n_0^3}(p_0^2 + q_0^2)^2 + \cdots \\ m_1 &= n_1 - \frac{1}{2n_1}(p_1^2 + q_1^2) - \frac{1}{8n_1^3}(p_1^2 + q_1^2)^2 + \cdots \end{aligned}\right\} \tag{40}$$

so that

$$\frac{1}{m_1 - m_0} = \frac{1}{n_1 - n_0}\left[1 - \frac{1}{2n_0(n_1 - n_0)}(p_0^2 + q_0^2) + \frac{1}{2n_1(n_1 - n_0)}(p_1^2 + q_1^2) + \cdots\right]. \tag{41}$$

Eq. (39) becomes, on substitution from (41):

$$\begin{aligned} T(p_0, q_0; p_1, q_1) = {} & n_1 a_1 - n_0 a_0 \\ & + \frac{r}{2(n_1 - n_0)}[(p_0 - p_1)^2 + (q_0 - q_1)^2] + \frac{a_0}{2n_0}(p_0^2 + q_0^2) - \frac{a_1}{2n_1}(p_1^2 + q_1^2) \\ & - \frac{r}{4(n_1 - n_0)^2}[(p_0 - p_1)^2 + (q_0 - q_1)^2]\left[\frac{p_0^2 + q_0^2}{n_0} - \frac{p_1^2 + q_1^2}{n_1}\right] \\ & - \frac{(1+b)r}{8(n_1 - n_0)^3}[(p_0 - p_1)^2 + (q_0 - q_1)^2]^2 \\ & + \frac{a_0}{8n_0^3}(p_0^2 + q_0^2)^2 - \frac{a_1}{8n_1^3}(p_1^2 + q_1^2)^2 + \cdots . \end{aligned} \tag{42}$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

The four variables p_0, q_0, p_1 and q_1 are seen to enter this expression only in the three combinations*

$$p_0^2 + q_0^2 = u^2, \quad p_1^2 + q_1^2 = v^2, \quad \text{and} \quad p_0 p_1 + q_0 q_1 = w^2. \tag{43}$$

With this substitution (42) becomes, on separating into orders,

$$T(p_0, q_0; p_1, q_1) = T^{(0)} + T^{(2)} + T^{(4)} + \cdots,$$

where

$$\left.\begin{aligned} T^{(0)} &= n_1 a_1 - n_0 a_0, \\ T^{(2)} &= \mathcal{A}u^2 + \mathcal{B}v^2 + \mathcal{C}w^2, \\ T^{(4)} &= \mathcal{D}u^4 + \mathcal{E}v^4 + \mathcal{F}w^4 + \mathcal{G}u^2v^2 + \mathcal{H}u^2w^2 + \mathcal{K}v^2w^2, \end{aligned}\right\} \tag{44}$$

and

$$\left.\begin{aligned} \mathcal{A} &= \frac{1}{2}\left[\frac{r}{n_1 - n_0} + \frac{a_0}{n_0}\right], \\ \mathcal{B} &= \frac{1}{2}\left[\frac{r}{n_1 - n_0} - \frac{a_1}{n_1}\right], \\ \mathcal{C} &= -\frac{r}{n_1 - n_0}, \\ \mathcal{D} &= -\frac{r}{4(n_1 - n_0)^2}\left[\frac{1+b}{2(n_1 - n_0)} + \frac{1}{n_0}\right] + \frac{a_0}{8n_0^3}, \\ \mathcal{E} &= -\frac{r}{4(n_1 - n_0)^2}\left[\frac{1+b}{2(n_1 - n_0)} - \frac{1}{n_1}\right] - \frac{a_1}{8n_1^3}, \\ \mathcal{F} &= \frac{-(1+b)r}{2(n_1 - n_0)^3}, \\ \mathcal{G} &= -\frac{r}{4(n_1 - n_0)^2}\left[\frac{1+b}{n_1 - n_0} + \frac{1}{n_0} - \frac{1}{n_1}\right], \\ \mathcal{H} &= \frac{r}{2(n_1 - n_0)^2}\left[\frac{1+b}{n_1 - n_0} + \frac{1}{n_0}\right], \\ \mathcal{K} &= \frac{r}{2(n_1 - n_0)^2}\left[\frac{1+b}{n_1 - n_0} - \frac{1}{n_1}\right]. \end{aligned}\right\} \tag{45}$$

* It can be shown more generally that the angle characteristic of any medium which is rotationally symmetrical about the z-axis depends on the four variables only through the three combinations (43). To see this, we use a result proved in §5.1, according to which any function $F(x_0, y_0; x_1, y_1)$ which is invariant with respect to rotation of axes about the origin in the x, y-plane depends only on the three scalar products

$$\mathbf{r}_0^2 = x_0^2 + y_0^2, \quad \mathbf{r}_1^2 = x_1^2 + y_1^2, \quad \mathbf{r}_0 \cdot \mathbf{r}_1 = x_0 x_1 + y_0 y_1,$$

of the two vectors $\mathbf{r}_0(x_0, y_0)$, $\mathbf{r}_1(x_1, y_1)$. Identifying $\mathbf{r}_0$ and $\mathbf{r}_1$ with the projections of the propagation vectors $\mathbf{g}_0(p_0, q_0, m_0)$ and $\mathbf{g}_1(p_1, q_1, m_1)$ on to the x, y-planes, the result follows.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

4.1.5 *Approximate form of the angle characteristic of a reflecting surface of revolution*

The expansion up to fourth degree for the angle characteristic associated with a reflecting surface of revolution can be derived in a similar manner. It is, however, not necessary to carry out the calculations in full. Using the same notation as in the preceding section (see Figs. 4.4 and 4.5) all the equations of §4.1.4 up to and including (39) apply without change in the present case; hence (39) is also the *angle characteristic of a reflecting surface of revolution, when regarded as a function of all the six ray components*. However, when m_0 and m_1 are eliminated from (39) with the help of the two identities connecting the ray components, different expressions for T (as a function of four ray components) are obtained in the two cases. Denoting by n the refractive index of the medium in which the rays are situated, we have in place of (40),

$$\left.\begin{aligned} m_0 &= n - \frac{1}{2n}(p_0^2 + q_0^2) - \frac{1}{8n^3}(p_0^2 + q_0^2)^2 + \cdots, \\ m_1 &= -\left[n - \frac{1}{2n}(p_1^2 + q_1^2) - \frac{1}{8n^3}(p_1^2 + q_1^2)^2 + \cdots\right]. \end{aligned}\right\} \tag{46}$$

In the second relation the negative square root $-\sqrt{n^2 - (p^2 + q^2)}$ has been taken, as we assume that the reflected ray returns into the region from which the light is propagated ($z < 0$); the direction cosine of the reflected ray with respect to the positive z direction, and consequently m, is therefore negative. Since (40) reduces to (46) on setting $n_0 = -n_1 = n$, it follows that *the angle characteristic, considered as a function of the four ray components* p_0, q_0, p_1 *and* q_1, *of a reflecting surface of revolution, can be obtained from the angle characteristic of a refracting surface of revolution by setting* $n_0 = -n_1 = n$. Hence, for the case of reflection, we have

$$\left.\begin{aligned} T^{(0)} &= -n(a_0 + a_1), \\ T^{(2)} &= \mathcal{A}'u^2 + \mathcal{B}'v^2 + \mathcal{C}'w^2, \\ T^{(4)} &= \mathcal{D}'u^4 + \mathcal{E}'v^4 + \mathcal{F}'w^4 + \mathcal{G}'u^2v^2 + \mathcal{H}'u^2w^2 + \mathcal{K}'v^2w^2, \end{aligned}\right\} \tag{47}$$

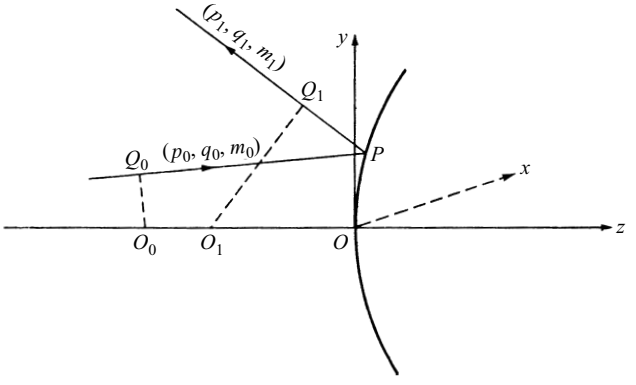

Fig. 4.5 The angle characteristic of a reflecting surface of revolution. (The points O_0, O_1, O, Q_0, P, Q_1 are not necessarily coplanar.)

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

where

$$\left.\begin{aligned}
\mathcal{A}' &= \frac{1}{2n}[-\tfrac{1}{2}r + a_0],\\
\mathcal{B}' &= \frac{1}{2n}[-\tfrac{1}{2}r + a_1],\\
\mathcal{C}' &= \frac{r}{2n},\\
\mathcal{D}' &= -\frac{r}{16n^3}\left[-\frac{1+b}{4}+1\right] + \frac{a_0}{8n^3},\\
\mathcal{E}' &= -\frac{r}{16n^3}\left[-\frac{1+b}{4}+1\right] + \frac{a_1}{8n^3},\\
\mathcal{F}' &= \frac{(1+b)r}{16n^3},\\
\mathcal{G}' &= \frac{r}{16n^3}\left[\frac{1+b}{2}-2\right],\\
\mathcal{H}' &= \frac{r}{8n^3}\left[-\frac{1+b}{2}+1\right],\\
\mathcal{K}' &= \frac{r}{8n^3}\left[-\frac{1+b}{2}+1\right].
\end{aligned}\right\} \tag{48}$$

4.2 Perfect imaging

Consider the propagation of light from a point source situated at a point P_0 in a medium specified by a refractive index function $n(x, y, z)$. An infinite number of rays will then proceed from P_0, but in general only a finite number will pass through any other point of the medium. In special cases, however, a point P_1 may be found through which an infinity of rays pass. Such a point P_1 is said to be a *stigmatic* (or a *sharp*) image of P_0.

In an ideal optical instrument every point P_0 of a three-dimensional region, called the *object space*, will give rise to a stigmatic image P_1. The totality of the image points defines the *image space*. The corresponding points in the two spaces are said to be *conjugate points*. In general not all the rays which proceed from P_0 will reach the image space; some, for example, will be excluded by the diaphragms of the instrument. Those rays which reach the image space will be said *to lie in the field of the instrument*. When P_0 describes a curve C_0 in the object space, P_1 will describe a conjugate curve C_1. The two curves will not necessarily be geometrically similar to each other. If *every* curve C_0 of the object space is geometrically similar to its image, we may say that the imaging between the two spaces is *perfect*. In a similar way we may define perfect imaging between two surfaces.

Optical instruments which are perfect in the sense just defined are of considerable interest, and accordingly we shall formulate some general theorems relating to perfect, or at least sharp, imaging of three-dimensional domains. Some results

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

relating to sharp imaging of two-dimensional domains (surfaces) will be briefly discussed in §4.2.3.

4.2.1 General theorems

An optical system $\mathcal{J}$ which images stigmatically a three-dimensional domain is often called an *absolute instrument*. It will be shown that *in an absolute instrument the optical length of any curve in the object space is equal to the optical length of its image*. This theorem was first put forward by Maxwell* in 1858 for the special case when both the object and the image space are homogeneous. More rigorous proofs were later given by H. Bruns (1895), F. Klein (1901) and H. Liebmann (1916).†

The theorem was later shown by Carathéodory‡ not to be restricted to homogeneous media, but to be valid also when the media are heterogeneous and anisotropic. In proving this theorem the method of Carathéodory will be used, but our discussion will be restricted to absolute instruments with isotropic (but generally heterogeneous) object and image spaces.§

Let A_0B_0 and A_1B_1 be ray-segments in the object and image spaces (Fig. 4.6) of a ray which lies in the field of an absolute instrument $\mathcal{J}$. Any other ray with a line-element which neither in position nor in direction departs appreciably from an element of $A_0B_0A_1B_1$ will also lie in the field of the instrument.

If each element of a curve (which is assumed to have a continuously turning tangent) coincides with an element of some ray which lies completely in the field of

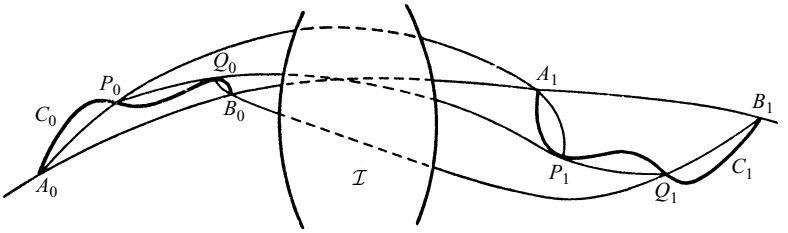

Fig. 4.6 An absolute optical instrument.

* J. C. Maxwell, *Quart. J. Pure Appl. Maths.*, **2** (1858), p. 233. Also his *Scientific Papers*, Vol. 1 (Cambridge, Cambridge University Press, 1890), p. 271.

† H. Bruns, "Das Eikonal", *Abh. Kgl. sächs. Ges. Wiss., math-phys. Kl.*, **21** (1895), 370; F. Klein, *Z. Math. Phys.*, **46** (1901), 376; *Ges. Math. Abh.*, **2** (1922), 607. (See also E. T. Whittaker, *The Theory of Optical Instruments* (Cambridge, Cambridge University Press, 1907), p. 47. H. Liebmann, *Sitzgsber. bayer. Akad. Wiss., Math-naturw. Abt.* (1916), 183. An account of these researches will also be found in an article by H. Boegehold in S. Czapski and O. Eppenstein, *Grundzüge der Optischen Instrumente nach Abbe* (Leipzig, Barth, 3rd edition, 1924), p. 213.)

‡ C. Carathéodory, *Sitzgsber. bayer. Akad. Wiss. Math-naturw. Abt.*, **56** (1926), 1.

§ In microwave optics the use of heterogeneous substances is common [see, for example, J. Brown, *Microwave Lenses* (London, Methuen, 1953)]. In light optics interest in them has developed more recently (see E. W. Marchand, *Progress in Optics*, Vol. 11, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, American Elsevier Publishing Company, 1973), p. 305 and E. W. Marchand, *Gradient Index Optics* (New York, Academic Press, 1978).)

$\mathcal{J}$, the curve will be said to lie *tangentially* in the field of $\mathcal{J}$. If a 'polygon' with a sufficient number of sides is inscribed into such a curve, each side of the polygon will coincide with an element of a ray which lies completely in the field of the instrument.

According to the principle of equal optical path (see §3.3.3), all the rays which join A_0 to its image A_1 have the same optical length. We shall denote this optical path length by $V(A_0)$ and will show that it is in fact independent of A_0.

Let B_0 and B_1 be another pair of conjugate points. Then (see Fig. 4.6)

$$[A_1B_1] = [A_0B_0] + V(B_0) - V(A_0). \tag{1}$$

Let C_0 be a curve which joins A_0 and B_0 and which lies tangentially in the field of the instrument, and let C_1 be its image. We inscribe into C_0 a polygon $A_0P_0Q_0B_0$ and denote the image points of P_0 and Q_0 by P_1 and Q_1. Then, on applying (1) to the side A_1P_1, we have

$$[A_1P_1] = [A_0P_0] + V(P_0) - V(A_0),$$

and similarly for the other sides:

$$[P_1Q_1] = [P_0Q_0] + V(Q_0) - V(P_0),$$

and

$$[Q_1B_1] = [Q_0B_0] + V(B_0) - V(Q_0).$$

Hence

$$[A_1P_1] + [P_1Q_1] + [Q_1B_1] = [A_0P_0] + [P_0Q_0] + [Q_0B_0] + V(B_0) - V(A_0).$$

Obviously this result can be extended to a polygon of any number of sides N. Proceeding to the limit, as $N \to \infty$ in such a way that the greatest side tends to zero, we obtain the relation

$$L_1 = L_0 + V(B_0) - V(A_0), \tag{2}$$

where

$$L_0 = \int_{C_0} n_0 \, \mathrm{d}s_0, \qquad L_1 = \int_{C_1} n_1 \, \mathrm{d}s_1 \tag{3}$$

are the optical lengths of the curves C_0 and C_1. Next, it will be shown that $V(B_0) = V(A_0)$.

The points on the two curves are in a one–one correspondence, which may be expressed by relations of the form

$$x_1 = f(x_0, y_0, z_0), \qquad y_1 = g(x_0, y_0, z_0), \qquad z_1 = h(x_0, y_0, z_0). \tag{4}$$

An element $\mathrm{d}s_1$ of C_1 is a function of the corresponding element $\mathrm{d}s_0$,

$$\mathrm{d}s_1 = \sqrt{\left(\frac{\mathrm{d}x_1}{\mathrm{d}s_0}\right)^2 + \left(\frac{\mathrm{d}y_1}{\mathrm{d}s_0}\right)^2 + \left(\frac{\mathrm{d}z_1}{\mathrm{d}s_0}\right)^2}\,\mathrm{d}s_0. \tag{5}$$

Hence

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$L_1 = \int_{C_1} F\left(x_1,\, y_1,\, z_1,\, \frac{\mathrm{d}x_1}{\mathrm{d}s_0},\, \frac{\mathrm{d}y_1}{\mathrm{d}s_0},\, \frac{\mathrm{d}z_1}{\mathrm{d}s_0}\right) \mathrm{d}s_0, \tag{6}$$

where

$$F\left(x_1,\, y_1,\, z_1,\, \frac{\mathrm{d}x_1}{\mathrm{d}s_0},\, \frac{\mathrm{d}y_1}{\mathrm{d}s_0},\, \frac{\mathrm{d}z_1}{\mathrm{d}s_0}\right) = n_1(x_1,\, y_1,\, z_1)\sqrt{\left(\frac{\mathrm{d}x_1}{\mathrm{d}s_0}\right)^2 + \left(\frac{\mathrm{d}y_1}{\mathrm{d}s_0}\right)^2 + \left(\frac{\mathrm{d}z_1}{\mathrm{d}s_0}\right)^2}$$

is a homogeneous function of the first degree in the derivatives $\mathrm{d}x_1/\mathrm{d}s_0$, $\mathrm{d}y_1/\mathrm{d}s_0$ and $\mathrm{d}z_1/\mathrm{d}s_0$; moreover, F remains unchanged when $\mathrm{d}x_1/\mathrm{d}s_0, \ldots,$ is replaced by $-\mathrm{d}x_1/\mathrm{d}s_0, \ldots$. Now from (4),

$$\frac{\mathrm{d}x_1}{\mathrm{d}s_0} = \frac{\partial f}{\partial x_0}\frac{\mathrm{d}x_0}{\mathrm{d}s_0} + \frac{\partial f}{\partial y_0}\frac{\mathrm{d}y_0}{\mathrm{d}s_0} + \frac{\partial f}{\partial z_0}\frac{\mathrm{d}z_0}{\mathrm{d}s_0}, \tag{7}$$

with similar expressions for $\mathrm{d}y_1/\mathrm{d}s_0$ and $\mathrm{d}z_1/\mathrm{d}s_0$. Hence using (7) and (4), F can be expressed in the form

$$F\left(x_1,\, y_1,\, z_1,\, \frac{\mathrm{d}x_1}{\mathrm{d}s_0},\, \frac{\mathrm{d}y_1}{\mathrm{d}s_0},\, \frac{\mathrm{d}z_1}{\mathrm{d}s_0}\right) = \Phi\left(x_0,\, y_0,\, z_0,\, \frac{\mathrm{d}x_0}{\mathrm{d}s_0},\, \frac{\mathrm{d}y_0}{\mathrm{d}s_0},\, \frac{\mathrm{d}z_0}{\mathrm{d}s_0}\right), \tag{8}$$

Φ being also a homogeneous function of the first degree in $\mathrm{d}x_0/\mathrm{d}s_0, \ldots;$ moreover, Φ remains unchanged when $\mathrm{d}x_0/\mathrm{d}s_0, \ldots,$ is replaced by $-\mathrm{d}x_0/\mathrm{d}s_0, \ldots,$

$$\Phi\left(x_0,\, y_0,\, z_0,\, -\frac{\mathrm{d}x_0}{\mathrm{d}s_0},\, -\frac{\mathrm{d}y_0}{\mathrm{d}s_0},\, -\frac{\mathrm{d}z_0}{\mathrm{d}s_0}\right) = \Phi\left(x_0,\, y_0,\, z_0,\, \frac{\mathrm{d}x_0}{\mathrm{d}s_0},\, \frac{\mathrm{d}y_0}{\mathrm{d}s_0},\, \frac{\mathrm{d}z_0}{\mathrm{d}s_0}\right). \tag{9}$$

From (2), (6) and (8) it follows that

$$\int_{C_0} (n_0 - \Phi)\, \mathrm{d}s_0 = V(A_0) - V(B_0), \tag{10}$$

showing that the value of the curvilinear integral in (10) depends only on the end points A_0, B_0 and not on the choice of C_0. The curve C_0 is, however, not quite arbitrary, for it must lie tangentially in the field of the instrument. Nevertheless, it may be concluded that the expression $(n_0 - \Phi)\,\mathrm{d}s_0$ must be a complete differential of some function Ψ,

$$n_0 - \Phi = \frac{\partial \Psi}{\partial x_0}\frac{\mathrm{d}x_0}{\mathrm{d}s_0} + \frac{\partial \Psi}{\partial y_0}\frac{\mathrm{d}y_0}{\mathrm{d}s_0} + \frac{\partial \Psi}{\partial z_0}\frac{\mathrm{d}z_0}{\mathrm{d}s_0}.$$

If now the derivatives $\mathrm{d}x_0/\mathrm{d}s_0, \ldots,$ are replaced by $-\mathrm{d}x_0/\mathrm{d}s_0$, the right-hand side will change sign, but the left-hand side will, on account of (9), remain unchanged. This is only possible if each side vanishes (Ψ = constant); hence

$$\Phi = n_0. \tag{11}$$

Eq. (10) shows that $V(A_0) = V(B_0)$ and consequently (2) reduces to the relation $L_1 = L_0$. Hence for any curve, whether or not it lies tangentially in the field, provided only it has an image,

$$\int_{C_0} n_0\, \mathrm{d}s_0 = \int_{C_1} n_1\, \mathrm{d}s_1. \tag{12}$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

This is *Maxwell's theorem for an absolute instrument*.*

From Maxwell's theorem a number of interesting conclusions can immediately be drawn. Consider a small triangle whose sides are of lengths $\mathrm{d}s_0^{(1)}$, $\mathrm{d}s_0^{(2)}$, $\mathrm{d}s_0^{(3)}$ and let $\mathrm{d}s_1^{(1)}$, $\mathrm{d}s_1^{(2)}$, $\mathrm{d}s_1^{(3)}$ be the sides of its image formed by an absolute instrument. Further let n_0 and n_1 be the refractive indices of the regions where the triangles are situated. By Maxwell's theorem,

$$n_0\,\mathrm{d}s_0^{(1)} = n_1\,\mathrm{d}s_1^{(1)}, \qquad n_0\,\mathrm{d}s_0^{(2)} = n_1\,\mathrm{d}s_1^{(2)}, \qquad n_0\,\mathrm{d}s_0^{(3)} = n_1\,\mathrm{d}s_1^{(3)}. \tag{13}$$

Hence the two triangles are similar to each other and corresponding sides are in the inverse ratio of the refractive indices. The angle between any two curves will therefore be preserved in the imaging, i.e. the imaging must be a *conformal transformation*. Now there is a general theorem due to Liouville† according to which a conformal transformation of a three-dimensional domain on to a three-dimensional domain can only be a projective transformation (collineation), an inversion‡ or the combination of

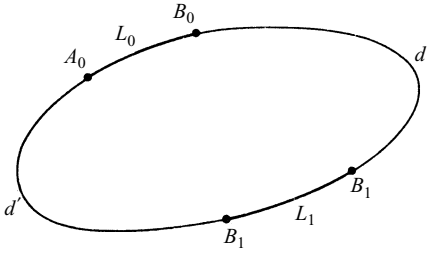

Fig. 4.7 Illustrating Lenz' proof of Maxwell's theorem for an absolute instrument.

* The following less general, but very simple and elegant proof of Maxwell's theorem was given by W. Lenz (contribution in *Probleme der Modernen Physik*, ed. P. Debye (Leipzig, Hirzel, 1928), p. 198.

Assume that all the rays from each object point reach the image and let (A_0, A_1) and (B_0, B_1) be two conjugate pairs. By hypothesis, the ray A_0B_0 must pass through A_1 and B_1. Likewise B_0A_0 must pass through these points. Hence each ray must be a closed curve and by the principle of equal optical path it follows that (see Fig. 4.7)

$$[A_0A_1]_{\text{clockwise}} = [A_0A_1]_{\text{anticlockwise}}$$

and

$$[B_0B_1]_{\text{clockwise}} = [B_0B_1]_{\text{anticlockwise}}.$$

Let

$$[A_0B_0] = L_0, \quad [A_1B_1] = L_1, \quad [B_0A_1] = d, \quad [B_1A_0] = d'.$$

The two equations then become

$$L_0 + d = d' + L_1,$$
$$d + L_1 = L_0 + d'$$

and on subtraction we obtain

$$L_0 = L_1.$$

This proves Maxwell's theorem for the special case when the curve is a portion of a ray. Generalization to an arbitrary curve may be obtained as in our main proof, by regarding the curve as a limiting form of a polygon formed by a large number of ray segments.

† See, for example, W. Blaschke, *Vorlesungen über Differential-Geometrie* I (Berlin, Springer, 2nd edition 1924), p. 68; (4th edition 1945), p. 101.

‡ An inversion transforms each point P_0 into a point P_1 on the line joining P_0 with a fixed origin O, and the product $OP_0 \cdot OP_1$ is constant.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

these two transformations. We have thus established the following theorem due to Carathéodory: *The imaging by an absolute instrument is a projective transformation, an inversion, or a combination of the two*.

Let us now briefly consider the case when the imaging between the two spaces is not only stigmatic but is perfect, i.e. it is such that any figure is transformed into one which is geometrically similar to it. Clearly the imaging must be a projective transformation, since it transforms lines into lines.* It then follows from (13) that *the magnification* $\mathrm{d}s_1/\mathrm{d}s_0$ *between any two conjugate linear elements is equal to the ratio* n_0/n_1 *of the refractive indices*. In particular, if $n_0 = n_1 = \text{constant}$, then $\mathrm{d}s_1/\mathrm{d}s_0 = 1$, so that a *perfect imaging between two homogeneous spaces of equal refractive indices is always trivial in the sense that it produces an image which is congruent with the object*. A plane mirror (or a combination of plane mirrors) is the only known instrument which produces such imaging.

These general considerations imply that, in order to obtain nontrivial imaging between homogeneous spaces of equal refractive indices, the requirement of exact stigmatism or of strict similarity between the object and the image must be dropped.

4.2.2 Maxwell's 'fish-eye'

A simple and interesting example of an absolute instrument is presented by the medium which is characterized by the refractive index function

$$n(r) = \frac{1}{1 + (r/a)^2}\, n_0, \tag{14}$$

where r denotes the distance from a fixed point O, and n_0 and a are constants. It is known as the 'fish-eye' and was first investigated by Maxwell.†

It was shown in §3.2 that in a medium with spherical symmetry the rays are plane curves which lie in planes through the origin, and that the equation of the rays may be written in the form [see §3.2 (11)]

$$\theta = c\int^{r} \frac{\mathrm{d}r}{r\sqrt{n^2(r)r^2 - c^2}},$$

c being a constant. On substituting from (14) and setting

$$\rho = \frac{r}{a}, \qquad K = \frac{c}{an_0}, \tag{15}$$

* Cf. F. Klein, *Elementary Mathematics from an Advanced Standpoint*, Vol. II (translated from third German edition, London, Macmillan, 1939, p. 89; reprinted by Dover Publications, New York).

† J. C. Maxwell, *Cambridge and Dublin Math. J.*, **8** (1854), 188; also *Scientific Papers*, Vol I (Cambridge, Cambridge University Press), p. 76.

Interesting generalizations of Maxwell's fish-eye were found by W. Lenz, contribution in *Probleme der Modernen Physik*, ed. P. Debye (Leipzig, Hirzel, 1928), p. 198 and R. Stettler, *Optik*, **12** (1955), 529. The latter paper also includes a generalization of the so-called *Luneburg lens* which, because of its wide angle scanning capabilities, has useful applications in microwave antenna design. This lens, first considered by R. K. Luneburg in his *Mathematical Theory of Optics* (University of California Press, Berkeley and Los Angeles, 1964) §29, is an inhomogeneous sphere with the refractive index function $n(r) = \sqrt{2 - r^2}$ $(0 \leqslant r \leqslant 1)$, which brings to a sharp focus every incident pencil of parallel rays. See also R. F. Rinehart, *J. Appl. Phys.*, **19** (1948), 860; A. Fletcher, T. Murphy and A. Young, *Proc. Roy. Soc.*, A **223** (1954), 216; and G. Toraldo di Francia, *Optica Acta*, **1** (1954–1955), 157.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

we obtain

$$\theta = \int^{\rho} \frac{K(1+\rho^2)\,\mathrm{d}\rho}{\rho\sqrt{\rho^2 - K^2(1+\rho^2)^2}}. \tag{16}$$

It may be verified that

$$\frac{K(1+\rho^2)}{\rho\sqrt{\rho^2 - K^2(1+\rho^2)^2}} = \frac{\mathrm{d}}{\mathrm{d}\rho}\left[\arcsin\left(\frac{K}{\sqrt{1-4K^2}}\frac{\rho^2-1}{\rho}\right)\right],$$

so that (16) becomes

$$\sin(\theta - \alpha) = \frac{c}{\sqrt{a^2 n_0{}^2 - 4c^2}}\frac{r^2 - a^2}{ar}, \tag{17}$$

where α is a constant of integration.

Eq. (17) is the polar equation of the rays. The one-parameter family of rays through a fixed point $P_0(r_0, \theta_0)$ is therefore given by

$$\frac{r^2 - a^2}{r\sin(\theta - \alpha)} = \frac{r_0{}^2 - a^2}{r_0\sin(\theta_0 - \alpha)}. \tag{18}$$

It is seen that whatever the value of α, this equation is satisfied by $r = r_1$, $\theta = \theta_1$, where

$$r_1 = \frac{a^2}{r_0}, \qquad \theta_1 = \pi + \theta_0, \tag{19}$$

showing that *all the rays from an arbitrary point P_0 meet in a point P_1 on the line joining P_0 to O; P_0 and P_1 are on opposite sides of O and $OP_0 \cdot OP_1 = a^2$*. Hence the fish-eye is an absolute instrument in which the imaging is an *inversion*.

We note that (17) is satisfied by $r = a$, $\theta = \alpha$ and $r = a$, $\theta = \pi + \alpha$; each ray therefore intersects the fixed circle $r = a$ in diametrically opposite points (see Fig. 4.8).

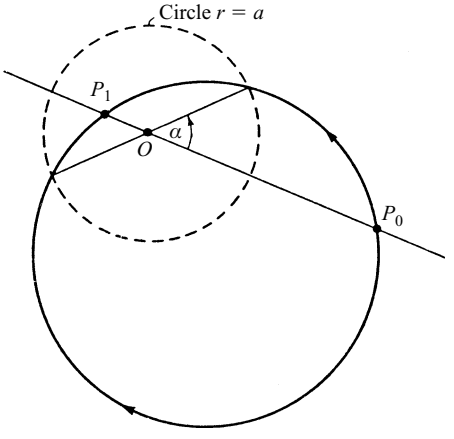

Fig. 4.8 Rays in Maxwell's 'fish-eye'.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

To obtain the equation of the rays in Cartesian coordinates, we put $x = r\cos\theta$, $y = r\sin\theta$ in (17), and find

$$y\cos\alpha - x\sin\alpha = \frac{c}{a\sqrt{a^2 n_0{}^2 - 4c^2}}(x^2 + y^2 - a^2),$$

or

$$(x + b\sin\alpha)^2 + (y - b\cos\alpha)^2 = a^2 + b^2 \tag{20}$$

where

$$b = \frac{a}{2c}\sqrt{a^2 n_0{}^2 - 4c^2}.$$

Eq. (20) shows that each ray is a circle.

4.2.3 Stigmatic imaging of surfaces

So far we have been concerned only with perfect or sharp imaging of three-dimensional domains. We saw that when the object space and image space are homogeneous and of equal refractive indices, perfect imaging can only be of a trivial kind, producing a mirror image of the object. It is natural to inquire whether nontrivial imaging may be obtained, when it is required that only certain *surfaces* should be imaged perfectly (or at least sharply) by the instrument. This question has been investigated by a number of authors,* who found that *when the object space and image space are homogeneous, not more than two surfaces may in general*† *be sharply imaged by a rotationally symmetrical system*. For the proof of this theorem the papers by Boegehold and Herzberger and by Smith should be consulted. Here we shall only consider in detail a simple case of sharp imaging of a spherical surface which is of particular interest in practice.

Consider the refraction at a solid homogeneous sphere S embedded in a homogeneous medium. Let O be the centre of the sphere, r its radius, and n and n' the refractive indices of the sphere and of the surrounding medium respectively. Further, let AQ be a ray incident upon the sphere. The refracted ray QB can easily be found by means of the following construction:

Let S_0 and S_1 be two spheres whose centres are at O and whose radii are

$$r_0 = \frac{n}{n'}r, \qquad r_1 = \frac{n'}{n}r. \tag{21}$$

If P_0 is the point of intersection of AQ with S_0, and P_1 is the point at which OP_0 meets S_1, then QP_1 is the refracted ray. For one has, by construction (see Fig. 4.9)

$$\frac{OQ}{OP_0} = \frac{OP_1}{OQ} = \frac{n'}{n}. \tag{22}$$

* H. Boegehold and M. Herzberger, *Compositio Mathematica*, **1** (1935), 448; M. Herzberger, *Ann. New York Acad. Sci.*, **48** (1946), Art. 1, 1; T. Smith, *Proc. Phys. Soc.*, **60** (1948), 293. See also C. G. Wynne, *Proc. Phys. Soc.*, **65B** (1952), 436.

† 'In general' implies here that certain degenerate cases in which the *whole* object space is imaged sharply (e.g. by reflection on a plane mirror) are excluded.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

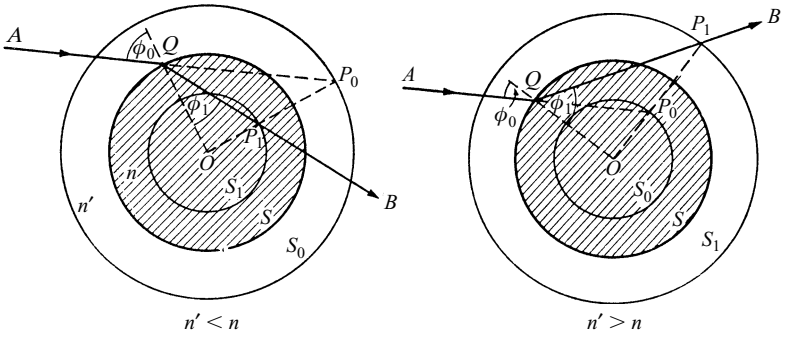

Fig. 4.9 Refraction at a spherical surface. Aplanatic points.

Moreover,

$$\widehat{QOP_0} = \widehat{QOP_1}. \tag{23}$$

Hence the triangles OQP_0 and OP_1Q are similar, and consequently

$$\frac{\sin\phi_0}{\sin\phi_1} = \frac{OP_0}{OQ} = \frac{n}{n'}, \tag{24}$$

where $\phi_0 = \angle OQP_0$ and $\phi_1 = \angle OQP_1$ are the angles of incidence and refraction respectively. ϕ_0 and ϕ_1 satisfy the law of refraction, and consequently QP_1 is the refracted ray.

The construction implies that all rays which diverge from a point P_0 on S_0 will form a (virtual) stigmatic image at the point P_1 in which the diameter OP_0 intersects S_1. Hence *the sphere S_1 is a stigmatic image of S_0, and vice versa*.

It will be useful to express (24) in a somewhat different form. If we denote by θ_0 and θ_1 the angles which the two conjugate rays make with the line P_0P_1, i.e. $\theta_0 = \angle OP_0Q$, $\theta_1 = \angle OP_1Q$, then, since the two triangles are similar, it follows that $\theta_0 = \phi_1$ and $\theta_1 = \phi_0$; hence

$$\frac{\sin\theta_1}{\sin\theta_0} = \frac{n}{n'} = \text{constant}. \tag{25}$$

Eq. (25) is a special case of the so-called *sine condition* whose significance will be explained in §4.5. In accordance with the terminology of §4.5, P_0 and P_1 are called *aplanatic points* of the spherical surface S.

The existence of the aplanatic points for refraction at a spherical surface is made use of, as will be shown in §6.6, in the construction of certain microscope objectives.

4.3 Projective transformation (collineation) with axial symmetry

It has been shown in the previous section that perfect imaging between three-dimensional domains must necessarily be a projective transformation, since it transforms lines into lines. But even when the requirements for perfect imaging are not strictly fulfilled, the properties of projective transformations are of great importance. For, as will be seen later, the relationship between the object and the image in any

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

optical system is, *at least to a first approximation*, a transformation of this kind. It will therefore be convenient to study the general properties of a projective transformation before deriving the laws of image formation in actual instruments. Though this preliminary discussion is essentially of a geometrical nature, it will be convenient to retain, where possible, the terminology of optics.

4.3.1 General formulae

Let (x, y, z) be the coordinates of a point P of the object space and (x', y', z') the coordinates of a point P' in the image space, both referred at present to the same set of Cartesian rectangular axes, chosen arbitrarily. A projective relationship between the two spaces is mathematically expressed by relations of the form

$$x' = \frac{F_1}{F_0}, \qquad y' = \frac{F_2}{F_0}, \qquad z' = \frac{F_3}{F_0}, \tag{1}$$

where

$$F_i = a_i x + b_i y + c_i z + d_i \qquad (i = 0, 1, 2, 3). \tag{2}$$

Pairs of points related by (1) will be said to form a *conjugate* pair.

Solving (1) for x, y, z, we obtain relations of the same form

$$x = \frac{F_1'}{F_0'}, \qquad y = \frac{F_2'}{F_0'}, \qquad z = \frac{F_3'}{F_0'}, \tag{3}$$

where

$$F_i' = a_i' x' + b_i' y' + c_i' z' + d_i'.$$

From (1) it follows that the image of any point situated in the plane $F_0 = 0$ will lie at infinity. Similarly from (3) it is seen that all object points whose images lie in the plane $F_0' = 0$ are at infinity. The plane $F_0 = 0$ is called the *focal plane* of the *object space* and the plane $F_0' = 0$ is known as the *focal plane* of the *image space*.* Rays which are parallel in the object space will be transformed into rays which intersect in a point on the focal plane $F_0' = 0$. Similarly rays from a point in the focal plane $F_0 = 0$ will transform into a pencil of parallel rays. In special cases both the focal planes may lie at infinity. The transformation is then said to be *affine* or *telescopic*. Then to finite values (x, y, z) there correspond finite values of (x', y', z'), so that in a telescopic transformation one always has $F_0 \neq 0$ and $F_0' \neq 0$. This clearly is only possible when $a_0 = b_0 = c_0 = 0$ and $a_0' = b_0' = c_0' = 0$.

Of special importance for optics is the case of *axial symmetry*, since the majority of optical systems consist of surfaces of revolution with a common axis (called usually *centred systems*). It then follows from symmetry that the image of each point P_0 lies in the plane which contains P_0 and the axis; in considering the properties of the associated projective transformation we may therefore restrict our discussion to points situated in such a *meridional plane*. Let the meridional plane be the y, z-plane and

* The terms 'focal plane' and 'focal points' have here a somewhat different meaning than in connection with normal congruences (§3.2.3) and astigmatic pencils (§4.6).

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

take the z-axis along the axis of symmetry. A point $(0, y, z)$ of the object space will then be transformed into a point $(0, y', z')$ of the image space, where

$$y' = \frac{b_2 y + c_2 z + d_2}{b_0 y + c_0 z + d_0}, \qquad z' = \frac{b_3 y + c_3 z + d_3}{b_0 y + c_0 z + d_0}. \tag{4}$$

Now it follows from symmetry that z' remains unchanged when y is changed into $-y$. This, in general, is only possible if $b_0 = b_3 = 0$. Further it follows that, if $y \to -y$, then $y' \to -y'$, which implies that $c_2 = d_2 = 0$. Hence (4) reduces to

$$y' = \frac{b_2 y}{c_0 z + d_0}, \qquad z' = \frac{c_3 z + d_3}{c_0 z + d_0}. \tag{5}$$

These equations contain five constants but only their ratios are significant. Hence *a projective transformation with axial symmetry is characterized by four parameters*.

Solving (5) for y and z, we obtain

$$y = \frac{c_0 d_3 - c_3 d_0}{b_2} \frac{y'}{c_0 z' - c_3}, \qquad z = \frac{-d_0 z' + d_3}{c_0 z' - c_3}. \tag{6}$$

From (5) and (6) it is seen that the focal planes are given by

$$F_0 \equiv c_0 z + d_0 = 0, \qquad F_0' \equiv c_0 z' - c_3 = 0;$$

hence the focal planes intersect the axis at right angles in the points whose abscissae are

$$z = -\frac{d_0}{c_0}, \qquad z' = \frac{c_3}{c_0}. \tag{7}$$

These points are called the *principal foci* and are denoted by F and F' in Fig. 4.10.

It will now be convenient to introduce a separate coordinate system for each of the two spaces, measuring the z coordinates from the principal foci; i.e. we set

$$\left.\begin{aligned} y = Y, \quad & c_0 z + d_0 = c_0 Z, \\ y' = Y', \quad & c_0 z' - c_3 = c_0 Z'. \end{aligned}\right\} \tag{8}$$

Eqs. (5) then become

$$Y' = \frac{b_2}{c_0} \frac{Y}{Z}, \qquad Z' = \frac{c_0 d_3 - c_3 d_0}{c_0^2 Z}.$$

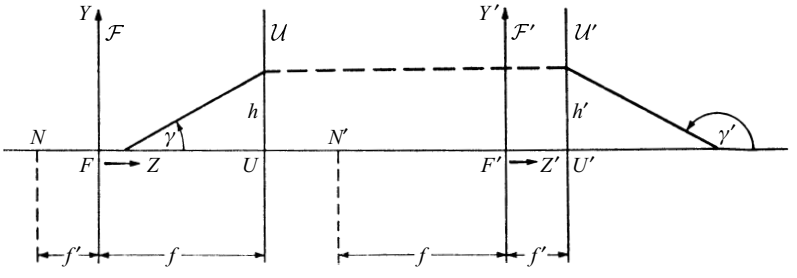

Fig. 4.10 The cardinal points and planes of an optical system: F, F' = foci; U, U' = unit (or principal) points; N, N' = nodal points; $\mathcal{F}$, $\mathcal{F}'$ = focal planes; $\mathcal{U}$, $\mathcal{U}'$ = unit (principal) planes.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

Further we set

$$f = \frac{b_2}{c_0}, \qquad f' = \frac{c_0 d_3 - c_3 d_0}{b_2 c_0}. \tag{9}$$

With this substitution, the equations of the transformation take the simple form

$$\frac{Y'}{Y} = \frac{f}{Z} = \frac{Z'}{f'}. \tag{10}$$

The second relation, $ZZ' = ff'$, is usually called *Newton's equation*. The constant f is known as *the focal length of the object space* and f' *as the focal length of the image space*.

From (10) it follows that, for fixed object and image planes,

$$\left(\frac{\mathrm{d}Y'}{\mathrm{d}Y}\right)_{Z=\text{constant}} = \frac{Y'}{Y} = \frac{f}{Z} = \frac{Z'}{f'}. \tag{11}$$

This quantity is known as the *lateral magnification*. Further, we have independently of $Y$and Y', the *longitudinal magnification*

$$\frac{\mathrm{d}Z'}{\mathrm{d}Z} = -\frac{Z'}{Z} = -\frac{ff'}{Z^2} = -\frac{Z'^2}{ff'}. \tag{12}$$

From (11) and (12) it is seen that the two magnifications are related by the formula

$$\frac{\mathrm{d}Z'}{\mathrm{d}Z} = -\frac{f'}{f}\left(\frac{\mathrm{d}Y'}{\mathrm{d}Y}\right)^2_{Z=\text{constant}}. \tag{13}$$

Since the lateral magnification depends on Z but not on Y it follows that an object which is situated in a plane perpendicular to the axis will be transformed into one which is geometrically similar to it.

The *lateral magnification* is equal to unity when $Z = f$ and $Z' = f'$. These planes are called the *unit* or *principal planes* and are denoted by the letters $\mathcal{U}$ and $\mathcal{U}'$ in Figs. 4.10 and 4.11. The points U and U' in which these planes intersect the axis are known as the *unit* or *principal points*.

Let h be the distance from the axis at which a ray from an axial point (0, 0, Z) meets the unit plane of the object space. The conjugate ray meets the other unit plane

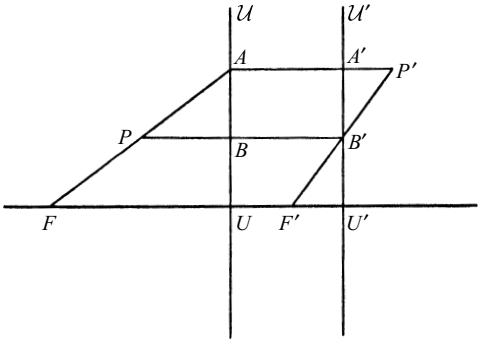

Fig. 4.11 Graphical determination of the image point.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

at the same distance (h) from the axis; and the angles γ and γ' which the two rays make with the axis are given by (see Fig. 4.10)

$$\tan\gamma = \frac{h}{f - Z}, \qquad \tan\gamma' = \frac{h}{f' - Z'}.$$

The ratio

$$\frac{\tan\gamma'}{\tan\gamma} = \frac{f - Z}{f' - Z'} = -\frac{Z}{f'} = -\frac{f}{Z'} \tag{14}$$

is known as the *angular magnification* or *convergence ratio*. It is independent of h and h' and is equal to unity when $Z = -f'$ and $Z' = -f$. The two axial points N and N' specified by these expressions are known as *nodal points*. They are the two conjugate points which are characterized by the property that conjugate rays passing through them are parallel to each other. The distance between the nodal points is equal to the distance between the unit points. When $f = -f'$ the nodal points and the unit points coincide. The foci, the unit points and the nodal points are called *cardinal* points of the transformation.

It is sometimes convenient to measure distances from the unit planes rather than from the focal planes. This means that in place of Z and Z' we use the variables

$$\zeta = Z - f, \qquad \zeta' = Z' - f'. \tag{15}$$

Newton's equation $ZZ' = ff'$ then becomes

$$\frac{f}{\zeta} + \frac{f'}{\zeta'} = -1. \tag{16}$$

Using the properties of the focal points and the principal planes, we can find the point P' conjugate to a given point P by a simple geometrical construction. Two lines are drawn through P, one parallel to the axis and the other through the focus F (Fig. 4.11). Let A and B be the points in which these lines intersect the unit plane $\mathcal{U}$. From the property of the unit planes it then follows that the points A' and B', which are conjugate to A and B, are the points of intersection with $\mathcal{U}'$ of the two lines through A and B drawn parallel to the axis. Moreover, since PA passes through the focus F, $P'A'$ must be parallel to the axis, and since PB is parallel to the axis, $P'B'$ must pass through the other focus F'. Hence the image point P' is the point of intersection of AA' with $B'F'$.

4.3.2 The telescopic case

We shall now consider the special case of telescopic (affine) collineation. As already explained it is characterized by having both focal planes at infinity. The coefficient c_0 in (5) and (6) then vanishes, and (5) reduces to

$$y' = \frac{b_2 y}{d_0}, \qquad z' = \frac{c_3 z + d_3}{d_0}. \tag{17}$$

We shall again choose separate coordinate systems for the object space and for the image space. As origins we take any conjugate axial pair of points. Referred to the new set of axes, the equations of the transformation take the simple form

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$Y' = \alpha Y, \qquad Z' = \beta Z, \tag{18}$$

where $\alpha = b_2/d_0$ and $\beta = c_3/d_0$.

The lateral as well as the longitudinal magnification is now constant. The angular magnification is also constant; to show this, consider two conjugate rays and take the origins of the two coordinate systems at the points in which these rays intersect the axis. The equations of the two rays then are

$$Y = Z\tan\gamma, \qquad Y' = Z'\tan\gamma'. \tag{19}$$

Hence

$$\frac{\tan\gamma'}{\tan\gamma} = \frac{Y'}{Z'}\frac{Z}{Y} = \frac{\alpha}{\beta} = \frac{b_2}{c_3}. \tag{20}$$

Although in a telescopic transformation both f and f' are infinite, the ratio f'/f must be regarded as finite. For by (9), as $c_0 \to 0$,

$$\frac{f'}{f} \to -\frac{c_3 d_0}{{b_2}^2} = -\frac{\beta}{\alpha^2}. \tag{21}$$

4.3.3 Classification of projective transformations

Projective transformations may be classified according to the signs of the focal lengths.

When the focal lengths are of opposite signs, $ff' < 0$, and according to (12) $\mathrm{d}Z'/\mathrm{d}Z > 0$. This implies that if the object is displaced in a direction parallel to the axis, the image will be displaced in the same direction. It will be seen later that such imaging occurs whenever the image is produced either by refraction alone, or by an even number of reflections, or by a combination of the two. Such imaging is called *concurrent* or *dioptric*.

When the focal lengths have equal signs, $ff' > 0$, $\mathrm{d}Z'/\mathrm{d}Z < 0$, so that to a displacement of the object in a direction parallel to the axis, there corresponds a displacement of the image in the opposite direction. This type of imaging is produced by means of an odd number of reflections, or by a combination of an odd number of reflections with any number of refractions. Imaging of this type is called *contracurrent* or *katoptric*.

In each group two types of transformations are distinguished, according to the sign of the focal length f. It is seen from (11) that for $Z > 0$ the lateral magnification is positive or negative according as $f > 0$ or $f < 0$. Hence an object situated in the right-hand half of the object space will have an image which is similarly oriented or inverted, according as f is positive or negative. In the former case the transformation is said to be *convergent*; in the latter, it is said to be *divergent*. This terminology is derived from the fact that an incident bundle of parallel rays, after passing the unit plane of the image space, is rendered convergent in the former case and divergent in the latter. The four cases are summarized in Table 4.1.

In the special case when the transformation is telescopic, the four types are distinguished by the signs of α and β. It follows from (21) that the transformation will be concurrent when β is positive and contracurrent when it is negative. Further it is

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

Table 4.1. *Classification of projective transformations (Cartesian sign-convention).*

	Convergent	Divergent
Concurrent (dioptric)	$f>0;\ f'<0$	$f<0;\ f'>0$
Contracurrent (katoptric)	$f>0;\ f'>0$	$f<0;\ f'<0$

seen from (18) that it will be convergent or divergent according as to α is positive or negative.

4.3.4 Combination of projective transformations

We now consider the combination of two successive projective transformations, which are rotationally symmetrical about the same axis.

Let the subscript 0 refer to the first transformation, and subscript 1 to the second. Then the equations specifying the two transformations are:

$$\left.\begin{aligned} \frac{Y_0'}{Y_0}=\frac{f_0}{Z_0}=\frac{Z_0'}{f_0'},\\ \frac{Y_1'}{Y_1}=\frac{f_1}{Z_1}=\frac{Z_1'}{f_1'}. \end{aligned}\right\} \tag{22}$$

Let c be the distance between the focal points F_0' and F_1. Since the image space of the first transformation coincides with the object space of the second,

$$Z_1 = Z_0' - c, \qquad Y_1 = Y_0'. \tag{23}$$

Elimination of the coordinates of the intermediate space from (22) by means of (23) gives

$$\left.\begin{aligned} Y_1' &= \frac{Z_1' Y_1}{f_1'} = \frac{Z_1' Y_0'}{f_1'} = \frac{Z_1' f_0 Y_0}{f_1' Z_0} = \frac{f_0 f_1 Y_0}{f_0 f_0' - c Z_0},\\ Z_1' &= \frac{f_1 f_1'}{Z_1} = \frac{f_1 f_1'}{Z_0' - c} = \frac{f_1 f_1'}{\dfrac{f_0 f_0'}{Z_0} - c} = \frac{f_1 f_1' Z_0}{f_0 f_0' - c Z_0}. \end{aligned}\right\} \tag{24}$$

Let

$$\left.\begin{aligned} Y = Y_0, \quad Z = Z_0 - \frac{f_0 f_0'}{c},\\ Y' = Y_1', \quad Z' = Z_1' + \frac{f_1 f_1'}{c}. \end{aligned}\right\} \tag{25}$$

Eqs. (25) express a change of coordinates, the origins of the two systems being shifted by distances $f_0' f_0/c$ and $-f_1 f_1'/c$, respectively, in the Z direction. In terms of these variables, the equations of the combined transformation become

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\frac{Y'}{Y} = \frac{f}{Z} = \frac{Z'}{f'}, \tag{26}$$

where

$$f = -\frac{f_0 f_1}{c}, \qquad f' = \frac{f_0' f_1'}{c}. \tag{27}$$

The distances between the origins of the new and the old systems of coordinates, i.e. the distances $\delta = F_0F$ and $\delta' = F_1'F'$ of the foci of the equivalent transformation from the foci of the individual transformations, are seen from (25) to be

$$\delta = \frac{f_0 f_0'}{c}, \qquad \delta' = -\frac{f_1 f_1'}{c}. \tag{28}$$

If $c = 0$, then $f = f' = \infty$ so that the equivalent collineation is telescopic. Eqs. (24) then reduce to

$$\left.\begin{aligned} Y_1' &= \frac{f_1}{f_0'} Y_0, \\ Z_1' &= \frac{f_1 f_1'}{f_0 f_0'} Z_0; \end{aligned}\right\} \tag{29}$$

the constants α and β in (18) of the equivalent transformation are therefore

$$\alpha = \frac{f_1}{f_0'}, \qquad \beta = \frac{f_1 f_1'}{f_0 f_0'}. \tag{30}$$

The angular magnification is now

$$\frac{\tan\gamma'}{\tan\gamma} = \frac{\alpha}{\beta} = \frac{f_0}{f_1'}. \tag{31}$$

If one or both of the transformations are telescopic, the above considerations must be somewhat modified.

4.4 Gaussian optics

We shall now study the elementary properties of lenses, mirrors and their combinations. In this elementary theory only those points and rays will be considered which lie in the immediate neighbourhood of the axis; terms involving squares and higher powers of off-axis distances, or of the angles which the rays make with the axis, will be neglected. The resulting theory is known as Gaussian optics.*

4.4.1 Refracting surface of revolution

Consider a pencil of rays incident on a refracting surface of revolution which separates two homogeneous media of refractive indices n_0 and n_1. To begin with, points and rays in both media will be referred to the same Cartesian reference system, whose origin

* As before, the usual sign convention of analytical geometry (Cartesian sign convention) is used. The various sign conventions employed in practice are very fully discussed in a Report on the Teaching of Geometrical Optics published by the Physical Society (London) in 1934.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

will be taken at the pole O of the surface, with the z direction along the axis of symmetry.

Let $P_0(x_0, y_0, z_0)$ and $P_1(x_1, y_1, z_1)$ be points on the incident and on the refracted ray, respectively. Neglecting terms of degree higher than first, it follows from §4.1 (29), §4.1 (40) and §4.1 (44) that the coordinates of these points and the components of the two rays are connected by the relations

$$\left.\begin{aligned} x_0 - \frac{p_0}{n_0} z_0 &= \frac{\partial T^{(2)}}{\partial p_0} = 2\mathcal{A} p_0 + \mathcal{C} p_1, \\ x_1 - \frac{p_1}{n_1} z_1 &= -\frac{\partial T^{(2)}}{\partial p_1} = -2\mathcal{B} p_1 - \mathcal{C} p_0, \end{aligned}\right\} \tag{1a}$$

$$\left.\begin{aligned} y_0 - \frac{q_0}{n_0} z_0 &= \frac{\partial T^{(2)}}{\partial q_0} = 2\mathcal{A} q_0 + \mathcal{C} q_1, \\ y_1 - \frac{q_1}{n_1} z_1 &= -\frac{\partial T^{(2)}}{\partial q_1} = -2\mathcal{B} q_1 - \mathcal{C} q_0, \end{aligned}\right\} \tag{1b}$$

where, according to §4.1 (45),

$$\mathcal{A} = \mathcal{B} = \frac{1}{2}\frac{r}{n_1 - n_0}, \qquad \mathcal{C} = -\frac{r}{n_1 - n_0}, \tag{2}$$

r being the paraxial radius of curvature of the surface.

Let us examine under what conditions all the rays from P_0 (which may be assumed to lie in the plane $x = 0$) will, after refraction, pass through P_1. The coordinates of P_1 will then depend only on the coordinates of P_0 and not on the components of the rays, so that when q_1 is eliminated from (1b), q_0 must also disappear.

Now from the first equation (1b)

$$q_1 = \frac{1}{\mathcal{C}}\left[y_0 - q_0\left(2\mathcal{A} + \frac{1}{n_0} z_0\right)\right], \tag{3}$$

and substituting this into the second equation, we obtain

$$y_1 = -\left(2\mathcal{B} - \frac{1}{n_1} z_1\right)\frac{1}{\mathcal{C}} y_0 + \left[\frac{1}{\mathcal{C}}\left(2\mathcal{B} - \frac{1}{n_1} z_1\right)\left(2\mathcal{A} + \frac{1}{n_0} z_0\right) - \mathcal{C}\right] q_0. \tag{4}$$

Hence P_1 will be a stigmatic image of P_0 if

$$\left(2\mathcal{A} + \frac{1}{n_0} z_0\right)\left(2\mathcal{B} - \frac{1}{n_1} z_1\right) = \mathcal{C}^2, \tag{5}$$

or, on substituting from (2), if

$$\left(\frac{r}{n_1 - n_0} + \frac{z_0}{n_0}\right)\left(\frac{r}{n_1 - n_0} - \frac{z_1}{n_1}\right) = \frac{r^2}{(n_1 - n_0)^2}. \tag{6}$$

Eq. (6) may be written in the form

$$n_0\left(\frac{1}{r} - \frac{1}{z_0}\right) = n_1\left(\frac{1}{r} - \frac{1}{z_1}\right). \tag{7}$$

It is seen that, within the present degree of approximation, every point gives rise to a

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

stigmatic image; and the distances of the conjugate planes from the pole O of the surface are related by (7). Moreover, (4), subject to (5), implies that the imaging is a *projective transformation*.

The expressions on either side of (7) are known as *Abbe's* (*refraction*) *invariant* and play an important part in the theory of optical imaging. Eq. (7) may also be written in the form

$$\frac{n_1}{z_1} - \frac{n_0}{z_0} = \frac{n_1 - n_0}{r}. \tag{8}$$

The quantity $(n_1 - n_0)/r$ is known as the *power* of the refracting surface and will be denoted by $\mathcal{P}$,

$$\mathcal{P} = \frac{n_1 - n_0}{r}. \tag{9}$$

According to (4) and (5) the lateral magnification y_1/y_0 is equal to unity when $z_1/n_1 = 2\mathcal{B} + \mathcal{C}$. But by (2), $2\mathcal{B} + \mathcal{C} = 0$. Hence the unit points U_0 and U_1 are given by $z_0 = z_1 = 0$, i.e. *the unit points coincide with the pole of the surface*. Further, from (8),

$$z_0 \to -\frac{n_0 r}{n_1 - n_0} \qquad \text{as } z_1 \to \infty$$

and

$$z_1 \to \frac{n_1 r}{n_1 - n_0} \qquad \text{as } z_0 \to -\infty,$$

so that the abscissae of the foci F_0 and F_1 are $-n_0 r/(n_1 - n_0)$ and $n_1 r/(n_1 - n_0)$, respectively. The focal lengths $f_0 = F_0 U_0$, $f_1 = F_1 U_1$ are therefore given by

$$\left.\begin{aligned} f_0 &= \frac{n_0 r}{n_1 - n_0} \\ \text{and} \qquad f_1 &= -\frac{n_1 r}{n_1 - n_0}; \end{aligned}\right\} \tag{10}$$

or, in terms of the power $\mathcal{P}$ of the surface,*

$$\frac{n_0}{f_0} = -\frac{n_1}{f_1} = \mathcal{P}. \tag{11}$$

Since f_0 and f_1 have different signs, the imaging is *concurrent* (see §4.3.3). If the surface is convex towards the incident light ($r > 0$) and if $n_0 < n_1$ then $f_0 > 0$, $f_1 < 0$, and the imaging is convergent. If $r > 0$ and $n_0 > n_1$ it is *divergent* (Fig. 4.12). When the surface is concave towards the incident light the situation is reversed.

Using the expressions (10) for the focal lengths, (8) becomes

* It will be seen later that the relation $n_0/f_0 = -n_1/f_1$ is not restricted to a single refracting surface, but holds in general for any centred system, the quantities with suffix zero referring to the object space, those with the suffix 1 to the image space. Eq. (11) may therefore be regarded as defining the *power* of a general centred system. The practical unit of power is a *dioptre*; it is the reciprocal of the focal length, when the focal length is expressed in metres. The power is positive when the system is convergent ($f_0 > 0$) and negative when it is divergent ($f_0 < 0$).

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

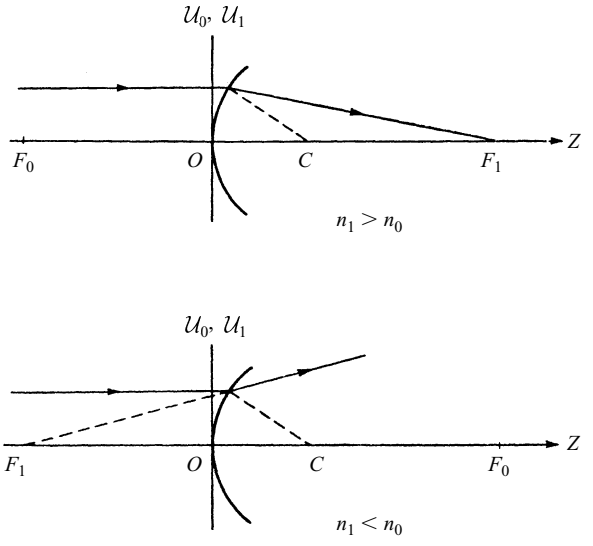

Fig. 4.12 Position of the cardinal points for refraction at a surface of revolution.

$$\frac{f_0}{z_0} + \frac{f_1}{z_1} = -1; \tag{12}$$

and the coefficients (2) may be written in the form

$$\mathcal{A} = \frac{f_0}{2n_0}, \qquad \mathcal{B} = -\frac{f_1}{2n_1}, \qquad \mathcal{C} = -\frac{f_0}{n_0} = \frac{f_1}{n_1}. \tag{13}$$

Let us introduce separate coordinate systems in the two spaces, with the origins at the foci, and with the axes parallel to those at O:

$$\begin{aligned} X_0 &= x_0, & X_1 &= x_1, \\ Y_0 &= y_0, & Y_1 &= y_1, \\ Z_0 &= z_0 + f_0, & Z_1 &= z_1 + f_1. \end{aligned}$$

Eqs. (12) and (4), subject to (5), then reduce to the standard form §4.3 (10):

$$\frac{Y_1}{Y_0} = \frac{f_0}{Z_0} = \frac{Z_1}{f_1}. \tag{14}$$

4.4.2 Reflecting surface of revolution

It is seen that with a notation strictly analogous to that used in the previous section, equations of the form (1a) and (1b) also hold when reflection takes place on a surface of revolution, the coefficients $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ now being replaced by the corresponding coefficients $\mathcal{A}'$, $\mathcal{B}'$ and $\mathcal{C}'$ of §4.1.5. It has been shown in §4.1.5 that $\mathcal{A}'$, $\mathcal{B}'$ and $\mathcal{C}'$ can be obtained from $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ on setting $n_0 = -n_1 = n$, n denoting the refractive index of the medium in which the rays are situated. Hence the appropriate formulae for reflection may be immediately written down by making this substitution in the preceding formulae. In particular, (7) gives

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

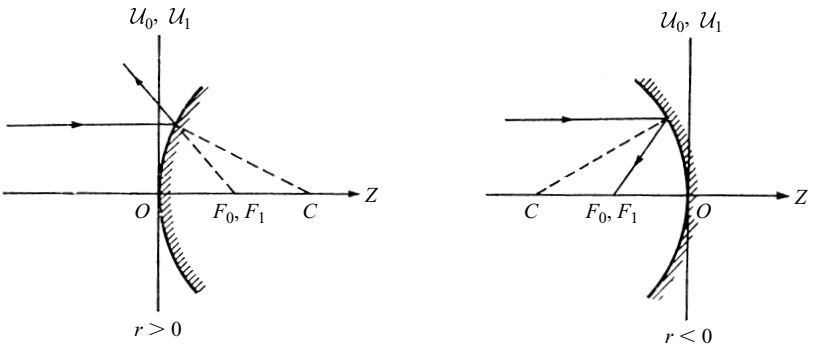

Fig. 4.13 Position of the cardinal points for reflection on a mirror of revolution.

$$\frac{1}{z_0} - \frac{1}{r} = -\frac{1}{z_1} + \frac{1}{r}, \tag{15}$$

the expression on either side of (15) being *Abbe's* (*reflection*) *invariant*. Eq. (15) may also be written as

$$\frac{1}{z_0} + \frac{1}{z_1} = \frac{2}{r}. \tag{16}$$

The focal lengths f_0 and f_1 are now given by

$$f_0 = f_1 = -\frac{r}{2}, \tag{17}$$

and the power $\mathcal{P}$ is

$$\mathcal{P} = -\frac{2n}{r}. \tag{18}$$

Since $f_0 f_1 > 0$, the imaging is *contracurrent* (katoptric). When the surface is convex towards the incident light ($r > 0$) $f_0 < 0$, and the imaging is then divergent; when it is concave towards the incident light ($r < 0$) $f_0 > 0$, and the imaging is convergent (Fig. 4.13).

4.4.3 The thick lens

Next we derive the Gaussian formulae relating to imaging by two surfaces which are rotationally symmetrical about the same axis.

Let n_0, n_1 and n_2 be the refractive indices of the three regions, taken in the order in which light passes through them, and let r_1 and r_2 be the radii of curvature of the surfaces at their axial points, measured positive when the surface is convex towards the incident light.

By (10), the focal lengths of the first surface are given by

$$f_0 = \frac{n_0 r_1}{n_1 - n_0}, \qquad f_0' = -\frac{n_1 r_1}{n_1 - n_0}, \tag{19}$$

and of the second surface by

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$f_1 = \frac{n_1 r_2}{n_2 - n_1}, \qquad f_1' = -\frac{n_2 r_2}{n_2 - n_1}. \tag{20}$$

According to §4.3 (27) the focal lengths of the combination are

$$f = -\frac{f_0 f_1}{c}, \qquad f' = \frac{f_0' f_1'}{c}, \tag{21}$$

where c is the distance between the foci F_0' and F_1. Let t be the axial thickness of the lens, i.e. the distance between the poles of the two surfaces; then (see Fig. 4.14)

$$c = t + f_0' - f_1. \tag{22}$$

On substituting into (22) for f_0' and for f_1 we obtain

$$c = \frac{D}{(n_1 - n_0)(n_2 - n_1)}, \tag{23}$$

where

$$D = (n_1 - n_0)(n_2 - n_1)t - n_1[(n_2 - n_1)r_1 + (n_1 - n_0)r_2]. \tag{24}$$

The required expressions for the focal lengths of the combination are now obtained on substituting for f_0, f_1, f_0', f_1' and c into (21). This gives

$$f = -n_0 n_1 \frac{r_1 r_2}{D}; \qquad f' = n_1 n_2 \frac{r_1 r_2}{D}. \tag{25}$$

Since ff' is negative, the imaging is concurrent. The power $\mathcal{P}$ of the lens is

$$\begin{aligned} \mathcal{P} &= \frac{n_0}{f} = -\frac{n_2}{f'} = -\frac{1}{n_1}\frac{D}{r_1 r_2} \\ &= \mathcal{P}_1 + \mathcal{P}_2 - \frac{t}{n_1}\mathcal{P}_1\mathcal{P}_2, \end{aligned} \tag{26}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the powers of the two surfaces.

By §4.3 (28), the distances $\delta = F_0 F$ and $\delta' = F_1' F'$ are seen to be

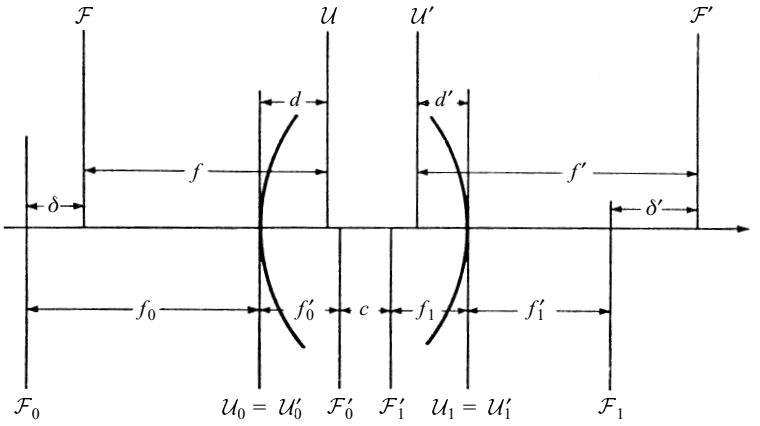

Fig. 4.14 The cardinal points of a combined system (thick lens).

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\delta = -n_0 n_1 \frac{n_2 - n_1}{n_1 - n_0}\frac{r_1^2}{D}, \qquad \delta' = n_1 n_2 \frac{n_1 - n_0}{n_2 - n_1}\frac{r_2^2}{D}. \tag{27}$$

The distances d and d' of the principal planes $\mathcal{U}$ and $\mathcal{U}'$ from the poles of the surfaces are (see Fig. 4.14)

$$\left.\begin{aligned} d &= \delta + f - f_0 = -n_0(n_2 - n_1)\frac{r_1 t}{D}, \\ d' &= \delta' + f' - f_1{}' = n_2(n_1 - n_0)\frac{r_2 t}{D}. \end{aligned}\right\} \tag{28}$$

Of particular importance is the case where the media on both sides of the lens are of the same refractive index, i.e. when $n_2 = n_0$. If we set $n_1/n_0 = n_1/n_2 = n$, the formulae then reduce to

$$\left.\begin{aligned} f &= -f' = -\frac{n r_1 r_2}{\Delta}, \\ \delta &= n\frac{r_1{}^2}{\Delta}, \qquad \delta' = -n\frac{r_2{}^2}{\Delta}, \\ d &= (n-1)\frac{r_1 t}{\Delta}, \qquad d' = (n-1)\frac{r_2 t}{\Delta}, \end{aligned}\right\} \tag{29}$$

where

$$\Delta = (n-1)[n(r_1 - r_2) - (n-1)t]. \tag{30}$$

Referred to axes at the foci F and F', the abscissae of the unit points are $Z = f$ and $Z' = f'$, and the nodal points are given by $Z = -f'$, $Z = -f$. Since $f = -f'$ the unit points and the nodal points now coincide. Formula §4.3 (16) which relates the distances ζ and ζ' of conjugate planes from the unit planes becomes

$$\frac{1}{\zeta} - \frac{1}{\zeta'} = -\frac{1}{f}. \tag{31}$$

The lens is *convergent* ($f > 0$) or *divergent* ($f < 0$) according as

$$f = -n\frac{r_1 r_2}{\Delta} \gtrless 0, \tag{32}$$

i.e. according as

$$\frac{1}{r_2} - \frac{1}{r_1} \lessgtr \frac{n-1}{n}\frac{t}{r_1 r_2}. \tag{33}$$

When $f = \infty$, we have the intermediate case of *telescopic* imaging. Then $\Delta = 0$, i.e.

$$r_1 - r_2 = \frac{n-1}{n}t. \tag{34}$$

The three cases may be illustrated by considering a double convex lens, $r_1 > 0$, $r_2 < 0$ (Fig. 4.15). If, for simplicity, we assume that both radii are numerically equal, i.e. $r_1 = -r_2 = r$, the imaging will be convergent or divergent according as $t \lessgtr 2nr/(n-1)$ and telescopic when $t = 2nr/(n-1)$.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

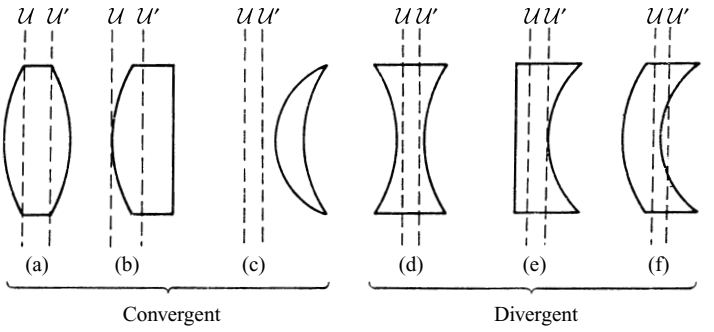

Fig. 4.15 Common types of lens: (a) double-convex; (b) plano-convex; (c) convergent meniscus; (d) double-concave; (e) plano-concave; (f) divergent meniscus. $\mathcal{U}$ and $\mathcal{U}'$ are the unit planes, light being assumed to be incident from the left.

4.4.4 The thin lens

The preceding formulae take a particularly simple form when the lens is so thin that the axial thickness t may be neglected. Then, according to (28), $d = d' = 0$, so that the unit planes pass through the axial point of the (infinitely thin) lens. Consequently, the rays which go through the centre of the lens will not suffer any deviation; this implies that *imaging by a thin lens is a central projection from the centre of the lens*.

From (26) it follows, on setting $t = 0$, that

$$\mathcal{P} = \mathcal{P}_1 + \mathcal{P}_2 = \frac{n_1 - n_0}{r_1} + \frac{n_2 - n_1}{r_2}, \tag{35}$$

i.e. the power of a thin lens is equal to the sum of the powers of the surfaces forming it.

If the media on the two sides of the lens are of equal refractive indices ($n_0 = n_2$), we have from (25)

$$\frac{1}{f} = -\frac{1}{f'} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right), \tag{36}$$

where as before, $n = n_1/n_0 = n_1/n_2$. Assuming that $n > 1$, as is usually the case, f is seen to be positive or negative according as the curvature $1/r_1$ of the first surface is greater or smaller than the curvature $1/r_2$ of the second surface (an appropriate sign being associated with the curvature). This implies that thin lenses whose thicknesses decrease from centre to the edge are convergent, and those whose thicknesses increase to the edge are divergent.

For later purposes we shall write down an expression for the focal lengths f and f' of a system consisting of two centred thin lenses, situated in air. According to §4.3 (27) we have, since $f_0 = -f_0'$, $f_1 = -f_1'$,

$$\frac{1}{f} = -\frac{1}{f'} = -\frac{c}{f_0 f_1}, \tag{37}$$

c being the distance between the foci F_0' and F_1 (see Fig. 4.16). If l is the distance between the two lenses, then

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

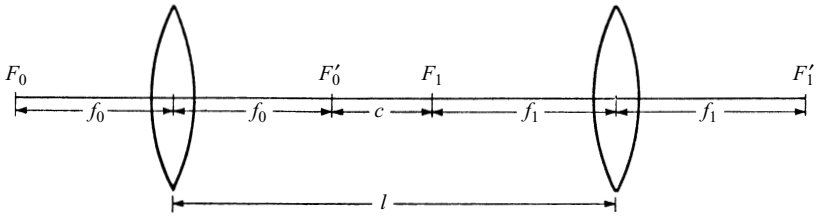

Fig. 4.16 System formed by two centred thin lenses.

$$l = f_0 + c + f_1. \tag{38}$$

Hence

$$\frac{1}{f} = -\frac{1}{f'} = \frac{1}{f_0} + \frac{1}{f_1} - \frac{l}{f_0 f_1}. \tag{39}$$

When the lenses are in contact ($l = 0$), (39) may also be written in the form $\mathcal{P} = \mathcal{P}_1 + \mathcal{P}_2$, so that the power of the combination is then simply equal to sum of the powers of the two lenses.

4.4.5 The general centred system

Within the approximations of Gaussian theory, a refraction or a reflection at a surface of revolution was seen to give rise to a projective relationship between the object and the image spaces.* Since according to §4.3 successive applications of projective transformations are equivalent to a single projective transformation, it follows that imaging by a centred system is, to the present degree of approximation, also a transformation of this type. The cardinal points of the equivalent transformation may be found by the application of the formulae of §4.4.1, §4.4.2 and §4.3.4. We shall mainly confine our discussion to the derivation of an important invariant relation valid (within the present degree of accuracy) for any centred system.

Let $S_1, S_2, \ldots S_m$ be the successive surfaces of the system, $f_0, f_1, \ldots f_{m-1}$ the corresponding focal lengths, and $n_0, n_1 \ldots n_m$ the refractive indices of the successive spaces (Fig. 4.17). Further, let $P_0, P_0^\star$ be two points in the object space, situated in a meridional plane, and let $P_1, P_1^\star, P_2, P_2^\star, \ldots,$ be their images in the successive surfaces. Referred to axes at the foci of the first surface, the coordinates of $P_0, P_0^\star$ and $P_1, P_1^\star$ are related by (14), viz.

$$\frac{Y_1}{Y_0} = \frac{f_0}{Z_0} = \frac{Z_1}{f_1}, \tag{40}$$

$$\frac{Y_1^\star}{Y_0^\star} = \frac{f_0}{Z_0^\star} = \frac{Z_1^\star}{f_1}. \tag{41}$$

* As in the case of a single surface, the object and the image spaces are regarded as superposed onto each other and extending indefinitely in all directions. The part of the object space which lies in front of the first surface (counted in the order in which the light traverses the system) is said to form the real portion of the object space and the portion of the image space which follows the last surface is called the real portion of the image space. The remaining portions of the two spaces are said to be *virtual*. In a similar manner we may define the real and virtual parts of any of the intermediate spaces of the system.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

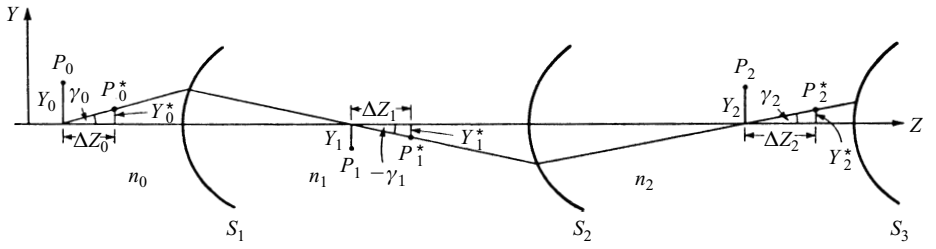

Fig. 4.17 Illustrating the Smith–Helmholtz formula.

Hence

$$\frac{Z_1^\star - Z_1}{f_1} = \frac{Y_1^\star}{Y_0^\star} - \frac{Y_1}{Y_0} = \frac{Y_0 Y_1^\star - Y_1 Y_0^\star}{Y_0 Y_0^\star}, \tag{42}$$

and

$$\frac{Z_0^\star - Z_0}{f_0} = \frac{Y_0^\star}{Y_1^\star} - \frac{Y_0}{Y_1} = \frac{Y_1 Y_0^\star - Y_0 Y_1^\star}{Y_1 Y_1^\star}. \tag{43}$$

Let

$$\left.\begin{aligned} Z_0^\star - Z_0 &= \Delta Z_0, \\ Z_1^\star - Z_1 &= \Delta Z_1. \end{aligned}\right\} \tag{44}$$

Then from (42) and (43),

$$\frac{f_0 Y_0 Y_0^\star}{\Delta Z_0} = -\frac{f_1 Y_1 Y_1^\star}{\Delta Z_1}. \tag{45}$$

Now by (10), $f_0/f_1 = -n_0/n_1$, so that this equation may be written as

$$\frac{n_0 Y_0 Y_0^\star}{\Delta Z_0} = \frac{n_1 Y_1 Y_1^\star}{\Delta Z_1}. \tag{46}$$

Similarly, for refraction at the second surface,

$$\frac{n_1 Y_1 Y_1^\star}{\Delta Z_1} = \frac{n_2 Y_2 Y_2^\star}{\Delta Z_2}, \tag{47}$$

and generally

$$\frac{n_{i-1} Y_{i-1} Y_{i-1}^\star}{\Delta Z_{i-1}} = \frac{n_i Y_i Y_i^\star}{\Delta Z_i} \qquad (1 \leqslant i \leqslant m). \tag{48}$$

Hence $n_i Y_i Y_i^\star / \Delta Z_i$ *is an invariant in the successive transformations*. This result plays an important part in the geometrical theory of image formation. If we set $Y_i^\star / \Delta Z_i = \tan \gamma_i$ (see Fig. 4.17), (48) becomes

$$n_{i-1} Y_{i-1} \tan \gamma_{i-1} = n_i Y_i \tan \gamma_i.$$

Since to the present degree of accuracy, $\tan \gamma$ and $\tan \gamma'$ may be replaced by γ and γ', respectively, we obtain the *Smith–Helmholtz formula*[*]

[*] This formula is also associated with the names of Lagrange and Clausius. It was actually known in more restricted forms to earlier writers, e.g. Huygens and Cotes. (See Lord Rayleigh, *Phil. Mag.*, (5) **21** (1886), 466.)

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$n_{i-1}Y_{i-1}\gamma_{i-1} = n_iY_i\gamma_i. \tag{49}$$

The quantity $n_iY_i\gamma_i$ is known as the *Smith–Helmholtz invariant*.

From (48) and (49) a number of important conclusions may be drawn. As one is usually interested only in relations between quantities pertaining to the first and the last medium (object and image space), we shall drop the suffixes and denote quantities which refer to these two spaces by unprimed and primed symbols respectively.

Let $(Y,\ Z)$ and $(Y+\delta Y,\ Z+\delta Z)$ be two neighbouring points in the object space and $(Y',\ Z')$ and $(Y'+\delta Y',\ Z'+\delta Z')$ the conjugate points. The Smith–Helmholtz formula gives, by successive application, the following relation:

$$\frac{nY(Y+\delta Y)}{\delta Z} = \frac{n'Y'(Y'+\delta Y')}{\delta Z'}. \tag{50}$$

In the limit as $\delta Y \to 0$ and $\delta Z \to 0$, this reduces to

$$\frac{\mathrm{d}Z'}{\mathrm{d}Z} = \frac{n'}{n}\frac{Y'^2}{Y^2}. \tag{51}$$

According to §4.3 (11), we may write $Y'/Y = (\mathrm{d}Y'/\mathrm{d}Y)_{Z=\text{constant}}$, and (51) becomes

$$\frac{\mathrm{d}Z'}{\mathrm{d}Z} = \frac{n'}{n}\left(\frac{\mathrm{d}Y'}{\mathrm{d}Y}\right)^2_{Z=\text{constant}}, \tag{52}$$

known as *Maxwell's elongation formula*. It implies that the *longitudinal magnification is equal to the square of the lateral magnification multiplied by the ratio n'/n of the refractive indices*. Now in §4.3 we derived an analogous formula ((13)) connecting the magnifications and the ratio of the focal lengths. On comparing these two formulae it is seen that

$$\frac{f'}{f} = -\frac{n'}{n}, \tag{53}$$

i.e. the *ratio of the focal lengths of the instrument is equal to the ratio n'/n of the refractive indices, taken with a negative sign*.

From the Smith–Helmholtz formula is also follows that

$$\frac{\mathrm{d}Y'}{\mathrm{d}Y}\frac{\gamma'}{\gamma} = \frac{n}{n'}, \tag{54}$$

showing that *the product of the lateral and the angular magnification is independent of the choice of the conjugate planes*.

It has been assumed so far that the system consists of refracting surfaces only. If one of the surfaces (say the ith one) is a mirror, we obtain in place of (48),

$$\frac{Y_{i-1}Y^{\star}_{i-1}}{\Delta Z_{i-1}} = -\frac{Y_iY_i^{\star}}{\Delta Z_i},$$

the negative sign arising from the fact that for reflection $f_{i-1}/f_i = 1$, whereas for refraction $f_{i-1}/f_i = -n_{i-1}/n_i$. In consequence n' must be replaced by $-n'$ in the final formulae. More generally n' must be replaced by $-n'$ when the system contains an odd number of mirrors; when it contains an even number of mirrors, the final formulae remain unchanged.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

4.5 Stigmatic imaging with wide-angle pencils

The laws of Gaussian optics were derived under the assumption that the size of the object and the angles which the rays make with the axis are sufficiently small. Often one also has to consider systems where the object is of small linear dimensions, but where the inclination of the rays is appreciable. There are two simple criteria, known as the *sine condition** and the *Herschel condition*† relating to the stigmatic imaging in such instruments.

Let O_0 be an axial object point and P_0 any point in its neighbourhood, not necessarily on the axis. Assume that the system images these two points stigmatically, and let O_1 and P_1 be the stigmatic images.

Let (x_0, y_0, z_0) and (x_1, y_1, z_1) be the coordinates of P_0 and P_1 respectively, P_0 being referred to rectangular axes at O_0 and P_1 to parallel axes at O_1, the z directions being taken along the axis of the system (Fig. 4.18). By the principle of equal optical path, the path lengths of all the rays joining P_0 and P_1 are the same. Hence, if V denotes the point characteristic of the medium,

$$V(x_0, y_0, z_0; x_1, y_1, z_1) - V(0, 0, 0; 0, 0, 0) = F(x_0, y_0, z_0; x_1, y_1, z_1), \qquad (1)$$

F being some function which is independent of the ray components. Using the basic relations §4.1 (7) which express the ray component in terms of the point characteristic, we have from (1), if terms above the first power in distances are neglected,

$$(p_1^{(0)}x_1 + q_1^{(0)}y_1 + m_1^{(0)}z_1) - (p_0^{(0)}x_0 + q_0^{(0)}y_0 + m_0^{(0)}z_0) = F(x_0, y_0, z_0; x_1, y_1, z_1), \qquad (2)$$

$(p_0^{(0)}, q_0^{(0)}, m_0^{(0)})$ and $(p_1^{(0)}, q_1^{(0)}, m_1^{(0)})$ being the ray components of any pair of corresponding rays through O_0 and O_1. It is to be noted that, although the points P_0 and P_1 are assumed to be in the neighbourhood of O_0 and O_1, no restriction as to the magnitude of the ray components is imposed.

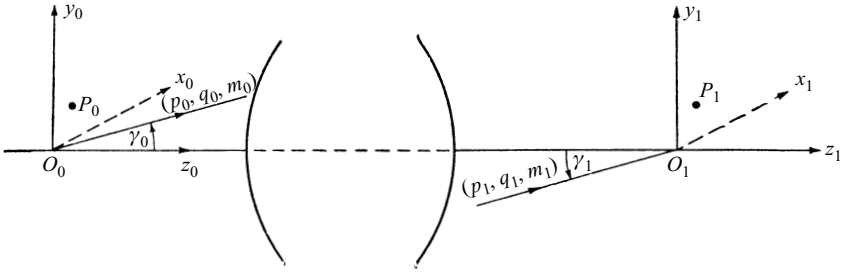

Fig. 4.18 Illustrating the sine condition and the Herschel condition.

* The sine condition was first derived by R. Clausius (*Pogg. Ann.*, **121** (1864), 1) and by H. Helmholtz (*Pogg. Ann. Jubelband* (1874), 557) from thermodynamical considerations. Its importance was, however, not recognized until it was rediscovered by E. Abbe (*Jenaisch. Ges. Med. Naturw.* 1879), 129; also *Carl. Repert. Phys.*, **16** (1880), 303).

The derivations given here are essentially due to C. Hockin, *J. Roy. Micro. Soc.* (2) **4** (1884), 337.

† J. F. W. Herschel, *Phil. Trans. Roy. Soc.*, **111** (1821), 226.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

Two cases are of particular interest, namely (i) when P_0 and P_1 lie in the planes $z_0 = 0$ and $z_1 = 0$, respectively, and (ii) when P_0 and P_1 are on the axis of symmetry. The two cases will be considered separately.

4.5.1 The sine condition

Without loss of generality we may again consider only points in a meridional plane ($x_0 = x_1 = 0$). If P_0 lies in the plane $z_0 = 0$ and P_1 in the plane $z_1 = 0$, (2) becomes

$$q_1^{(0)} y_1 - q_0^{(0)} y_0 = F(0,\ y_0,\ 0;\ 0,\ y_1,\ 0). \tag{3}$$

This relation holds for each pair of conjugate rays. In particular it must, therefore, hold for the axial pair $p_0^{(0)} = q_0^{(0)} = 0$, $p_1^{(0)} = q_1^{(0)} = 0$. Hence

$$F(0,\ y_0,\ 0;\ 0,\ y_1,\ 0) = 0. \tag{4}$$

Relation (3) becomes

$$q_1^{(0)} y_1 = q_0^{(0)} y_0, \tag{5}$$

or, more explicitly,

$$n_1 y_1 \sin\gamma_1 = n_0 y_0 \sin\gamma_0, \tag{6}$$

γ_0 and γ_1 being the angles which the corresponding rays through O_0 and O_1 make with the z-axis, and n_0 and n_1 being the refractive indices of the object and image spaces. Eq. (6) is known as the *sine condition*, and is the required condition under which a small region of the object plane in the neighbourhood of the axis is imaged sharply by a pencil of any angular divergence. If the angular divergence is sufficiently small, $\sin\gamma_0$ and $\sin\gamma_1$ may be replaced by γ_0 and γ_1, respectively, and the sine condition reduces to the Smith–Helmholtz formula, §4.4 (49).

If the object lies at infinity, the sine condition takes a different form. Assume first that the axial object point is at a great distance from the first surface. If Z_0 is the abscissa of this point referred now to axes at the first focus, and h_0 is the height above the axis at which a ray from the axial point meets the first surface, then $\sin\gamma_0 \sim -h_0/Z_0$; more precisely $Z_0 \sin\gamma_0/h_0 \to -1$ as $Z_0 \to -\infty$ whilst h_0 is kept fixed. Hence, if Z_0 is large enough, (6) may be written as

$$\frac{n_0}{n_1} h_0 = -\frac{y_1}{y_0} Z_0 \sin\gamma_1. \tag{7}$$

But by §4.3 (10), $y_1 Z_0/y_0 = f_0$, and by §4.4 (53) $n_0/n_1 = -f_0/f_1$, so that in the limit (6) reduces to (see Fig. 4.19)

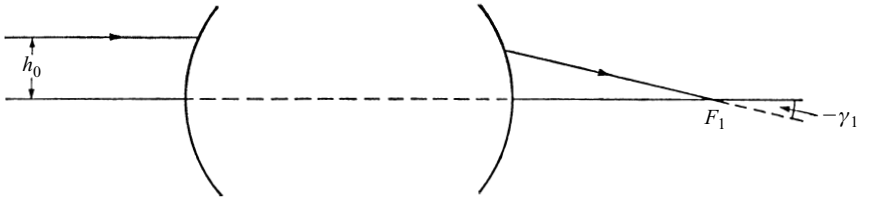

Fig. 4.19 The sine condition when the object is at infinity.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\frac{h_0}{\sin\gamma_1} = f_1. \tag{8}$$

This implies that each ray which is incident in the direction parallel to the axis intersects its conjugate ray on a sphere of radius f_1, which is centred at the focus F_1.

Axial points which are stigmatic images of each other, and which, in addition, have the property that conjugate rays which pass through them satisfy the sine condition, are said to form an *aplanatic* pair. We have already encountered such point pairs when studying the refraction at a spherical surface (§4.2.3).

In the terminology of the theory of aberrations (see Chapter V), axial stigmatism implies the absence of all those terms in the expansion of the characteristic function which do not depend on the off-axis distance of the object, i.e. it implies the absence of spherical aberration of all orders. If, in addition, the sine condition is satisfied, then all terms in the characteristic function which depend on the first power of the off-axis distance must also vanish; these terms represent aberrations known as *circular coma*.

Since the sine condition gives information about the quality of the off-axis image in terms of the properties of axial pencils it is of great importance for optical design.

4.5.2 The Herschel condition

Next consider the case when P_0 and P_1 lie on the axis of the system ($x_0 = y_0 = 0$, $x_1 = y_1 = 0$). The condition (2) for sharp imaging reduces to

$$m_1^{(0)} z_1 - m_0^{(0)} z_0 = F(0, 0, z_0; 0, 0, z_1), \tag{9}$$

or, in terms of γ_0 and γ_1,

$$n_1 z_1 \cos\gamma_1 - n_0 z_0 \cos\gamma_0 = F(0, 0, z_0; 0, 0, z_1). \tag{10}$$

In particular for the axial ray this gives

$$F(0, 0, z_0; 0, 0, z_1) = n_1 z_1 - n_0 z_0. \tag{11}$$

Hence (10) may be written as

$$n_1 z_1 \sin^2(\gamma_1/2) = n_0 z_0 \sin^2(\gamma_0/2). \tag{12}$$

This is one form of the *Herschel condition*. Since the distances from the origin are assumed to be small, we have, by §4.4 (52),

$$\frac{z_1}{z_0} = \frac{n_1}{n_0}\left(\frac{y_1}{y_0}\right)^2, \tag{13}$$

so that the Herschel condition may also be written in the form

$$n_1 y_1 \sin(\gamma_1/2) = n_0 y_0 \sin(\gamma_0/2). \tag{14}$$

When the Herschel condition is satisfied, an element of the axis in the immediate neighbourhood of O_0 will be imaged sharply by a pencil of rays, irrespective of the angular divergence of the pencil.

It is to be noted that the sine condition and the Herschel condition cannot hold simultaneously unless $\gamma_1 = \gamma_0$. Then $y_1/y_0 = z_1/z_0 = n_0/n_1$, i.e. the longitudinal and lateral magnifications must then be equal to the ratio of the refractive indices of the object and image space.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

4.6 Astigmatic pencils of rays

Rectilinear rays which have a point in common are said to form a *homocentric pencil*. The associated wave-fronts are then spherical, centred on their common point of intersection. It was with such pencils that we were concerned in the preceding sections.

In general, the homocentricity of a pencil is destroyed on refraction or reflection. It will therefore be useful to study the properties of more general pencils of rectilinear rays.

4.6.1 Focal properties of a thin pencil

Let S denote one of the orthogonal trajectories (wave-fronts) of a pencil of rectilinear rays, and let P be any point on it (Fig. 4.20). We take a plane through the ray at P and denote by C the curve in which this plane intersects S. Since the rays of the pencil are all normal to S, the centre of the circle of curvature at P will be on the ray through P.

If now the plane is gradually rotated around the ray, the curve C and consequently the radius of curvature will change continuously. When the plane has undergone a rotation of 180°, the radius of curvature will have passed through its maximum and minimum values. It can be shown by elementary geometry* that the two planes which contain the shortest and the longest radius of curvature are perpendicular to each other. These two planes are known as the *principal planes*† at P and the corresponding radii are called *principal radii of curvature*. The curves on S which have the property that at each point they are tangential to the principal planes form two mutually orthogonal families of curves, called the *lines of curvature*.

In general two normals at adjacent points of a surface do not intersect, to first order. But if they are drawn from adjacent points on a line of curvature they will intersect to this order, and their point of intersection is a focus of the congruence formed by the normals (rays). Hence, in agreement with the general conclusions of §3.2.3 there are two foci on each normal, these being the two principal centres of curvature. The caustic surface of a pencil of rectilinear rays has therefore, in general, two branches and is the *evolute* of the wave-fronts; conversely the wave-fronts are the *involutes* of the caustic surface. If the wave-fronts are surfaces of revolution, one branch of the

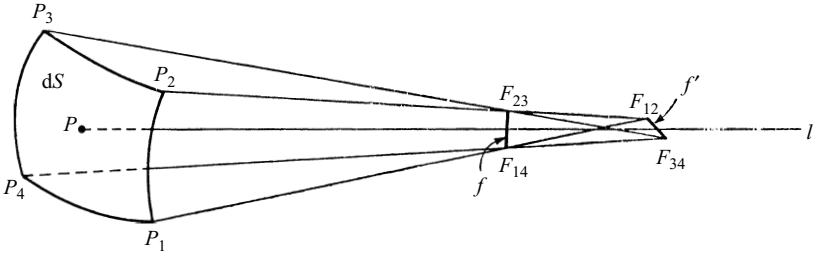

Fig. 4.20 A thin pencil of rays.

* See, for example, C. E. Weatherburn, *Differential Geometry of Three Dimensions* (Cambridge, Cambridge University Press, 1927), p. 185.

† The terms principal planes and focal planes have now a different meaning than in connection with projective transformations discussed in §4.3.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

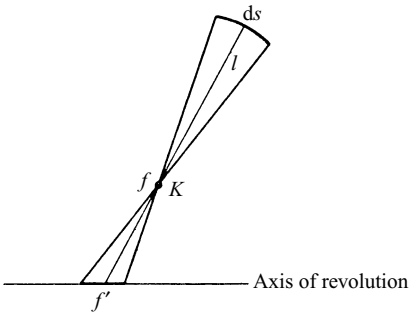

Fig. 4.21 Focal lines of a wave-front with cylindrical symmetry.

caustic surface degenerates into a segment of the axis of revolution; the other is a surface of revolution whose meridional section is the evolute of the meridional section of the wave-front.

Let us consider a thin pencil consisting of all rays whch intersect an element dS of the wave-front. It will be convenient to take as boundary of dS two pairs of lines of curvature, which may be assumed to be arcs of circles. Two of them (P_1P_2 and P_4P_3) may be taken to be vertical, and the other two (P_1P_4 and P_2P_3) to be horizontal (Fig. 4.20).

To the first order in small quantities, all the rays which pass through the arc P_1P_2 will intersect in a focus F_{12}, and those through P_3P_4 will intersect at F_{34}. The line f' joining F_{12} and F_{34} is known as a *focal line* of the pencil and is seen to be horizontal. Similarly the rays through P_1P_4 and P_2P_3 will give rise to a vertical focal line f.

If the lines of curvature are drawn through any point on dS, the corresponding two foci will lie on the two focal lines, and conversely. Hence *an approximate model of a thin pencil of rays is obtained by joining all pairs of points on two mutually orthogonal elements of lines*.

The ray l through the centre point P is called the *central* (or *principal*) *ray* of the pencil, and the distance between the focal lines, measured along this ray, is called the *astigmatic focal distance* of the pencil. The two planes specified by f and l, and by f' and l, are known as the *focal planes* of the pencil, and are mutually perpendicular. It is not, however, necessarily true (as is often incorrectly asserted in the literature) that the focal lines are perpendicular to the central ray. Consider for example a family of wave-fronts which have cylindrical symmetry about a common axis (Fig. 4.21). Let dS be a surface element (assumed not to contain an axial point) of one of the wave-fronts, and let ds be the curve of intersection of dS with a plane which contains the axis. Then, clearly, the focal line f at the centre of curvature K of ds is perpendicular to this plane. The other focal line, f', coincides with a portion of the axis, namely with that portion which is bounded by the normals at the end points of ds. In general this focal line is not perpendicular to l.

4.6.2 Refraction of a thin pencil

It was seen that a thin pencil of rays is completely specified by its central ray and its two focal lines. Suppose that a pencil specified in this way is incident on a refracting surface. It is of importance to determine the central ray and the focal lines of the

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

refracted pencil. We shall consider the case of particular importance in practice, namely, when one of the principal planes of the incident pencil coincides with a principal plane of curvature of the surface at the point O at which the central ray meets it (Fig. 4.22).

Take Cartesian axes at O, with Oz along the normal to the surface T and with Ox and Oy in the direction of the principal lines of curvature of T.

Further, let θ_0 and θ_1 be the angles which the central rays l_0 and l_1 of the two pencils make with Oz. Let F_0 and F_0' be the foci of the incident pencil, situated at $z = \zeta_0$ and $z = \zeta_0'$ respectively. The focal line at F_0 will be assumed to be perpendicular to the plane of incidence; F_0 is then said to be a *primary focus*, and the corresponding focal line f_0 *the primary focal line*. The focus F_0' is called the *secondary focus*; the corresponding focal line f_0' (which lies in the plane of incidence) is known as the *secondary focal line*. In the case of a centred system, the primary and the secondary foci of a pencil whose central ray lies in the meridional plane are known as the *tangential focus* and the *sagittal focus* respectively.

To find the focal lines of the refracted pencil it is necessary to write down first an expression for the angle characteristic of the refracting surface. If the radii of curvature of the surface in the principal directions Ox and Oy are r_x and r_y respectively, the equation of the surface is

$$z = \frac{x^2}{2r_x} + \frac{y^2}{2r_y} + \cdots . \tag{1}$$

Now according to §4.1 (34), the angle characteristic referred to a set of axes at O is (taking $a_0 = a_1 = 0$)

$$T = (p_0 - p_1)x + (q_0 - q_1)y + (m_0 - m_1)z. \tag{2}$$

From the law of refraction it follows, in a way similar to that described in §4.1 (which corresponds to the case $r_x = r_y$), when terms of the lowest degree only are retained, that

$$\left.\begin{aligned} x &= -r_x \frac{p_0 - p_1}{m_0 - m_1}, \\ y &= -r_y \frac{q_0 - q_1}{m_0 - m_1}. \end{aligned}\right\} \tag{3}$$

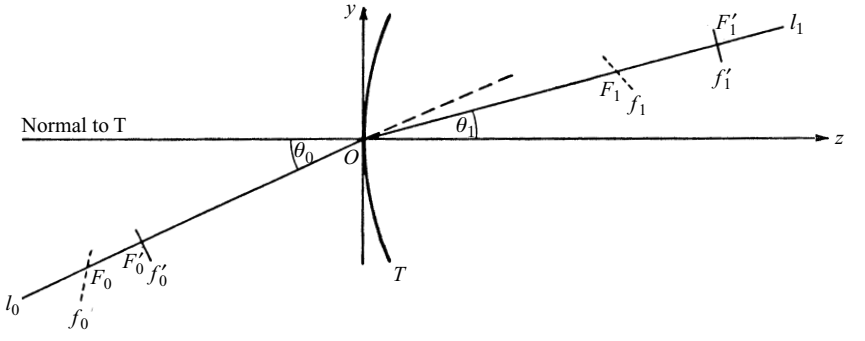

Fig. 4.22 Refraction of a thin astigmatic pencil.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

Substitution into (1) gives

$$z = \frac{1}{2\mu^2}[r_x(p_0 - p_1)^2 + r_y(q_0 - q_1)^2], \tag{4}$$

with

$$\mu = m_1 - m_0 = n_1 \cos\theta_1 - n_0 \cos\theta_0. \tag{5}$$

Substitution from (3) and (4) then leads to the required expression for the angle characteristic:

$$T(p_0, q_0; p_1, q_1) = \frac{1}{2\mu}[r_x(p_0 - p_1)^2 + r_y(q_0 - q_1)^2] + \cdots. \tag{6}$$

On using this expression in §4.1 (29), the equations of the incident and refracted ray are obtained:

$$x_0 - \frac{p_0}{m_0} z_0 = \frac{1}{\mu} r_x(p_0 - p_1), \tag{7}$$

$$x_1 - \frac{p_1}{m_1} z_1 = \frac{1}{\mu} r_x(p_0 - p_1), \tag{8}$$

the corresponding equations involving the y coordinates being strictly analogous.

Consider now the changes in the various quantities as we pass from the central ray to a neighbouring ray. From (7) and (8)

$$\delta x_0 - \frac{p_0}{m_0}\delta z_0 = z_0 \delta\left(\frac{p_0}{m_0}\right) + r_x\left[\delta\left(\frac{p_0}{\mu}\right) - \delta\left(\frac{p_1}{\mu}\right)\right], \tag{9}$$

$$\delta x_1 - \frac{p_1}{m_1}\delta z_1 = z_1 \delta\left(\frac{p_1}{m_1}\right) + r_x\left[\delta\left(\frac{p_0}{\mu}\right) - \delta\left(\frac{p_1}{\mu}\right)\right]. \tag{10}$$

Now the components of the central ray of the incident pencil are

$$p_0 = 0, \qquad q_0 = n_0 \sin\theta_0, \qquad m_0 = n_0 \cos\theta_0, \tag{11}$$

so that

$$\begin{aligned} \delta\left(\frac{p_0}{m_0}\right) &= \frac{1}{{m_0}^2}(m_0 \delta p_0 - p_0 \delta m_0) \\ &= \frac{1}{n_0} \sec\theta_0 \delta p_0, \end{aligned} \tag{12}$$

and

$$\begin{aligned} \delta\left(\frac{q_0}{m_0}\right) &= \frac{1}{{m_0}^2}(m_0 \delta q_0 - q_0 \delta m_0) \\ &= \frac{1}{n_0} \sec^3\theta_0 \delta q_0. \end{aligned} \tag{13}$$

Here use was made of the identity $m_0\delta m_0 + p_0\delta p_0 + q_0\delta q_0 = 0$.

Eqs. (9) and (10) become

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\delta x_0 = \frac{z_0}{n_0}\sec\theta_0\delta p_0 - \frac{1}{\mu}r_x(\delta p_1 - \delta p_0), \tag{14}$$

$$\delta x_1 = \frac{z_1}{n_1}\sec\theta_1\delta p_1 - \frac{1}{\mu}r_x(\delta p_1 - \delta p_0). \tag{15}$$

In deriving (15) use was also made of the fact that $p_1 = 0$; this result follows from the law of refraction and the assumption that $p_0 = 0$. In a similar way we find

$$\delta y_0 - (\tan\theta_0)\delta z_0 = \frac{z_0}{n_0}(\sec^3\theta_0)\delta q_0 - r_y\left[\delta\left(\frac{q_1}{\mu}\right) - \delta\left(\frac{q_0}{\mu}\right)\right], \tag{16}$$

$$\delta y_1 - (\tan\theta_1)\delta z_1 = \frac{z_1}{n_1}(\sec^3\theta_1)\delta q_1 - r_y\left[\delta\left(\frac{q_1}{\mu}\right) - \delta\left(\frac{q_0}{\mu}\right)\right]. \tag{17}$$

Consider now those rays of the pencil which pass through the focus F_0. Then $z_0 = \zeta_0$, $\delta x_0 = \delta y_0 = \delta z_0 = 0$. Since all these rays also intersect the focal line f_0', $\delta p_0 = 0$. With this substitution (14) and (16) give

$$\delta p_1 = 0, \tag{14a}$$

and

$$\frac{\zeta_0}{n_0}\sec^3\theta_0\delta q_0 - \frac{1}{\mu}r_y(\delta q_1 - \delta q_0) = 0. \tag{16a}$$

Eq. (14a) shows that the corresponding refracted rays lie in the y, z-plane. Now since all the rays from F_0 pass through the focus F_1 ($z_1 = \zeta_1$), (17) must hold with $z_1 = \zeta_1$, $\delta x_1 = \delta y_1 = \delta z_1 = 0$, whatever the value of δq_0, so that

$$\frac{\zeta_1}{n_1}\sec^3\theta_1\delta q_1 - \frac{1}{\mu}r_y(\delta q_1 - \delta q_0) = 0. \tag{17a}$$

Eqs. (16a) and (17a) can be satisfied simultaneously for an arbitrary value of δq_0 only if

$$\frac{n_0\cos^3\theta_0}{\zeta_0} - \frac{n_1\cos^3\theta_1}{\zeta_1} = \frac{n_0\cos\theta_0 - n_1\cos\theta_1}{r_y}. \tag{18}$$

This relation gives the position of the focus F_1 of the refracted rays. From (14a) it is seen that the focal line through F_1 is perpendicular to the y, z-plane so that F_1 is a *primary focus*.

To find the position of the other focus, consider the rays which proceed from F_0'. Then $z_0 = \zeta_0'$, $\delta x_0 = \delta y_0 = \delta z_0 = 0$. Since all these rays intersect the focal line f_0, $\delta q_0 = \delta m_0 = 0$. Eqs. (14) and (16) now give

$$\frac{\zeta_0'}{n_0}\sec\theta_0\delta p_0 - \frac{1}{\mu}r_x(\delta p_1 - \delta p_0) = 0, \tag{14b}$$

and

$$\delta q_1 = 0. \tag{16b}$$

Eq. (16b) shows that the refracted rays now lie in the x, z-plane. All these rays will pass through the other focus $F_1'(z_1 = \zeta_1')$, so that (15) must be satisfied with $z_1 = \zeta_1'$, $\delta x_1 = \delta y_1 = \delta z_1 = 0$, whatever the value of δp_0. Hence,

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\frac{\zeta_1'}{n_1}\sec\theta_1\delta p_1 - \frac{1}{\mu}r_x(\delta p_1 - \delta p_0) = 0. \tag{15b}$$

Since (15b) and (14b) hold simultaneously for any arbitrary value of δp_0, it follows that

$$\frac{n_0\cos\theta_0}{\zeta_0'} - \frac{n_1\cos\theta_1}{\zeta_1'} = \frac{n_0\cos\theta_0 - n_1\cos\theta_1}{r_x}. \tag{19}$$

This relation gives the position of the *secondary focus* F_1'.

It is often convenient to specify the position of the foci by means of their distances from O rather than by means of their z coordinates. If $OF_0 = d_0^{(t)}$, $OF_0' = d_0^{(s)}$, $OF_1 = d_1^{(t)}$, $OF_1' = d_1^{(s)}$ (in Fig. 4.22 $d_0^{(t)} < 0$, $d_0^{(s)} < 0$, $d_1^{(t)} > 0$, $d_1^{(s)} > 0$), then

$$\left.\begin{aligned} \zeta_0 &= d_0^{(t)}\cos\theta_0, \quad & \zeta_1 &= d_1^{(t)}\cos\theta_1, \\ \zeta_0' &= d_0^{(s)}\cos\theta_0, \quad & \zeta_1{}' &= d_1^{(s)}\cos\theta_1, \end{aligned}\right\} \tag{20}$$

and the two relations (18) and (19) become

$$\frac{n_0\cos^2\theta_0}{d_0^{(t)}} - \frac{n_1\cos^2\theta_1}{d_1^{(t)}} = \frac{n_0\cos\theta_0 - n_1\cos\theta_1}{r_y}, \tag{21}$$

and

$$\frac{n_0}{d_0^{(s)}} - \frac{n_1}{d_1^{(s)}} = \frac{n_0\cos\theta_0 - n_1\cos\theta_1}{r_x}. \tag{22}$$

The corresponding relations for reflection may be obtained by setting $n_1 = -n_0$.

4.7 Chromatic aberration. Dispersion by a prism

In Chapter II it was shown that the refractive index is not a material constant but depends on colour, i.e. on the wavelength of light. We shall now discuss some elementary consequences of this result in relation to the performance of lenses and prisms.

4.7.1 Chromatic aberration

If a ray of polychromatic light is incident upon a refracting surface, it is split into a set of rays, each of which is associated with a different wavelength. In traversing an optical system, light of different wavelengths will therefore, after the first refraction, follow slightly different paths. In consequence, the image will not be sharp and the system is said to suffer from *chromatic aberration*.

We shall again confine our attention to points and rays in the immediate neighbourhood of the axis, i.e. it will be assumed that the imaging in each wavelength obeys the laws of Gaussian optics. The chromatic aberration is then said to be of the first order, or primary. If Q_α and Q_β are the images of a point P in two different wavelengths (Fig. 4.23), the projections of $Q_\alpha Q_\beta$ in the directions parallel and perpendicular to the axis are known as *longitudinal* and *lateral* chromatic aberration respectively.

Consider the change δf in the focal length of a thin lens, due to a change δn in the

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

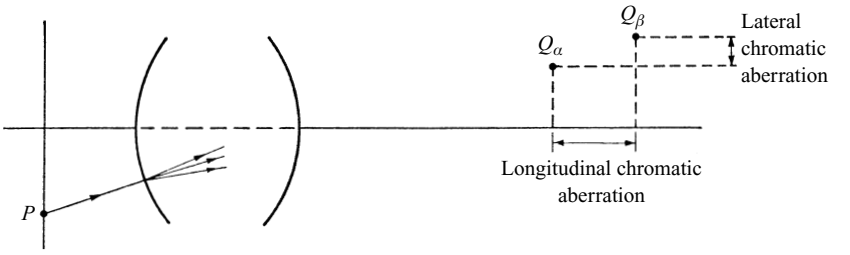

Fig. 4.23 Longitudinal and lateral chromatic aberration.

refractive index. According to §4.4 (36) the quantity $(n-1)f$ will, for a given lens, be independent of the wavelength. Hence

$$\frac{\delta f}{f}+\frac{\delta n}{n-1}=0. \tag{1}$$

The quantity

$$\Delta=\frac{n_F-n_C}{n_D-1}, \tag{2}$$

where n_F, n_D and n_C are the refractive indices for the Fraunhofer F, D and C lines ($\lambda=4861$ Å, 5893 Å, 6593 Å respectively) is a rough measure of the dispersive properties of the glass, and is called the *dispersive power*. From (1) it is seen that it is approximately equal to the distance between the red and blue images divided by the focal length of the lens, when the object is at infinity. The variation with wavelength of the refractive index of the usual types of glass employed in optical systems is shown in Fig. 4.24. The corresponding values of Δ lie between about $1/60$ and $1/30$.

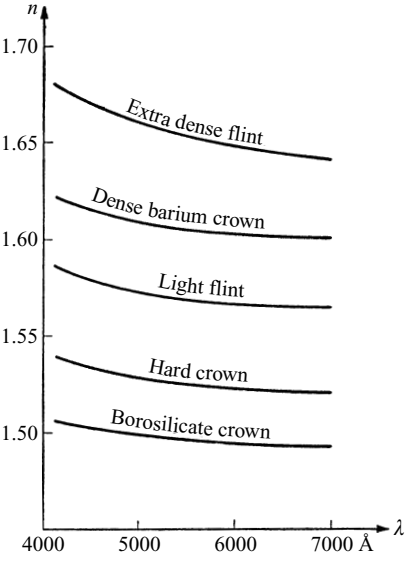

Fig. 4.24 Typical dispersion curves of various types of glass.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

To obtain an image of good quality, the monochromatic as well as the chromatic aberrations must be small. Usually a compromise has to be made, since in general it is impossible to eliminate all the aberrations simultaneously. Often it is sufficient to eliminate the chromatic aberration for two selected wavelengths only. The choice of these wavelengths will naturally depend on the purpose for which the system is designed; for example, since the ordinary photographic plate is more sensitive to the blue region than is the human eye, photographic objectives are usually 'achromatized' for colours nearer to the blue end of the spectrum than is the case in visual instruments. Achromatization with respect to two wavelengths does, of course, not secure a complete removal of the colour error. The remaining chromatic aberration is known as *the secondary spectrum*.

Let us now examine under what conditions two thin lenses will form an achromatic combination with respect to their focal lengths. According to §4.4 (39) the reciprocal of the focal length of a combination of two thin lenses separated by a distance l is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{l}{f_1 f_2}. \tag{3}$$

It is seen that $\delta f = 0$ when

$$\frac{\delta f_1}{f_1^2} + \frac{\delta f_2}{f_2^2} - \frac{l}{f_1 f_2}\left(\frac{\delta f_1}{f_1} + \frac{\delta f_2}{f_2}\right) = 0. \tag{4}$$

If the achromatization is made for the C and F lines, we have, using (1) and (2),

$$l = \frac{\Delta_1 f_2 + \Delta_2 f_1}{\Delta_1 + \Delta_2}, \tag{5}$$

where Δ_1 and Δ_2 are the dispersive powers of the two lenses.

One method of reducing the chromatic aberration is to employ two thin lenses in contact (Fig. 4.25), one made of crown glass, and the other of flint glass. In this case, since $l = 0$, we have, according to (5),

$$\frac{\Delta_1}{f_1} + \frac{\Delta_2}{f_2} = 0, \tag{6}$$

or, using (3),

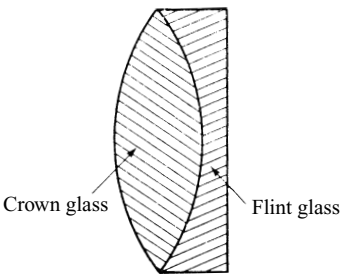

Fig. 4.25 An achromatic doublet.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\frac{1}{f_1} = \frac{1}{f}\frac{\Delta_2}{\Delta_2 - \Delta_1}, \qquad \frac{1}{f_2} = -\frac{1}{f}\frac{\Delta_1}{\Delta_2 - \Delta_1}. \tag{7}$$

Now for a given glass, and with a fixed value of the focal length f, (7) specifies f_1 and f_2 uniquely. But f_1 and f_2 depend on three radii of curvature; hence one of the radii may be chosen arbitrarily. This degree of freedom is sometimes used to make the spherical aberration as small as possible.

Another method of obtaining an achromatic system is to employ two thin lenses made of the same glass ($\Delta_1 = \Delta_2$) and separated by a distance equal to half of the sum of their focal lengths

$$l = \tfrac{1}{2}(f_1 + f_2). \tag{8}$$

That such a combination is achromatic follows immediately from (5).

An instrument consisting of several components cannot, in general, be made achromatic with respect to both position and magnification, unless each component is itself achromatic in this sense. We shall prove this for the case of two centred thin lenses separated by a distance l. Since according to §4.4.4 the imaging by a thin lens is a central projection from the centre of the lens, we have (see Fig. 4.26)

$$\frac{Y_1'}{Y_1} = -\frac{\zeta_1'}{\zeta_1}, \qquad \frac{Y_2'}{Y_2} = -\frac{\zeta_2'}{\zeta_2}. \tag{9}$$

Since $Y_2 = Y_1'$, the magnification is

$$\frac{Y_2'}{Y_1} = \frac{\zeta_1'}{\zeta_1}\frac{\zeta_2'}{\zeta_2}. \tag{10}$$

If now the wavelength is altered, ζ_1 will remain unchanged, and if the position of the image is assumed to be achromatized, ζ_2' will also remain unchanged. Hence the condition for achromatization of the magnification may be expressed by the formula

$$\delta\left(\frac{\zeta_1'}{\zeta_2}\right) = \frac{1}{\zeta_2^2}(\zeta_2\delta\zeta_1' - \zeta_1'\delta\zeta_2) = 0. \tag{11}$$

Since $\zeta_1' + \zeta_2 = l$, $\delta\zeta_1' = -\delta\zeta_2$, and it is seen that (11) can only be satisfied if $\delta\zeta_1' = \delta\zeta_2 = 0$, i.e. if each of the lenses is achromatized.

So far we have only considered the primary chromatic aberration of a thin lens and of a combination of two such lenses. Expressions for the primary chromatic aberration of a general centred system will be derived in Chapter V.

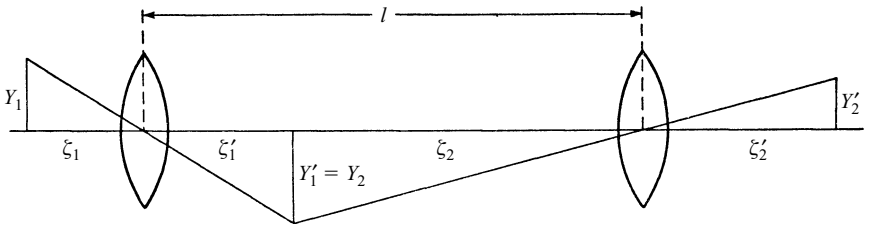

Fig. 4.26 Achromatization of two thin lenses.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

4.7.2 Dispersion by a prism

We shall now briefly discuss the passage of light through a prism.

Let α be the angle between the two faces of the prism. It is assumed that the edge A in which the two faces meet is perpendicular to the plane which contains the incident, transmitted and emergent rays (Fig. 4.27). To begin with the light will be assumed to be strictly monochromatic.

Let B_1 and B_2 be the points of intersection of the incident and the emergent ray with the two faces, ϕ_1 and ψ_1 the angles of incidence and refraction at B_1, and ψ_2 and ϕ_2 the inner and outer angles at B_2 (i.e. the angles which the ray B_1B_2 and the emergent ray make with the normal at B_2). Further let C be the point of intersection of the normals to the prism at B_1 and B_2, and D the point of intersection of the incident and the emergent rays, when these are prolonged sufficiently far.

If ε is the *angle of deviation*, i.e. the angle which the emergent ray makes with the incident ray, then

$$\phi_1 + \phi_2 = \varepsilon + \alpha, \tag{12}$$

$$\psi_1 + \psi_2 = \alpha. \tag{13}$$

Further, by the law of refraction,

$$\left.\begin{aligned} \sin\phi_1 &= n\sin\psi_1, \\ \sin\phi_2 &= n\sin\psi_2, \end{aligned}\right\} \tag{14}$$

where n is the refractive index of the glass with respect to the surrounding air. The deviation ε will have an extremum when

$$\frac{\mathrm{d}\varepsilon}{\mathrm{d}\phi_1} = 0. \tag{15}$$

Using (12) this implies that

$$\left(\frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1}\right)_{\text{extr.}} = -1. \tag{16}$$

Now we have from (13) and (14),

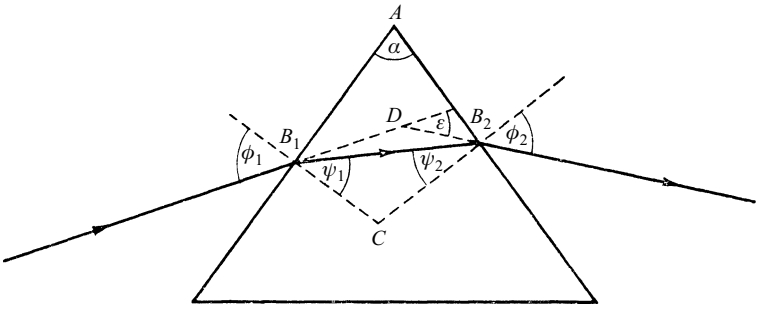

Fig. 4.27 Passage of a ray through a prism.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\left.\begin{aligned} \frac{\mathrm{d}\psi_1}{\mathrm{d}\phi_1} &= -\frac{\mathrm{d}\psi_2}{\mathrm{d}\phi_1}, \\ \cos\phi_1 &= n\cos\psi_1\frac{\mathrm{d}\psi_1}{\mathrm{d}\phi_1}, \\ \cos\phi_2\frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1} &= n\cos\psi_2\frac{\mathrm{d}\psi_2}{\mathrm{d}\phi_1}, \end{aligned}\right\} \tag{17}$$

and hence, on elimination

$$\frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1} = -\frac{\cos\phi_1\cos\psi_2}{\cos\psi_1\cos\phi_2}. \tag{18}$$

From (16) and (18) it follows that, for an extremum,

$$\frac{\cos\phi_1\cos\psi_2}{\cos\psi_1\cos\phi_2} = 1, \tag{19}$$

whence, on squaring and using (14)

$$\frac{1-\sin^2\phi_1}{n^2-\sin^2\phi_1} = \frac{1-\sin^2\phi_2}{n^2-\sin^2\phi_2}. \tag{20}$$

This equation is satisfied by

then

$$\left.\begin{aligned} \phi_1 &= \phi_2; \\ \psi_1 &= \psi_2. \end{aligned}\right\} \tag{21}$$

To determine the nature of the extremum we must evaluate $\mathrm{d}^2\varepsilon/\mathrm{d}{\phi_1}^2$. From (12) and (18),

$$\begin{aligned} \frac{\mathrm{d}^2\varepsilon}{\mathrm{d}{\phi_1}^2} &= \frac{\mathrm{d}^2\phi_2}{\mathrm{d}{\phi_1}^2} = \frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1}\frac{\mathrm{d}}{\mathrm{d}\phi_1}\left[\ln\left(-\frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1}\right)\right] \\ &= \frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1}\left[-\tan\phi_1 - \tan\psi_2\frac{\mathrm{d}\psi_2}{\mathrm{d}\phi_1} + \tan\psi_1\frac{\mathrm{d}\psi_1}{\mathrm{d}\phi_1} + \tan\phi_2\frac{\mathrm{d}\phi_2}{\mathrm{d}\phi_1}\right]. \end{aligned} \tag{22}$$

When $\phi_1 = \phi_2$, $\psi_1 = \psi_2$, this becomes with the help of (16), (17) and (14),

$$\left(\frac{\mathrm{d}^2\varepsilon}{\mathrm{d}{\phi_1}^2}\right)_{\text{extr.}} = 2\tan\phi_1 - 2\tan\psi_1\frac{\cos\phi_1}{n\cos\psi_1} = 2\tan\phi_1\left(1-\frac{\tan^2\psi_1}{\tan^2\phi_1}\right). \tag{23}$$

Since $n>1$, $\phi_1>\psi_1$; also since $0<\phi_1<\pi/2$, $\tan\phi_1>0$. Hence $(\mathrm{d}^2\varepsilon/\mathrm{d}{\phi_1}^2)>0$, so that *the deviation is a minimum*. According to (21) it takes place when the passage of the rays through the prism is *symmetrical*. The minimal value of the deviation then is

$$\varepsilon_{\min} = 2\phi_1 - \alpha. \tag{24}$$

In terms of $\varepsilon_{\min}$ and α, the angle of incidence and the angle of refraction at the first face of the prism are

$$\phi_1 = \tfrac{1}{2}(\varepsilon_{\min}+\alpha), \qquad \psi_1 = \tfrac{1}{2}\alpha, \tag{25}$$

so that

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$n = \frac{\sin \phi_1}{\sin \psi_1} = \frac{\sin\left[\frac{1}{2}(\varepsilon_{\min} + \alpha)\right]}{\sin\left(\frac{1}{2}\alpha\right)}. \tag{26}$$

This formula is often used in determinations of the refractive index of glass. One measures $\varepsilon_{\min}$ and α with the help of a spectrometer and evaluates n from (26).

Instead of a single ray, let us now consider the passage of a pencil of parallel rays through the prism, for example from a point source P placed in the focal plane of a lens L_1 (Fig. 4.28), the light still being assumed to be monochromatic. Let B_1' and B_2' be the feet of the perpendiculars dropped from B_1 and B_2 on to the rays which pass through the edge A. Then B_1B_1' and B_2B_2' are the lines of intersection of two wave-fronts with the plane of incidence (plane of the figure). These two lines are inclined to each other at the angle of deviation ε. We set

$$B_1B_1' = l_1, \qquad B_2B_2' = l_2', \qquad B_1B_2 = t. \tag{27}$$

Consider now a parallel beam of polychromatic instead of monochromatic light. If the lens L_1 is corrected for chromatic aberration, B_1B_1' will still be in the wave-front of the incident pencil. On the other hand the line B_2B_2' will no longer be unique, but will depend on the wavelength λ. For the refractive index of the prism is a function of the wavelength,

$$n = n(\lambda), \tag{28}$$

and consequently the deviation ε also depends on λ:

$$\varepsilon = \varepsilon(\lambda). \tag{29}$$

The quantity

$$\frac{\mathrm{d}\varepsilon}{\mathrm{d}\lambda} = \frac{\mathrm{d}\varepsilon}{\mathrm{d}n}\frac{\mathrm{d}n}{\mathrm{d}\lambda} \tag{30}$$

formed for a constant value of the angle of incidence ϕ_1 is often called the *angular dispersion of the prism*. In (30) the first factor on the right depends entirely on the geometry of the arrangement, whilst the second factor characterizes the dispersive

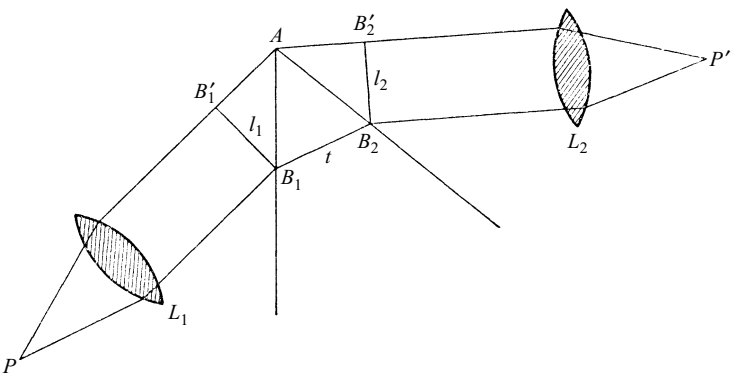

Fig. 4.28 Dispersion by a prism.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

power of the glass of which the prism is made. Since $\phi_1 =$ constant, we have from (12) and (13)

$$\frac{\mathrm{d}\varepsilon}{\mathrm{d}n} = \frac{\mathrm{d}\phi_2}{\mathrm{d}n}, \qquad \frac{\mathrm{d}\psi_1}{\mathrm{d}n} = -\frac{\mathrm{d}\psi_2}{\mathrm{d}n}, \tag{31}$$

and from (14)

$$\sin\psi_1 + n\cos\psi_1 \frac{\mathrm{d}\psi_1}{\mathrm{d}n} = 0,$$

$$\cos\phi_2 \frac{\mathrm{d}\phi_2}{\mathrm{d}n} = \sin\psi_2 + n\cos\psi_2 \frac{\mathrm{d}\psi_2}{\mathrm{d}n}, \tag{32}$$

whence on elimination

$$\frac{\mathrm{d}\varepsilon}{\mathrm{d}n} = \frac{\sin(\psi_1+\psi_2)}{\cos\phi_2\cos\psi_1} = \frac{\sin\alpha}{\cos\phi_2\cos\psi_1}. \tag{33}$$

From the triangle AB_1B_2, we have, by the sine theorem,

$$AB_2 = \frac{\cos\psi_1}{\sin\alpha}\, t, \tag{34}$$

and from the triangle AB_2B_2',

$$AB_2 = l_2 \sec\phi_2. \tag{35}$$

Using (34) and (35), (33) gives

$$\frac{\mathrm{d}\varepsilon}{\mathrm{d}\lambda} = \frac{t}{l_2}\frac{\mathrm{d}n}{\mathrm{d}\lambda}. \tag{36}$$

In the position of minimum deviation, we have by symmetry $l_1 = l_2$. If, moreover, the lenses are so large that the pencil completely fills the prism, then t will be equal to the length b of the base of the prism. Eq. (36) then gives for the *angular dispersion* $\delta\varepsilon$, i.e. for the angle by which the emergent wave-front is rotated when the wavelength is changed from λ to $\lambda + \delta\lambda$, the following expression:

$$\delta\varepsilon = \frac{b}{l_1}\frac{\mathrm{d}n}{\mathrm{d}\lambda}\delta\lambda. \tag{37}$$

4.8 Radiometry and apertures

The branch of optics concerned with the measurement of light is called *radiometry*. Strictly, radiometry is not part of geometrical optics, but it seems appropriate to include a short account of it in the present chapter, since in many practical applications the approximate geometrical picture of an optical field forms an adequate basis for radiometric investigations. We shall, therefore, take over the geometrical model according to which light is regarded as the flow of luminous energy along the geometrical rays, subject to the geometrical law of conservation of energy. This states [see §3.1 (31)] that the energy which is transmitted in unit time through any cross-section of a tube of rays is constant.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

4.8.1 Basic concepts of radiometry[*]

In radiometry we are concerned essentially with the light energy emerging from a portion of a surface S. This surface may be fictitious, or it may coincide with the actual radiating surface of a source, or with an illuminated surface of a solid. If the latter is opaque, it is the reflected light which is considered; if it is transparent or semitransparent (in which case the light is partly absorbed or scattered), it is the transmitted light which is usually measured.

Let $P(\xi, \eta)$ be a typical point on S referred to any convenient set of curvilinear coordinates on the surface. The (time averaged) amount of energy which emerges per unit time from the element δS of the surface at P and which falls within an element $\delta\Omega$ of the solid angle around a direction specified by the polar angles (α, β), may sometimes be expressed in the form

$$\delta F = B \cos\theta \delta S \delta\Omega, \tag{1}$$

which is a generalization of §3.1 (54)†. Here θ is the angle which the direction (α, β) makes with the normal to the surface element (see Fig. 4.29), and B is a factor which in general depends on (ξ, η) and (α, β),

$$B = B(\xi, \eta; \alpha, \beta). \tag{2}$$

The factor $\cos\theta$ is introduced in (1) since it is the projection of δS on to a plane normal to the direction (α, β), rather than δS itself, which is the physically significant quantity. B is called the *brightness*, also sometimes called the *radiance* or the *specific*

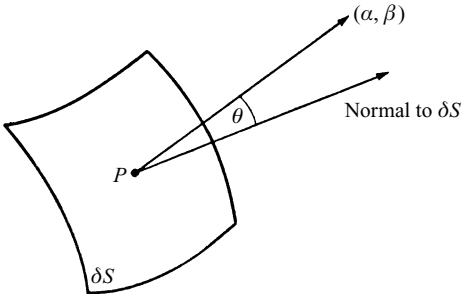

Fig. 4.29 Energy transfer from a surface element.

[*] Only the fundamental notions of radiometry will be discussed. For further information and for the description of instruments used for measuring light see, for example, J. W. T. Walsh, *Photometry* (London, Constable, 2nd edition, 1953) or W. E. Forsythe (ed.), *Measurement of Radiant Energy* (New York and London, McGraw-Hill, 1937) or R. W. Boyd, *Radiometry and the Detection of Optical Radiation* (New York, J. Wiley and Sons, 1983).

† As already mentioned in §3.1 in the transition from §3.1 (52) to §3.1 (54), the expression for the energy flux δF does not hold, in general, even for fields of very short wavelengths. For a discussion of this point and of the difficulties of providing a rigorous basis for radiometry see E. Wolf, *J. Opt. Soc. Amer.* **68** (1978), 6. Many papers concerned with the foundations of radiometry are reprinted in A. T. Friberg (ed.) *Selected Papers on Coherence and Radiometry* (Bellingham, WA, SPIE Optical Engineering Press, 1993). The foundations of radiometry under somewhat restricted conditions are discussed in Sec. 5.7 of L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge, Cambridge University Press, 1995).

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

intensity, at the point (ξ, η) in the direction (α, β). It must be distinguished from the visual sensation of brightness, from which it will in general differ, because the eye is not equally sensitive to all colours;* this point will be discussed more fully later.

δF is usually decomposed in two different ways, to show the explicit dependence on $\delta\Omega$ and δS:

$$\delta F = \delta I \delta\Omega = \delta E \delta S. \tag{3}$$

Comparison of (1) and (3) gives

$$\delta I = \frac{\delta F}{\delta\Omega} = B \cos\theta \delta S, \tag{4}$$

$$\delta E = \frac{\delta F}{\delta S} = B \cos\theta \delta\Omega. \tag{5}$$

The integral

$$I(\alpha, \beta) = \int B \cos\theta \, \mathrm{d}S \tag{6}$$

taken over a piece of surface is called the *radiant intensity*† *in the direction* (α, β), and the integral

$$E(\xi, \eta) = \int B \cos\theta \, \mathrm{d}\Omega \tag{7}$$

taken throughout a solid angle is called the *radiometric illumination at the point* (ξ, η).

The variation of B with direction will depend on the nature of the surface, especially on whether it is rough or smooth, whether it is self-luminous, or whether it transmits or reflects other light. Often it is permissible to assume that, to a good approximation, B is independent of the direction. The radiation is then said to be *isotropic*. If the radiation is isotropic and if the radiating surface is plane, (6) reduces to

$$I(\alpha, \beta) = I_0 \cos\theta, \tag{8}$$

where

$$I_0 = \int B \, \mathrm{d}S.$$

The photometric intensity in any direction then varies as the cosine of the angle between that direction and the normal to the surface. Eq. (8) is known as *Lambert's* (*cosine*) *law*, and when satisfied one speaks of *diffuse emission* or *diffuse reflection*, according to whether the surface is emitting or reflecting.

The measurement of the quantities F, B, I and E involves the determination of a

* One often uses the adjective 'photometric' when one wishes to stress that a particular quantity is evaluated with regard to its visual, rather than its true physical effects.

† In Chapter I the light intensity was defined as the time average of the amount of energy which crosses per second a unit area perpendicular to the direction of the flow. This quantity must not be confused with the *radiant* intensity as defined by (6). It is unfortunate that the same word is used to denote two different quantities. Except in the present section we shall always understand by 'intensity' the quantity introduced in Chapter I. If the surface element δS at P is orthogonal to the Poynting vector, the intensity (in the sense of Chapter I) is equal to the illumination δE at P.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

time interval, an area, a direction, a solid angle and an energy. The averages involved are often small and consequently sensitive instruments have to be used. They are essentially of two kinds. First, those which react to the heat developed in an absorbing medium (e.g. bolometer, thermo-couple, etc.) and are mainly used in studying heat radiation (infra-red); secondly, those which are based on the (surface) photoelectric effect, i.e. on the phenomenon of emission of electrons from a metal, caused by the incidence of light on the surface of the metal (in monochromatic light the number of emitted electrons, i.e. the current produced, is proportional to the energy of the incident light). Instruments of this kind are used, for instance, as exposure meters in photography.

In technical radiometry, however, indirect methods are used. One defines and constructs a standard source of light, and expresses its photometric data in absolute energy units. Measurements are then made relative to this source, often using the eye as a null indicator, namely as an indicator of equal brightness. The comparison is based on a simple law which holds for the illumination due to a very small source* Q:

Let δS be a surface element at P and let $QP = r$. If θ is the angle which QP makes with the normal to δS (see Fig. 4.30), then the energy which the source sends through δS per unit time is $I\delta\Omega$, where I is the radiant intensity of the source in the direction QP, and $\delta\Omega$ is the solid angle which δS subtends at Q. Now by elementary geometry,

$$\cos\theta\delta S = r^2\delta\Omega. \tag{9}$$

Hence, using (3), we have

$$E = \frac{I\cos\theta}{r^2}. \tag{10}$$

Eq. (10) is the basic equation of practical radiometry. It expresses the so-called *cosine law of illumination* (E proportional to $\cos\theta$) and the *inverse square law* (E inversely proportional to r^2), and enables a comparison of the intensities of sources with the help of simple geometry. If a surface element δS is illuminated by two point sources Q_1 and Q_2 of intensities I_1 and I_2, and if the lines connecting δS with the sources make angles θ_1 and θ_2 with the normal to δS (Fig. 4.31) then, when the illuminations are equal, we have

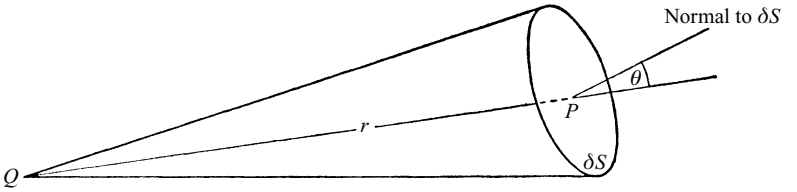

Fig. 4.30 Illumination from a point source.

* The radiant intensity I of a point source is defined by a procedure often used in connection with limiting concepts. Let us suppose that the area of the surface decreases towards zero, while at the same time B increases towards infinity, in such a way that the integral (6) remains finite. Then I is a function of α and β and of the position of the point source.

In calculating the radiometric illumination, the finite extension of the source is usually neglected when its linear dimensions are less than about $1/20$ of its distance from the illuminated surface.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

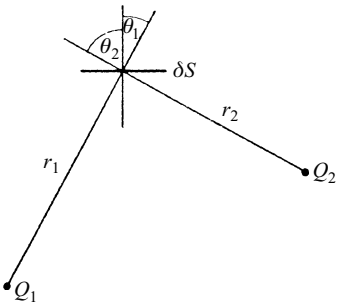

Fig. 4.31 Comparison of the intensities of two point sources.

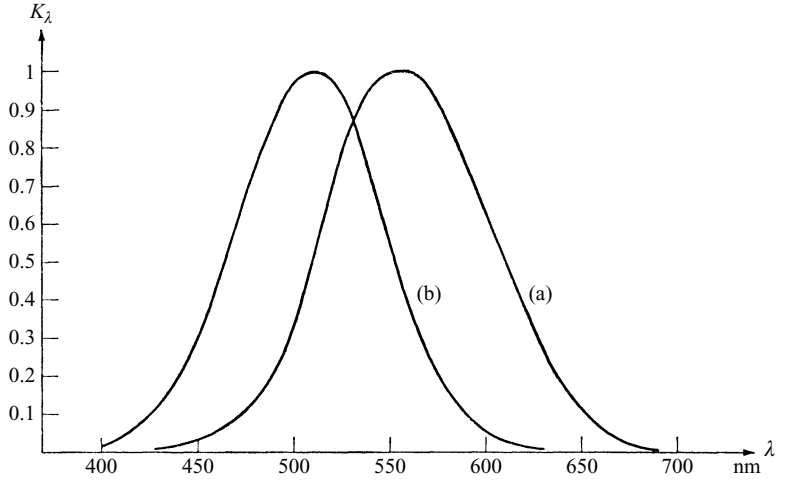

Fig. 4.32 The relative visibility curve K_λ of the average human eye: (a) for bright light and (b) for feeble light.

$$I_1 : I_2 = \frac{r_1^2}{\cos\theta_1} : \frac{r_2^2}{\cos\theta_2}. \tag{11}$$

With the help of (11), the strength of a given source may be determined by a comparison with a standard source.

The equality of the illumination produced by two sources can be detected either by physical means, or directly, by the eye. Comparison by the eye is relatively easy when the light from the two sources is monochromatic and of the same frequency, but in general one has to compare sources which emit light of different spectral compositions. This simple procedure can then no longer be used, since the eye is not equally sensitive to light of all wavelengths. Filters, each of which transmits light in a narrow range around a known wavelength of the spectrum, must be used instead; the determination of equality of illumination for different colours is then reduced to the determination of the relative energy values, taking into account the *relative visibility curve* of the eye. This is the curve obtained by plotting the reciprocal K_λ of that amount of flux which produces the same sensation of brightness, against the wavelength. The curve depends to some extent on the strength of the illumination. For bright light it has a maximum near 550 nm. (See Fig. 4.32 and Table 4.2.) With

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

Table 4.2. *The relative visibility factor K_λ of an average human eye (bright light)*

λ (in nm)	K_λ	λ (in nm)	K_λ
400	0.0004	600	0.631
410	0.0012	610	0.503
420	0.0040	620	0.381
430	0.0116	630	0.265
440	0.023	640	0.175
450	0.038	650	0.107
460	0.060	660	0.061
470	0.091	670	0.032
480	0.139	680	0.017
490	0.208	690	0.0082
500	0.323	700	0.0041
510	0.503	710	0.0021
520	0.710	720	0.00105
530	0.862	730	0.00052
540	0.954	740	0.00025
550	0.995	750	0.00012
560	0.995	760	0.00006
570	0.952		
580	0.870		
590	0.757		

decreasing strength of illumination the curve retains its shape but the maximum shifts towards the blue end of the spectrum, being at about 507 nm for very faint light. This phenomenon is known as the *Purkinje effect*.

If the flux of energy is evaluated with respect to the visual sensation which it produces, rather than with regard to its true physical magnitude, one speaks of *the luminous energy F'*:

$$F' = \frac{\int K_\lambda F_\lambda \, \mathrm{d}\lambda}{\int K_\lambda \, \mathrm{d}\lambda}, \tag{12}$$

$F_\lambda \, \mathrm{d}\lambda$ being the amount of energy in the range $(\lambda, \lambda + \mathrm{d}\lambda)$, K_λ the relative visibility and the integration being taken throughout the spectral range. The quantities B', I' and E', which bear the same relations to B, I and E as F' bears to F, are usually called the *luminance*, the *luminous intensity* (or *candle power*) and the *illumination* respectively. These concepts are extensively used in photometry.

There are practical units for each of the four quantities F', B', I' and E'. Since it is easier to maintain a standard of luminous intensity rather than of luminous flux, the unit of luminous intensity is usually considered to be the basic photometric unit, and those for F', B' and E' are expressed in terms of it. The adopted standard of luminous intensity was at one time the *international candle*, a standard preserved by a number

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

of carbon lamps kept at various national laboratories. In more recent times it has been replaced by a new standard called the *candela* (cd); this is defined as one-sixtieth of the luminous intensity per square centimetre of a black body radiator at the temperature of solidification of platinum (2042 K approx.). The value of the luminous intensity of light of a different spectral composition must be evaluated by the procedure already explained, taking into account the relative visibility curve.

The unit of luminous flux is called the *lumen*. It is the luminous flux emitted within a unit solid angle by a uniform point source of luminous intensity 1 candela.

The unit of illumination depends on the unit of length employed. The metric unit of illumination is the *lux* (lx), sometimes also called the *metre-candle*; it is the illumination of a surface area of one square metre receiving a luminous flux of one lumen. The old British unit is the lumen per square foot, formerly called the *foot-candle* (f.c.).

The unit of luminance is the candela per square centimetre, termed the *stilb* (sb), and the candela per square metre termed the *nit*. In the old British system, the units used are the candela per square inch or candela per square foot.

Other units are also used, but for their definitions and their relation to the units here discussed we must refer to books on radiometry and photometry.

4.8.2 Stops and pupils[*]

The amount of light which reaches the image space of an optical system depends not only on the brightness of the object but also on the dimensions of the optical elements (lenses, mirrors) and of the stops. A stop (or diaphragm) is an opening, usually circular, in an opaque screen. The opaque parts of the screen prevent some of the severely aberrated rays from reaching the image. For the purposes of the present discussion it will be convenient to include the edges of the lenses and mirrors as well as diaphragms in the term 'stop'.

Consider all the rays from the axial object point P_0. The stop which determines the cross-section of the image-forming pencil is called the *aperture stop*. To determine its position, the Gaussian image of each stop must be found in the part of the system which precedes it; the image which subtends the smallest angle at P_0 is called the *entrance pupil*. The physical stop which gives rise to the entrance pupil is the aperture stop. (If it lies in front of the first surface it is identical with the entrance pupil.) The angle $2\theta_0$ which the diameter of the entrance pupil subtends at P_0 is called the *angular aperture on the object side*, or simply *angular aperture* (Fig. 4.33).

The image of the aperture stop formed by the part of the system which follows it (also the image of the entry pupil by the whole system) is known as the *exit pupil*; the angle $2\theta_1$ which its diameter subtends at the image P_1 may be called the *angular aperture on the image side* (sometimes also the *projection angle*).

In the pencil of rays which proceeds from each object point, there will be a ray which passes through the centre of the entrance pupil. This special ray is known as the *principal ray* (also the *chief* or the *reference ray*) of the pencil, and is of particular importance in the theory of aberrations. In the absence of aberrations, the principal ray

[*] The theory of stops was formulated by E. Abbe, *Jena Z. Naturwiss.*, **6** (1871), 263, and was extended by M. von Rohr, *Zentr. Ztg. Opt. Mech.*, **41** (1920), 145, 159, 171.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

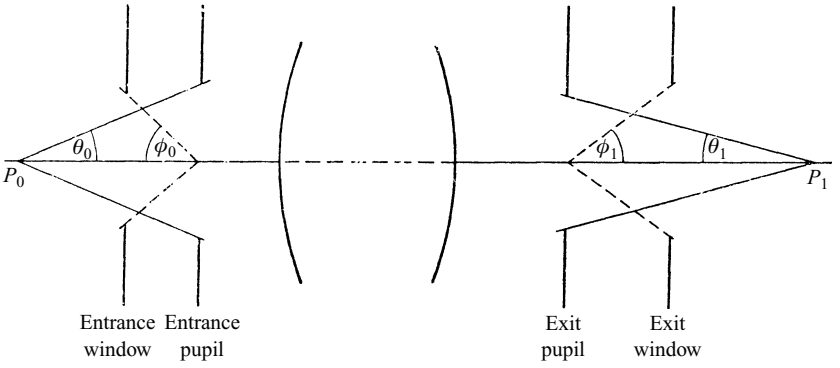

Fig. 4.33 Stops and pupils.

will also pass through the centre of the aperture stop and through the centre of the exit pupil.

If the aperture stop is situated in the back focal plane of that part of the system which precedes it, then the entrance pupil will be at infinity and all the principal rays in the object space will be parallel to the axis. Such a system is said to be *telecentric on the object side*. If the aperture stop is in the front focal plane of the part of the system which follows it, then the exit pupil will be at infinity and the principal rays in the image space will be parallel to the axis; the system is then said to be *telecentric on the image side*. Telecentric arrangements are useful in measurements of the size of the object.

If other parameters are kept fixed, the angular aperture is a measure of the amount of light which traverses the system. There are other quantities which are frequently used to specify the 'light-gathering power' of an optical system, for example the *numerical aperture* (NA) of a microscope objective. This is defined as the sine of the angular semiaperture in the object space multiplied by the refractive index of the object space,

$$\mathrm{NA} = n \sin \theta_0. \tag{13}$$

When a system is designed to work with objects at a great distance, as in the case of telescopes or certain photographic lenses, a convenient measure of its light-gathering power is the so-called '*F number*' or '*nominal focal ratio*'. It is the ratio of the focal length f of the system to the diameter d of the entrance pupil:

$$F = f/d. \tag{14}$$

Thus for a lens of 10 cm focal length which works with an aperture of 2 cm, $F = 5$. It is called an $f/5$ lens and is said to work at a 'speed' of $f/5$.

Quantities such as the angular aperture, the numerical aperture, and the F number, which may be taken as the measure of the light-gathering power of the instrument, are often called *relative apertures*.

In addition to aperture stops, optical systems also possess *field stops*; they determine what proportion of the surface of an extended object is imaged by the instrument. The distinction between the two types of stops is illustrated in Fig. 4.34.

To determine the field stop, we again find first the image of each stop in the part of

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

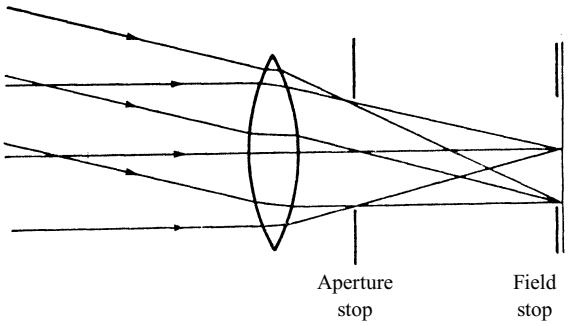

Fig. 4.34 Illustrating the distinction between the aperture stop and a field stop.

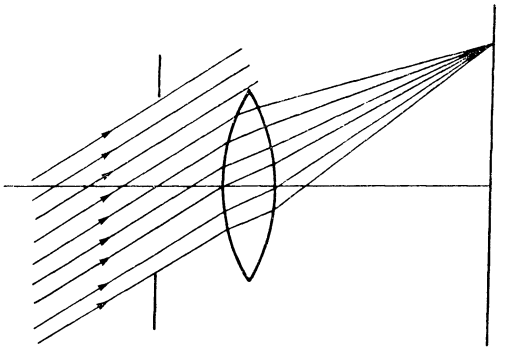

Fig. 4.35 Vignetting.

the system which precedes it. That image which subtends the smallest angle ($2\phi_0$ in Fig. 4.33) at the centre of the entrance pupil is called the *entrance window*, and the angle $2\phi_0$ is called the *field angle* or *the angular field of view*. The physical stop which corresponds to the entrance window is then the required field stop.

The image of the entrance window by the instrument (that is also the image of the field stop by the part of the system which follows it) is called the *exit window*. The angle ($2\phi_1$ in Fig. 4.33) which the diameter of the exit window subtends at the centre of the exit pupil is sometimes called the *image field angle*.

It may happen that, although the aperture stop is smaller than the lenses, some of the rays miss part of a lens entirely and parts of a lens may receive no light at all from certain regions of the object. This effect is known as *vignetting*, and is illustrated in Fig. 4.35. It is seldom encountered in systems like telescopes which have a relatively small field of view, but is of importance in other instruments, such as photographic objectives. Designers sometimes rely on vignetting to obliterate undesirable off-axis aberrations.

4.8.3 Brightness and illumination of images

We shall now briefly consider the relations between the basic radiometric quantities which characterize the radiation in the image and object space.

Assume that the object is a small plane element of area δS_0, perpendicular to the

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

axis and radiating in accordance with Lambert's law. The brightness B_0 is then independent of direction. The amount of energy δF_0 which falls per unit time on to an annular element of the entry pupil centred on the axis is

$$\delta F_0 = B_0 \cos\gamma_0 \delta S_0 \delta\Omega_0, \tag{15}$$

where

$$\delta\Omega_0 = 2\pi \sin\gamma_0 \delta\gamma_0, \tag{16}$$

γ_0 being the angle which a typical ray passing through the annulus makes with the axis. Hence if θ_0 denotes, as before, the angular semiaperture on the object side, the total flux of energy which falls on to the entrance pupil per unit time is

$$\begin{aligned} F_0 &= 2\pi B_0 \delta S_0 \int_0^{\theta_0} \sin\gamma_0 \cos\gamma_0 \, \mathrm{d}\gamma_0 \\ &= \pi B_0 \delta S_0 \sin^2\theta_0. \end{aligned} \tag{17}$$

The energy flux F_1 emerging from the exit pupil may be expressed in a similar form:

$$F_1 = \pi B_1 \delta S_1 \sin^2\theta_1. \tag{18}$$

F_1 cannot exceed F_0 and can only be equal to it if there are no losses due to reflection, absorption or scattering within the system; hence

$$B_1 \sin^2\theta_1 \delta S_1 \leqslant B_0 \sin^2\theta_0 \delta S_0. \tag{19}$$

Now the ratio $\delta S_1/\delta S_0$ is equal to the square of the lateral magnification M:

$$\frac{\delta S_1}{\delta S_0} = M^2. \tag{20}$$

If further it is assumed that the system obeys the sine condition (§4.5, (6)),

$$\frac{n_0 \sin\theta_0}{n_1 \sin\theta_1} = M. \tag{21}$$

On substituting from (20) and (21) into (19) it follows that

$$B_1 \leqslant \left(\frac{n_1}{n_0}\right)^2 B_0. \tag{22}$$

In particular, *if the refractive indices of the object and image spaces are equal*, then according to (22), *the brightness of the image cannot exceed that of the object, and can only be equal to it if the losses of light within the system are negligible*.

From (18) and (22) it follows, assuming the losses to be negligible, that

$$F_1 = \pi \left(\frac{n_1}{n_0}\right)^2 B_0 \delta S_1 \sin^2\theta_1, \tag{23}$$

so that the radiometric illumination $E_1 = F_1/\delta S_1$ at the axial point P_1 of the image is

$$E_1 = \pi \left(\frac{n_1}{n_0}\right)^2 B_0 \sin^2\theta_1. \tag{24}$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

If θ_1 is small, the solid angle Ω_1, which the exit pupil subtends at P_1, is approximately equal to $\pi \sin^2 \theta_1$, so that (24) may then be written as

$$E_1 = \left(\frac{n_1}{n_0}\right)^2 B_0 \Omega_1. \tag{25}$$

This relation applies to the axial image, but the off-axis image may be treated in a similar way. If ϕ_1 is the angle which the principal ray CQ_1 makes with the axis (see Fig. 4.36), then we have in place of (25),

$$E_1 = \left(\frac{n_1}{n_0}\right)^2 B_0 \Omega_1' \cos \phi_1, \tag{26}$$

Ω_1' being the solid angle which the exit pupil subtends at Q_1. With the help of (9) it follows that

$$\frac{\Omega_1'}{\Omega_1} = \cos \phi_1 \left(\frac{CP_1}{CQ_1}\right)^2 = \cos^3 \phi_1, \tag{27}$$

so that

$$E_1 = \left(\frac{n_1}{n_0}\right)^2 B_0 \Omega_1 \cos^4 \phi_1. \tag{28}$$

This formula shows that *the illumination in the image decreases as the fourth power of the cosine of the angle which the principal ray through the image point makes with the axis*, it being assumed that the object radiates according to Lambert's law, that there are no losses of light within the system and that the angular semiaperture θ_1 is small.

In applying the preceding formulae, it should be remembered that they have been derived with the help of the laws of geometrical optics. For a very small source or a source which is not highly incoherent they may no longer give a good approximation. For example, the image of a point source is not a point but a bright disc surrounded by rings (the Airy pattern, see §8.5.2); the light is then distributed over the whole diffraction pattern, and consequently the illumination at the geometrical focus is smaller than that given by (24).

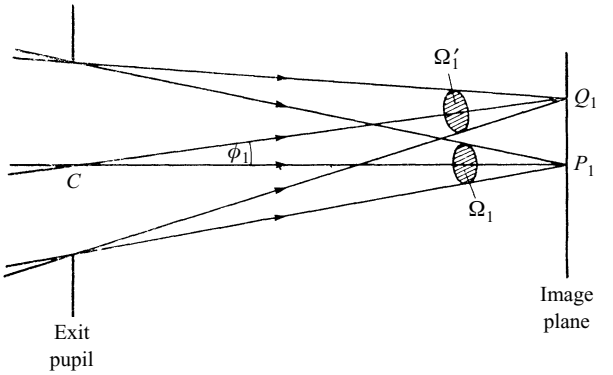

Fig. 4.36 Illumination at an off-axis image point.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

4.9 Ray tracing*

In designing optical instruments it is necessary to determine the path of the light with a greater accuracy than that given by Gaussian optics. This may be done by algebraic analysis, taking into account the higher-order terms in the expansion of the characteristic function, a procedure discussed in the next chapter. Alternatively one may determine the path of the light rays accurately with the help of elementary geometry, by successive application of the law of refraction (or reflection); this method, which will now be briefly described, is known as *ray tracing* and is extensively employed in practice.

4.9.1 Oblique meridional rays†

We consider first the tracing of an oblique meridional ray, i.e. a meridional ray from an extra-axial object point. Let A be the pole of the first surface of the system. The surface will be assumed to be a spherical refracting surface of radius r centred at a point C, and separating media of refractive indices n and n'. An incident ray OP (see Fig. 4.37) in the meridional plane is specified by the angle U which it makes with the axis, and by the distance $L = AB$ between the pole of A and the point B at which it meets the axis. Let I be the angle between the incident ray and the normal PC. The corresponding quantities relating to the refracted ray are denoted by primed symbols.

The following sign convention is used: The quantities r, L and L' are taken to be positive when C, B and B' are to the right of A, the light being assumed to be incident

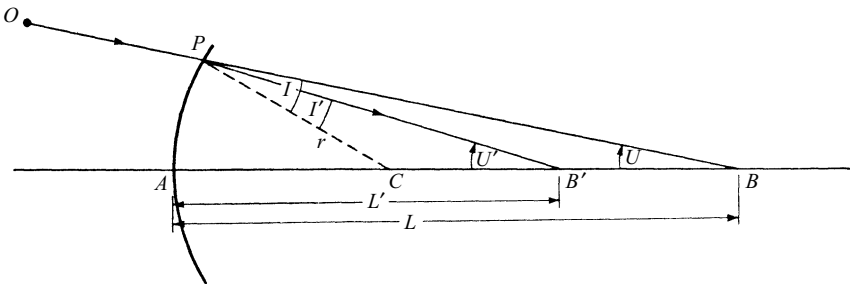

Fig. 4.37 Notation used in an oblique meridional ray trace.

* For a fuller discussion of ray tracing, see, for example, A. E. Conrady, *Applied Optics and Optical Design*, Part I (Oxford, Oxford University Press, 1929; reprinted by Dover Publications, Inc., New York, 1957); M. von Rohr, *The Geometrical Investigation of Formation of Images in Optical Systems*, translated from German by R. Kanthack (London, HM Stationery Office, 1920), and M. Herzberger, *Modern Geometrical Optics* (New York, Interscience Publishers, 1958). A method for the tracing of rays with the help of electronic computing machines was described by G. Black, *Proc. Phys. Soc.* B, **68** (1954), 569. A method for the tracing of rays through nonspherical surfaces was proposed by T. Smith, *Proc. Phys. Soc.*, **57** (1945), 286; see also W. Weinstein, *Proc. Phys. Soc.* B, **65** (1952), 731.

† In §4.9.1, §4.9.2, and §4.10 the notation and the sign convention usually employed in a meridional ray trace are used. The sign convention for angles differs from the Cartesian sign convention used throughout the rest of the book; to revert to the Cartesian sign convention set $U = -\gamma$, $U' = -\gamma'$.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

from the left. The angles U and U' are considered to be positive if the axis can be brought into coincidence with the rays PB and PB' by a clockwise rotation of less than 90° about B or B' respectively. The angles I and I' are taken to be positive if the incident and the refracted rays may be made to coincide with the normal PC by a clockwise rotation of less than 90° about the point P of incidence.

The quantities L and U which specify the incident ray may be assumed to be known. It is then necessary to calculate L' and U'. Assuming also for the moment that both L and r are finite, we have, from the triangle PCB:

$$\sin I = \frac{L - r}{r} \sin U. \tag{1}$$

By the law of refraction

$$\sin I' = \frac{n}{n'} \sin I. \tag{2}$$

Also from the figure

$$U' = U + I - I'. \tag{3}$$

Finally, from the triangle PCB',

$$L' = \frac{\sin I'}{\sin U'} r + r. \tag{4}$$

By successive application of the *refraction equations* (1)–(4), the quantities L' and U', which specify the refracted ray PB', are obtained.

The refracted ray PB' now becomes the incident ray for the second surface. Writing L_1' in place of L' and U_1' in place of U', and denoting by L_2, U_2 the corresponding values, with L_2 referred to the pole of the second surface, we have the *transfer equations*

$$L_2 = L_1' - d, \tag{5}$$

$$U_2 = U_1', \tag{6}$$

where $d > 0$ is the distance between the poles of the two surfaces.

Next the 'incident values' given by (5) and (6) are substituted into the refraction equations (1)–(4). Solving for the primed quantities, the ray is then traced through the second surface. In this way, by the repeated application of the refraction and the transfer equations, the values L' and U', relating to the ray in the image space, are obtained. The point of intersection of this ray with the image plane may then be determined. In practice, one would naturally trace not a single ray, but a number of suitably selected rays through the system; their intersection points with the image plane then give a rough estimate of the performance of the system.

If one of the surfaces (say the kth) is a mirror, the appropriate formulae to be used may be formally deduced from the preceding ones by setting $n_k' = -n_k$. Then d_k must be considered to be negative. Moreover, all the remaining refractive indices and the subsequent d values must also be considered to be negative, unless a second reflection takes place, when they revert to positive signs.

Next consider the two special cases which were so far excluded. If the incident ray is parallel to the axis ($L = \infty$) the equation

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$\sin I = \frac{Y}{r}, \tag{7}$$

is used in place of (1), where Y is the distance of the ray from the axis [see Fig. 4.38(a)].

If the surface is plane ($r = \infty$) we have, in place of (1)–(4), the following set of equations [see Fig. 4.38(b)]:

$$I = -U, \tag{8}$$

$$\sin U' = \frac{n}{n'} \sin U, \tag{9}$$

$$I' = -U', \tag{10}$$

$$L' = \frac{\tan U}{\tan U'} L. \tag{11}$$

On account of (9), (11) may also be written in the form

$$L' = \frac{n'}{n} \frac{\cos U'}{\cos U} L, \tag{11a}$$

which is more convenient for computation than (11) if the angles are small.

It is useful to determine also the coordinates (Y_k, Z_k) of the point P_k of incidence at the kth surface, and the distance $D_k = P_k P_{k+1}$ between two successive points of incidence (see Fig. 4.39).

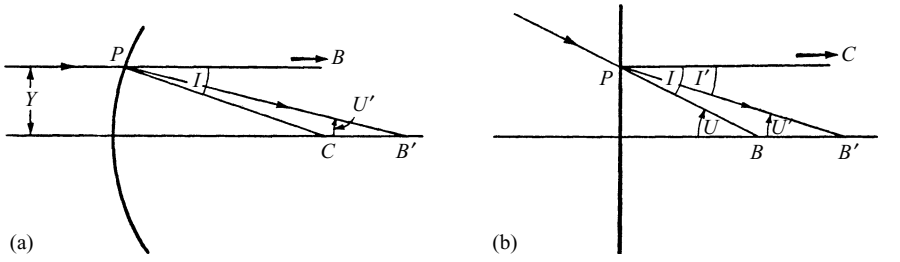

Fig. 4.38 The special cases: (a) $L = \infty$; (b) $r = \infty$.

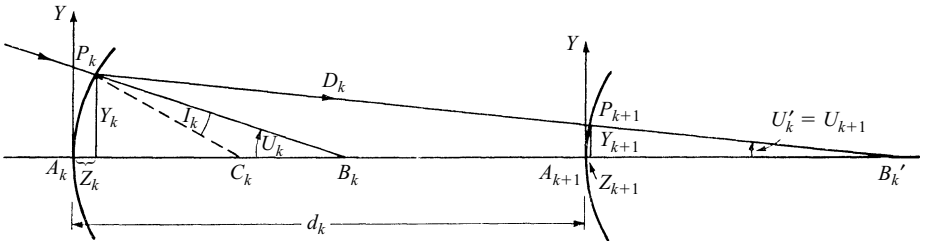

Fig. 4.39 Oblique meridional ray trace through two successive refracting surfaces.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

From the figure,

$$Y_k = r_k \sin(U_k + I_k), \tag{12}$$

$$Z_k = r_k - r_k \cos(U_k + I_k) = \frac{{Y_k}^2}{[1 + \cos(U_k + I_k)]r_k}, \tag{13}$$

$$D_k = (d_k + Z_{k+1} - Z_k)\sec U'_k. \tag{14}$$

In terms of Y_k, Z_k and U'_k,

$$L'_k = Z_k + \frac{Y_k}{\tan U'_k}. \tag{15}$$

This relation may be used as a check on the value computed from (11a) or (11).

In the special case when L_k is infinite, $U_k = 0$, and (12)–(15) still hold. If r_k is infinite, Y_k may be computed from the relation

$$Y_k = L_k \tan U_k, \tag{16}$$

Z_k then being zero.

4.9.2 Paraxial rays

If the inclination of a ray to the axis is sufficiently small, the sines of the various angles may be replaced, in the preceding formulae, by the angles themselves. The formulae then reduce to the Gaussian approximation for the path of the light. Such '*paraxial* ray-tracing formulae' are used in practice for computing the Gaussian magnification and the focal length of the system. A brief summary of these formulae will therefore be given here.

It is customary to denote quantities which refer to the paraxial region by small letters. The refraction equations (1)–(4) become

$$i = \frac{l - r}{r} u, \tag{17}$$

$$i' = \frac{n}{n'} i, \tag{18}$$

$$u' = u + i - i', \tag{19}$$

$$l' = \frac{i'}{u'} r + r. \tag{20}$$

The transfer equations (5) and (6) take the form

$$l_2 = l'_1 - d, \tag{21}$$

$$u_2 = u'_1. \tag{22}$$

In a similar way the paraxial equations for the cases $L = \infty$ and $r = \infty$ are obtained from (7)–(11a).

The paraxial equation for the incidence height, needed later, follows from (12):

$$y_k = r_k(u_k + i_k). \tag{23}$$

Although the relations (17)–(20) involve the angles which the incident ray and the refracted ray make with the axis, l' is independent of these quantities. This result,

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

established in a different manner in §4.4.1 follows when i, i' and u' are eliminated from the above relations. It is then found that u also disappears, and we obtain

$$n'\left(\frac{1}{r}-\frac{1}{l'}\right)=n\left(\frac{1}{r}-\frac{1}{l}\right). \tag{24}$$

This will be recognized as Abbe's relation §4.4 (7).

To determine the lateral Gaussian magnification M, it is only necessary to trace a paraxial ray from the axial object point. Then according to §4.4 (54),

$$M=\frac{n_1}{n_l}\frac{u_1}{u_l}, \tag{25}$$

where the subscripts 1 and l refer to the first and last medium.

The focal length f' of the system may be obtained by tracing a paraxial ray at any desired height y_1 from an infinitely distant object. The equation of the conjugate ray in the image space, referred to axes at the second focal point, is then $y_l/z_l=-u_l'$, and it follows from §4.3 (10) that

$$f'=-\frac{y_1}{u_l'}. \tag{26}$$

4.9.3 Skew rays

So far only rays which lie in a meridional plane were considered. We shall now briefly discuss the tracing of *skew* rays, i.e. rays which are not coplanar with the axis. The tracing of such rays is much more laborious and is usually carried out only in the design of systems with very high apertures.

A ray will now be specified by its direction cosines and by the coordinates of the point at which it meets a particular surface of the system. We take Cartesian rectangular axes at the pole A_1 of the first surface, with the Z-axis along the axis of the system. Let L_1, M_1, N_1, $(L_1^2+M_1^2+N_1^2=1)$ be the direction cosines of a ray incident at the point $P_1(X_1,\ Y_1,\ Z_1)$ (see Fig. 4.40).

The first step is to calculate the cosine of the angle I_1 of incidence. If r_1 is the radius of the first surface, and direction cosines of the normal at P_1 are

$$\overline{L}_1=-\frac{X_1}{r_1},\qquad \overline{M}_1=-\frac{Y_1}{r_1},\qquad \overline{N}_1=\frac{r_1-Z_1}{r_1},$$

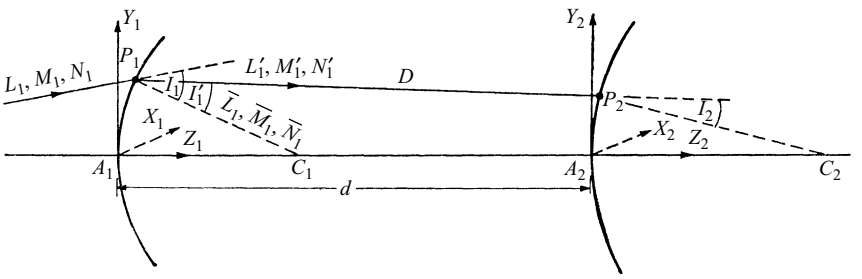

Fig. 4.40 Tracing a skew ray.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

so that

$$\cos I_1 = L_1\overline{L}_1 + M_1\overline{M}_1 + N_1\overline{N}_1$$
$$= N_1 - \frac{1}{r_1}(L_1X_1 + M_1Y_1 + N_1Z_1). \tag{27}$$

The next step is to determine the direction cosines L_1', M_1', N_1' of the refracted ray. This is done in two stages: One calculates first the cosine of the angle of refraction, using the law of refraction in the form

$$n'\cos I_1' = \sqrt{n'^2 - n^2 + n^2\cos^2 I_1}. \tag{28}$$

One then uses the fact that the refracted ray lies in the plane specified by the incident ray and the surface normal. Denoting by $\mathbf{s}_1$, $\mathbf{s}_1'$ and $\bar{\mathbf{s}}_1$ the unit vectors along the incident ray, the refracted ray, and the normal, i.e. the vectors with components (L_1, M_1, N_1), (L_1', M_1', N_1') and $(\overline{L}_1, \overline{M}_1, \overline{N}_1)$, the coplanarity condition gives

$$\mathbf{s}_1' = \lambda\mathbf{s}_1 + \mu\bar{\mathbf{s}}_1, \tag{29}$$

where λ and μ are certain scalar functions. To determine λ and μ we first multiply (29) scalarly by $\mathbf{s}_1$, and use the fact that (see Fig. 4.40) $\mathbf{s}_1 \cdot \mathbf{s}_1' = \cos(I_1 - I_1')$, $\mathbf{s}_1 \cdot \bar{\mathbf{s}}_1 = \cos I_1$. This gives

$$\cos(I_1 - I_1') = \lambda + \mu\cos I_1.$$

Next we multiply (29) scalarly by $\bar{\mathbf{s}}_1$, and use the relations $\bar{\mathbf{s}}_1 \cdot \mathbf{s}_1' = \cos I_1'$, $\bar{\mathbf{s}}_1 \cdot \mathbf{s}_1 = \cos I_1$. We then obtain

$$\cos I_1' = \lambda\cos I_1 + \mu.$$

From the last two relations it follows that

$$\lambda = \frac{\sin I_1'}{\sin I_1} = \frac{n}{n'}, \qquad \mu = \frac{1}{n'}(n'\cos I_1' - n\cos I_1),$$

and (29) then gives the following three equations for the direction cosines of the refracted ray:

$$\left.\begin{aligned} n'L_1' &= nL_1 - \sigma X_1, \\ n'M_1' &= nM_1 - \sigma Y_1, \\ n'N_1' &= nN_1 - \sigma(Z_1 - r_1), \end{aligned}\right\} \tag{30}$$

where

$$\sigma = \frac{1}{r_1}(n'\cos I_1' - n\cos I_1). \tag{31}$$

This completes the ray trace through the first surface, by means of the *refraction equations* (27), (28), (30) and (31).

The refracted ray now becomes the incident ray for the second surface. Taking parallel axes at the pole A_2 of the second surface, we have the *transfer equations for the direction cosines*

$$L_2 = L_1', \qquad M_2 = M_1', \qquad N_2 = N_1'. \tag{32}$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

Let d be the distance A_1A_2 between the poles of the two surfaces. Denoting by X_1^+, Y_1^+, Z_1^+ the coordinates of the point P_1 referred to the axes at A_2, we have the *transfer equations for the coordinates*

$$X_1^+ = X_1, \qquad Y_1^+ = Y_1, \qquad Z_1^+ = Z_1 - d. \tag{33}$$

Next the coordinates (X_2, Y_2, Z_2) of the point P_2 in which the refracted ray meets the second surface must be determined. If D denotes the distance from P_1 and P_2, then

$$\left.\begin{aligned} X_2 &= X_1^+ + L_2 D, \\ Y_2 &= Y_1^+ + M_2 D, \\ Z_2 &= Z_1^+ + N_2 D. \end{aligned}\right\} \tag{34}$$

To determine D we use the fact that P_2 lies on the second surface. If r_2 is the radius of this surface, then

$$X_2^2 + Y_2^2 + Z_2^2 - 2Z_2 r_2 = 0. \tag{35}$$

On substituting into this equation from (34) the following equation for D is obtained:

$$D^2 - 2Fr_2 D + Gr_2 = 0, \tag{36}$$

where

$$\left.\begin{aligned} F &= N_2 - \frac{1}{r_2}(L_2 X_1^+ + M_2 Y_1^+ + N_2 Z_1^+), \\ G &= \frac{1}{r_2}[(X_1^+)^2 + (Y_1^+)^2 + (Z_1^+)^2] - 2Z_1^+. \end{aligned}\right\} \tag{37}$$

Eq. (36) gives the following expression for D, if we also make use of the fact that $D = d$ for the axial ray:

$$D = r_2\left(F - \sqrt{F^2 - \frac{1}{r_2}G}\right). \tag{38}$$

The coordinates of P_2 are determined on substituting for D into (34).

Finally, to complete the cycle, the cosine of the angle I_2 of incidence at the second surface must be calculated. It is given by a relation strictly analogous to (27):

$$\cos I_2 = N_2 - \frac{1}{r_2}(L_2 X_2 + M_2 Y_2 + N_2 Z_2), \tag{39}$$

or, using (34) and (38)

$$\begin{aligned} \cos I_2 &= F - \frac{1}{r_2} D \\ &= \sqrt{F^2 - \frac{1}{r_2}G}. \end{aligned} \tag{40}$$

Hence the last stage consists of the evaluation of the formulae (40), (38) and (34).

Because of the labour involved in the tracing of a general skew ray, the calculations are sometimes restricted to skew rays which lie in the immediate neighbourhood of a selected principal ray. Such skew rays may be traced through the system with the help

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

of simplified schemes,* which are similar to those used for paraxial ray tracing and are adequate for determining the position of the sagittal focal surface.

4.10 Design of aspheric surfaces

In the great majority of optical systems lenses and mirrors are employed, the surfaces of which have plane, spherical or paraboloidal form. The restriction to surfaces of such simple form is mainly due to the practical difficulties encountered in the production of surfaces of more complicated shapes with the high degree of precision required in optics. The restriction to surfaces of simple form naturally imposes limitations on the ultimate performance which systems of conventional design can attain. For this reason, in spite of the difficulties of manufacture, surfaces of more complicated form, called *aspheric* surfaces, are employed in certain systems. As early as 1905, K. Schwarzschild† considered a class of telescope objectives consisting of two aspheric mirrors, and showed that such systems can be made aplanatic.

In 1930, Bernhard Schmidt, an optician of Hamburg, constructed a telescope of a new type, which consisted of a spherical mirror and a suitably designed aspheric lens placed at its centre of curvature. The performance of this system (considered more fully in §6.4) was found to be quite outstanding. By means of such a telescope it is possible to photograph on one plate a very large region of the sky, many hundred times larger than can be obtained with telescopes of conventional design. The *Schmidt camera* has since become an important tool in astronomical research. Aspheric systems which use the principle of the Schmidt camera are also employed in certain projection-type television receivers,‡ in X-ray fluorescent screen photography, and certain fast low-dispersion spectrographs. Aspheric surfaces find also useful application in microscopy (see §6.6).

By making one surface of any centred system aspherical, it is possible, in general, to ensure exact axial stigmatism; by means of two aspheric surfaces any centred system may in general be made aplanatic. In this section formulae will be derived for the design of such aspheric surfaces.

4.10.1 Attainment of axial stigmatism§

Consider the rays from an axial object point P. In the image space of the system the rays from different zones of the exit pupil will in general intersect the axis at different points. Let $S^{(0)}$ be the last surface of the system and O its axial point (Fig. 4.41). It will be shown that it is possible to modify the profile of $S^{(0)}$ in such a way as to compensate exactly for the departure from the homocentricity of the image-forming

* See H. H. Hopkins, *Proc. Phys. Soc.*, **58** (1946), 663; also his *Wave Theory of Aberrations* (Oxford, Clarendon Press, 1950), pp. 59, 65.

† K. Schwarzschild, *Astr. Mitt. Königl. Sternwarte Göttingen* (1905). Reprinted from *Abh. Königl. Ges. Wiss. Göttingen, Math. Phys. Klasse*, **4** (1905–1906), No. 2.

Two telescopes of this type were later constructed, one with an aperture of 24 in at the University of Indiana and another, with a 12-in aperture, at Brown University.

‡ See, for example, I. G. Malloff and D. W. Epstein, *Electronics* **17** (1944), 98.

§ The methods here described are due to E. Wolf and W. S. Preddy, *Proc. Phys. Soc.*, **59** (1947), 704; also E. Wolf, *Proc. Phys. Soc.*, **61** (1948), 494. Similar formulae were also given by R. K. Luneburg in his *Mathematical Theory of Optics* (Berkeley and Los Angeles, University of California Press, 1964), §24.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

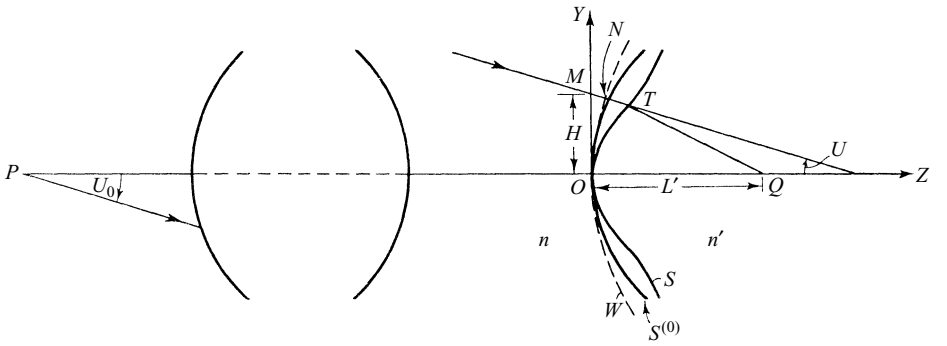

Fig. 4.41 Design of an aspheric surface to attain axial stigmatism.

pencil; more precisely we shall show that it is possible in general to replace $S^{(0)}$ by a new surface S which will ensure that all the rays in the image space meet the axis at any prescribed point Q.

On account of symmetry only meridional rays need be considered. Each ray in the space which precedes the last surface will be specified by the following parameters: the angle U which it makes with the axis and the distance $H = OM$ at which it intersects the Y-axis (see Fig. 4.41). It will be convenient to label the rays: Let t be any convenient parameter, for example the angle U_0 which the corresponding ray in the object space makes with the axis, or the height at which it meets the first surface. The pencil may then be completely specified by the two functional relations

$$U = U(t), \qquad H = H(t). \tag{1}$$

It may be assumed that $t = 0$ for the axial ray. In general the relations (1) will not be known in an explicit form; a table of values of U and H for any prescribed set of t values may, however, be obtained from a ray trace.

Let W be the orthogonal surface (wave-front) through O of the pencil (Fig. 4.41), and let N be the point in which the ray through M meets it. Assuming that a correcting surface S with the required property can be found, it follows that in the corrected system the optical path length from N to Q must be equal to the optical path length from O to Q. Hence, if $T(Y, Z)$ is the point in which the ray meets S,

$$[NT] + [TQ] = [OQ]. \tag{2}$$

If n is the refractive index of the space which precedes S, and n' the refractive index of the image space, we have*

$$\left.\begin{aligned} [NT] &= [MT] - [MN] = nZ\sec U - [MN], \\ [TQ] &= n'\sqrt{(L' - Z)^2 + (H - Z\tan U)^2}, \\ [OQ] &= n'L', \end{aligned}\right\} \tag{3}$$

L' denoting the distance from O to Q.

The optical path $[MN]$ which occurs in the expression for $[NT]$ may also be

* As in §4.9.1, the angle U is considered to be positive if the axis can be brought into coincidence with the ray by a clockwise rotation of less than 90° about the axial point.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

evaluated in terms of the given quantities. By applying Lagrange's invariant relation §3.3 (1) to the curve formed by the segment OM of the Y-axis, the segment MN of the ray and the curve NO on the wave-front W, we have

$$\int_{OM} n\mathbf{s}\cdot\mathrm{d}\mathbf{r}+\int_{MN} n\mathbf{s}\cdot\mathrm{d}\mathbf{r}+\int_{NO} n\mathbf{s}\cdot\mathrm{d}\mathbf{r}=0, \tag{4}$$

where $\mathbf{s}$ is the unit vector along the ray and $\mathrm{d}\mathbf{r}$ an element of the path of integration. Now from the figure,

$$\left.\begin{aligned}&\int_{OM} n\mathbf{s}\cdot\mathrm{d}\mathbf{r}=-n\int_0^H \sin U\,\mathrm{d}H=-n\int_0^t \sin U\frac{\mathrm{d}H}{\mathrm{d}t}\,\mathrm{d}t,\\&\int_{MN} n\mathbf{s}\cdot\mathrm{d}\mathbf{r}=[MN],\\&\int_{NO} n\mathbf{s}\cdot\mathrm{d}\mathbf{r}=0,\end{aligned}\right\} \tag{5}$$

and hence (4) gives

$$[MN]=n\int_0^t \sin U\frac{\mathrm{d}H}{\mathrm{d}t}\,\mathrm{d}t. \tag{6}$$

Eq. (6) expresses $[MN]$ as an integral which may be evaluated numerically from a table of U and H values.*

Substitution from (3) and (6) into (2) leads to the following equation for Z:

$$\mathcal{A}Z^2+2\mathcal{B}Z\cos U+\mathcal{C}\cos^2 U=0, \tag{7}$$

with

$$\left.\begin{aligned}&\mathcal{A}=n'^2-n^2,\\&\mathcal{B}=n^2\int_0^t \sin U\frac{\mathrm{d}H}{\mathrm{d}t}\,\mathrm{d}t-n'^2(L'\cos U+H\sin U)+nn'L',\\&\mathcal{C}=n'^2H^2-\left(n\int_0^t \sin U\frac{\mathrm{d}H}{\mathrm{d}t}\,\mathrm{d}t\right)\left(n\int_0^t \sin U\frac{\mathrm{d}H}{\mathrm{d}t}\,\mathrm{d}t+2n'L'\right).\end{aligned}\right\} \tag{8}$$

Hence

$$Z=\frac{\cos U}{\mathcal{A}}[-\mathcal{B}\pm\sqrt{\mathcal{B}^2-\mathcal{A}\mathcal{C}}]. \tag{9}$$

Also, from the figure,

$$Y=H-Z\tan U. \tag{10}$$

Since $Z=U=H=0$ when $t=0$, and since we assume L' to be positive (see Fig. 4.41), the positive sign must be taken in front of the square root in (9). Finally, combining (9) and (10), we obtain

* It is, of course, possible to evaluate $[MN]$ directly from a ray trace by making use of the property that the optical path from P to N is equal to the optical path from P to O. This gives

$$[MN]=[PO]-[PM].$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$Z + \mathrm{i}Y = \frac{-\mathcal{B} + \sqrt{\mathcal{B}^2 - \mathcal{A}\mathcal{C}}}{\mathcal{A}}\,\mathrm{e}^{-\mathrm{i}U} + \mathrm{i}H. \tag{11}$$

Eq. (11) is an exact parametric equation of the aspheric surface S in terms of the free parameter t.

The special case when the focus Q is at infinity ($L' = \infty$) is also of interest. To derive the appropriate formula we note that both $\mathcal{B}$ and $\mathcal{C}$ contain L' in the first power only. Hence for sufficiently large L',

$$\begin{aligned} -\mathcal{B} + \sqrt{\mathcal{B}^2 - \mathcal{A}\mathcal{C}} &= \mathcal{B}\left[-1 + \sqrt{1 - \frac{\mathcal{A}\mathcal{C}}{\mathcal{B}^2}}\right] \\ &= -\frac{1}{2}\frac{\mathcal{A}\mathcal{C}}{\mathcal{B}} - \frac{1}{8}\frac{\mathcal{A}^2\mathcal{C}^2}{\mathcal{B}^3} - \cdots \\ &= -\frac{1}{2}\frac{\mathcal{A}\mathcal{C}}{\mathcal{B}} + O\left(\frac{1}{L'}\right). \end{aligned}$$

In the limit, as $L' \to \infty$, (11) therefore reduces to

$$Z + \mathrm{i}Y = \frac{n\mathrm{e}^{-\mathrm{i}U}}{n - n'\cos U}\int_0^t \sin U \frac{\mathrm{d}H}{\mathrm{d}t}\,\mathrm{d}t + \mathrm{i}H. \tag{12}$$

We have only considered the case when the aspheric surface is the last surface of the system but the method may be extended to the design of an aspheric surface situated in the interior of an optical system. The computation is, however, much more laborious in such cases and will not be considered here.*

4.10.2 Attainment of aplanatism†

We have seen that by making one surface of a system aspherical, it is possible to attain exact axial stigmatism. We shall now consider the design of two aspheric surfaces which will ensure not only axial stigmatism but also the satisfaction of the sine condition.

Let S and S' be the two aspheric surfaces, the profiles of which are to be determined. It will be assumed that S and S' are neighbouring surfaces in the system.‡ They may, however, be separated from the object or image points by any number of refracting or reflecting surfaces. Again, we shall be concerned only with the final correction of the system, and assume that all the design data, save the profiles of S and S', are known.

We introduce two sets of Cartesian rectangular axes, with origins at the poles O and O' of S and S', and with the Z-axes along the axis of the system. The surface S will be referred to the axes at O, and S' to the axes at O'.

* For details see E. Wolf, *Proc. Phys. Soc.*, **61** (1948), 494. Other methods were described by M. Herzberger and H. O. Hoadley, *J. Opt. Soc. Amer.*, **36** (1946), 334; and D. S. Volosov *J. Opt. Soc. Amer.*, **37** (1947), 342.

† The methods described in this section are due to G. D. Wassermann and E. Wolf, *Proc. Phys. Soc.* B, **62** (1949), 2.

‡ A generalization to systems where S and S' are not optical neighbours was described by E. M. Vaskas, *J. Opt. Soc. Amer.*, **47** (1957), 669.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

In the space which precedes S, the pencil of rays from the axial object point P will again be specified by a relation of the form (see Fig. 4.42)

$$U = U(t), \qquad H = H(t). \tag{13}$$

The pencil of rays in the space which follows S' (in the corrected system) will be specified by a similar relation:

$$U' = U'(t'), \qquad H' = H'(t'). \tag{14}$$

Eq. (14) may be obtained in a tabulated form by tracing rays backwards from the selected axial image point P'.

If the object and image are at finite distances we choose as the parameters t and t' the sines of the angles which the corresponding rays in the object and image spaces make with the axis of the system:*

$$t = \sin U_0, \qquad t' = \sin U_1. \tag{15}$$

If the object is at infinity, we choose as the t parameter the distance H_0 of the corresponding ray in the object space from the axis; if the image is at infinity we take $t' = H_1$, H_1 being the distance from the axis of the corresponding ray in the image space. In either of these cases, the sine condition demands that

$$\frac{t}{t'} = \text{constant}. \tag{16}$$

Our problem may now be formulated as follows: Given (13) and (14), find two surfaces S and S' which ensure that the pencil (U, H) goes over into the pencil (U', H') by successive refractions at the two surfaces; and, moreover, the corresponding rays in the two pencils must satisfy the relation (16).

Let n be the refractive index of the space which precedes S, n' the refractive index of the space which follows S', and $n^\star$ that of the space between them. Further let $\mathbf{s}$ be the unit vector along the ray incident at the point $T(Z, Y)$ and $\mathbf{s}^\star$ the unit vector along the refracted ray (see Fig. 4.42).

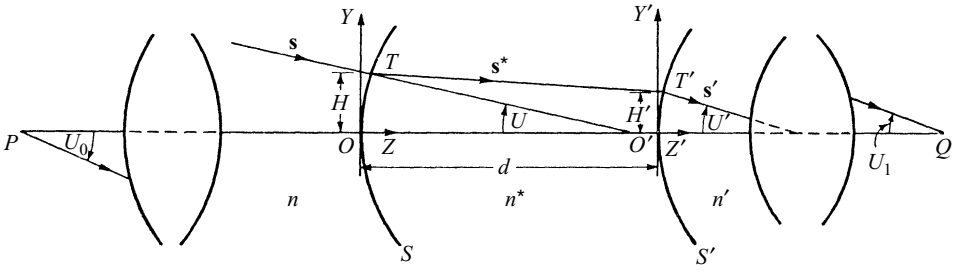

Fig. 4.42 Design of two aspheric surfaces to attain aplanatism.

* If instead of the sine condition we wished (in addition to axial stigmatism) to satisfy the Herschel condition, we would choose

$$t = \sin U_0/2, \qquad t' = \sin U_1/2.$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

According to the law of refraction (§3.2.2) the vector $\mathbf{N} = n\mathbf{s} - n^\star\mathbf{s}^\star$ must be in the direction of the surface normal at T. Hence if τ is the unit tangent at T of the meridional section of the surface,

$$(n\mathbf{s} - n^\star\mathbf{s}^\star)\cdot\tau = 0. \tag{17}$$

Now the components of the vectors s, $s^\star$ and τ are:

$$\left.\begin{array}{lccc} \mathbf{s}: & 0, & -\sin U, & \cos U, \\ \mathbf{s}^\star: & 0, & -\sin U^\star, & \cos U^\star, \\ \tau: & 0, & \dfrac{\dot{Y}}{\sqrt{\dot{Y}^2+\dot{Z}^2}}, & \dfrac{\dot{Z}}{\sqrt{\dot{Y}^2+\dot{Z}^2}}; \end{array}\right\} \tag{18}$$

here $U^\star$ is the angle which the refracted ray TT' makes with the axis of the system, and the dot denotes differentiation with respect to the parameter t. Eq. (17) becomes

$$n(\dot{Z}\cos U - \dot{Y}\sin U) = n^\star(\dot{Z}\cos U^\star - \dot{Y}\sin U^\star). \tag{19}$$

Now if D is the distance from T to T', D_y and D_z the projections of D on to the Y and Z-axes, and d the axial distance from O to O',

$$\cos U^\star = \frac{D_z}{D}, \qquad \sin U^\star = \frac{D_y}{D}, \tag{20}$$

with

$$D_y = Y - Y', \qquad D_z = d + Z' - Z, \qquad D = \sqrt{{D_y}^2 + {D_z}^2}. \tag{21}$$

Also, from the figure,

$$Y = H - Z\tan U, \tag{22}$$

$$Y' = H' - Z'\tan U'. \tag{23}$$

On substituting in (19) for $\cos U^\star$ and $\sin U^\star$ from (20), and for $\dot{Y}$ from (22), we find that

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \left(\frac{nD\cos U - n^\star D_z}{nD\sin U - n^\star D_y} + \tan U\right)^{-1}\left(\frac{\mathrm{d}H}{\mathrm{d}t} - Z\frac{\mathrm{d}}{\mathrm{d}t}(\tan U)\right). \tag{24}$$

Similarly

$$\frac{\mathrm{d}Z'}{\mathrm{d}t'} = \left(\frac{n'D\cos U' - n^\star D_z}{n'D\sin U' - n^\star D_y} + \tan U'\right)^{-1}\left(\frac{\mathrm{d}H'}{\mathrm{d}t'} - Z'\frac{\mathrm{d}}{\mathrm{d}t'}(\tan U')\right). \tag{25}$$

Eqs. (21)–(25), subject to the relation (16) and the boundary conditions $Z = Z' = 0$ when $t = t' = 0$, enable a complete computation of the two correcting surfaces to be carried out. For, using (21), (22) and (23) we may eliminate Y and Y' from (24) and (25); and, using (16), we then obtain two first-order simultaneous differential equations for Z and Z' of the type

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = f(Z, Z', t), \qquad \frac{\mathrm{d}Z'}{\mathrm{d}t} = g(Z, Z', t). \tag{26}$$

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

These may be integrated by standard methods.* Since, however, it is necessary to determine not only Z and Z' but also Y and Y' for a selected range of the parameter t, it is preferable not to eliminate Y and Y' but rather to solve for the unknown quantities step by step.

4.11 Image reconstruction from projections (computerized tomography)

4.11.1 Introduction

The theory of imaging which we have considered so far has been mainly concerned with the formation of images of planar objects. If the object is three-dimensional and semi-transparent, the image of each of its planar cross-sections does not provide reliable information about the object; this is so because on propagating to the image plane the light must first pass through the portion of the object in front of the cross-section of interest [see Fig. 4.43(b)] and the information which it carries then becomes distorted. Only if the interaction of the light with the portion of the object through which it must pass is sufficiently weak (as is frequently but not always the case in

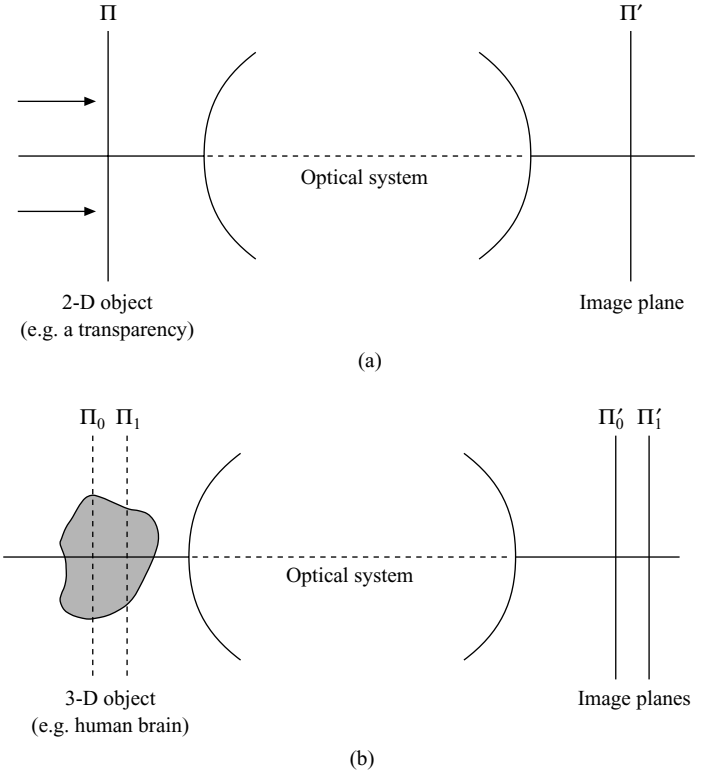

Fig. 4.43 Imaging of (a) a two-dimensional and (b) a three-dimensional object.

* For example, by Adams' method (see E. T. Whittaker and G. Robinson, *The Calculus of Observations* (Glasgow, Blackie & Son, 4th edition, 1946), p. 363), or by the method of Runge and Kutta [see C. Runge and H. König, *Numerisches Rechnen*, (Berlin, Springer, 1924)].

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

microscopy) can one ignore such distortions.

The problem of obtaining reliable information about the structure of three-dimensional objects arises in many fields. It is of particular importance in diagnostic medicine but such problems also frequently arise in many other fields.

Because direct imaging cannot provide reliable information about the structure of three-dimensional objects, some indirect methods must be used. The oldest, and undoubtedly the most successful one for some applications, is so-called computerized (or computed) axial tomography, or just computerized tomography, often abbreviated as CAT or CT. This technique was originally developed for use in medical diagnostics with X-rays, but was later adapted for other applications, using in many cases other kinds of radiation and even elementary particles.

In computerized tomography the underlying physical model for propagation of radiation through the object is still that of geometrical optics but the reconstruction process requires much more sophisticated mathematical methods than are used in the usual theory of optical image formation. In this section we will describe this reconstruction technique.

4.11.2 Beam propagation in an absorbing medium

Before considering the main problem of reconstruction of three-dimensional objects by computerized tomography, we will derive an intensity law for beams which propagate in a weakly absorbing medium whose real refractive index n is constant. Were the medium non-absorbing, the variation of the space-dependent parts of a (complex) monochromatic electromagnetic field of frequency ω along each ray in the beam would be given by the formula [see §3.1 (6)]

$$\mathbf{E}_2 = \mathbf{E}_1 e^{ik_0 nl}, \qquad \mathbf{H}_2 = \mathbf{H}_1 e^{ik_0 nl}, \tag{1}$$

where $k_0 = \omega/c$ is the free-space wave number and l is the distance between two typical points P_1 and P_2 on the ray (Fig. 4.44); $\mathbf{E}_1$ and $\mathbf{E}_2$ are the space-dependent parts of the electric fields at the points P_1 and P_2, respectively. $\mathbf{H}_1$ and $\mathbf{H}_2$ are the space-dependent parts of the magnetic fields at these points.

Suppose next that the medium is weakly absorbing. Formally, absorption may be taken into account by replacing the real refractive index n by a complex one, which we denote by $\hat{n}$,

$$\hat{n} = n(1 + i\kappa), \tag{2}$$

where κ, just like n, is a real constant. The constant κ is usually called the *attenuation*

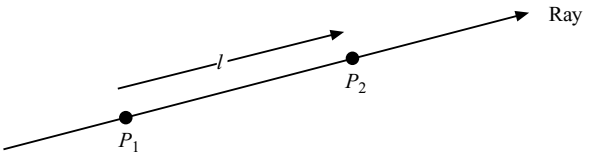

Fig. 4.44 Illustrating the formula (1).

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

index of the medium. The theory of the complex refractive index $\hat{n}$ will be discussed in §14.1.

As before we identify the intensity of the field with the absolute value of the Poynting vector [see §3.1 (28) and §1.4 (56)], i.e.

$$I = \frac{c}{8\pi}|\mathcal{R}(\mathbf{E} \times \mathbf{H}^\star)|, \tag{3}$$

where $\mathcal{R}$ denotes the real part. It follows from (1), with n replaced by $\hat{n} = n(1 + \mathrm{i}\kappa)$, that

$$\mathbf{E}_2 \times \mathbf{H}_2^\star = \mathbf{E}_1 \times \mathbf{H}_1^\star \mathrm{e}^{-2k_0 n\kappa l}. \tag{4}$$

From (4) and (3) we see at once that the intensities at the points P_1 and P_2 are related by the formula

$$I_2 = I_1 \mathrm{e}^{-\alpha l}, \tag{5}$$

where the constant

$$\alpha = 2k_0 n\kappa \tag{6}$$

is called the *absorption coefficient* of the medium.

Suppose next that the absorption coefficient is not a constant but that it varies with position. Then in place of (5) one has the more general formula

$$I_2 = I_1 \exp\left(-\int_{P_1}^{P_2} \alpha(\mathbf{r})\,\mathrm{d}l\right), \tag{7}$$

where the integral is taken along the ray from the initial point P_1 to the final point P_2. Formula (7) is often called *Beer's law* and is frequently derived by quasi-geometrical arguments of the theory of radiative energy transfer.*

4.11.3 Ray integrals and projections

Suppose that a narrow beam of X-rays traverses a portion of the human body. Because the wavelengths of X-rays are generally very short on the scale of variation of the physical properties of the human body, geometrical optics will adequately describe the propagation, and Beer's law (7) may be expected to apply.

One of the principal quantities which provides information about many organs in the human body is the absorption coefficient $\alpha(\mathbf{r})$. From the knowledge of its distribution, radiologists and physicians can deduce a good deal of information about the anatomy of the organ. Suppose that a beam of nearly monochromatic X-rays is passed through some organ or tissue of the human body from a source point P_0 to a detector located at a point P (see Fig. 4.45), which measures the intensity, denoted by I. It follows from Beer's law (7) that

$$\int_{P_0}^{P} \alpha(\mathbf{r})\mathrm{d}s = -\ln \frac{I(P)}{I(P_0)}. \tag{8}$$

* See, for example, L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge, Cambridge University Press, 1995), Sec. 5.7.4.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

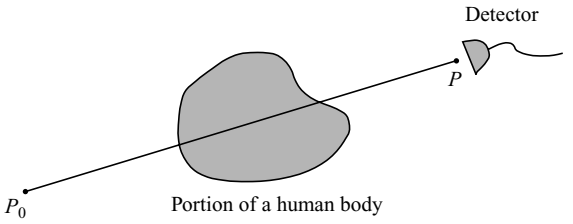

Fig. 4.45 Illustrating the concept of a ray integral defined by (8).

Hence from measurements of the intensity $I(P)$ at the detector and from knowledge of the intensity $I(P_0)$ at the source point P_0 one can deduce at once the value of the integral of the absorption coefficient from P_0 to P along the ray, which appears on the left-hand side of (8). The integral is called the *ray integral* and a suitable set of ray integrals is said to form a *projection*. Two kinds of projections, called fan-beam projection and parallel projection, are shown in Figs. 4.46(a) and (b).

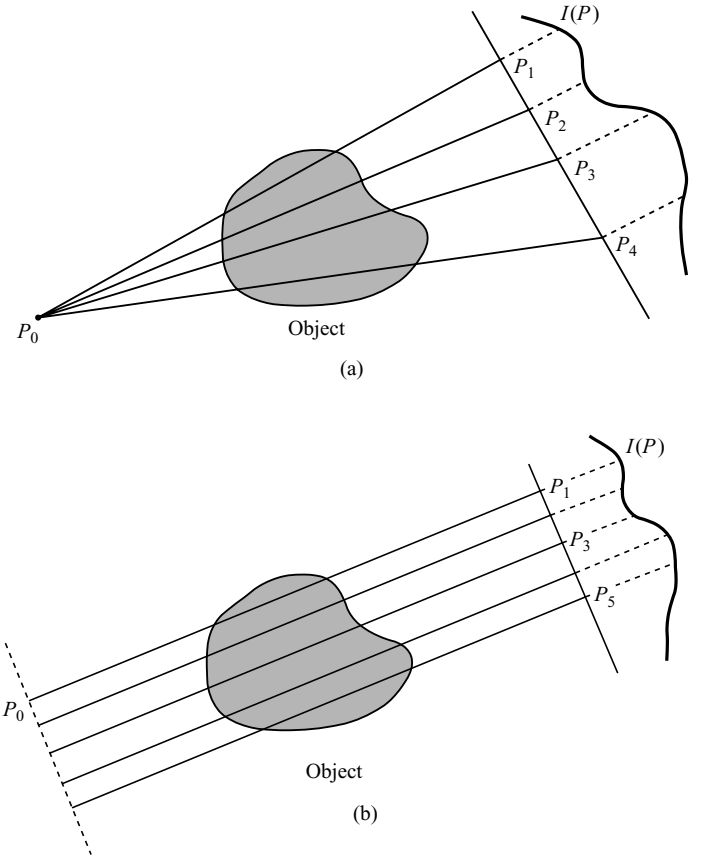

Fig. 4.46 Types of projections commonly used in computerized tomography:
(a) fan-beam projection; (b) parallel projection.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

The basic question of computerized tomography is the following: Can one deduce, from a sufficient number of projections, the spatial distribution of the absorption coefficient $\alpha(\mathbf{r})$ throughout an organ of the human body or throughout some other object? The answer to this question is yes. The mathematical solution to this problem was obtained many years ago by J. Radon*, but this was not known to the pioneers of computerized tomography, which was developed about half a century later. We will briefly mention the historical background and development of the subject at the end of §4.11.5.

We will now describe the essence of Radon's work which is relevant to computerized tomography.

4.11.4 The N-dimensional Radon transform

In medical diagnostic applications of computerized tomography utilizing X-rays, one generally wishes to obtain information about planar sections of a portion of the human body. In applications which use electron beams, on the other hand, one usually requires information about three-dimensional objects. To cover the different cases it will be convenient to describe the reconstruction procedure for a domain with an arbitrary number of spatial dimensions. Later we will specialize the results to two dimensions.

Let $\mathbf{x}$ and $\boldsymbol{\xi}$ be N-dimensional vectors,

$$\mathbf{x} \equiv (x_1, x_2, \ldots x_N),$$

$$\boldsymbol{\xi} \equiv (\xi_1, \xi_2, \ldots \xi_N),$$

and let

$$\boldsymbol{\xi} \cdot \mathbf{x} \equiv \xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_N x_N$$

be the scalar product of the two vectors. Further let $f(\mathbf{x})$ be an arbitrary function of $\mathbf{x}$. The *Radon transform* of $f(\mathbf{x})$ is defined by the formula

$$F(\boldsymbol{\xi}, p) = \int f(\mathbf{x}) \delta(p - \boldsymbol{\xi} \cdot \mathbf{x}) \, \mathrm{d}^N x, \tag{9}$$

where δ is the one-dimensional Dirac delta function, p is an arbitrary scalar and the integration extends over the whole $\mathbf{x}$-space.

The presence of the Dirac delta function in the integrand in (9) implies that the function $F(\boldsymbol{\xi}, p)$ contains only those contributions from $f(\mathbf{x})$ for which the end points of the vector $\mathbf{x}$ are constrained to lie on the locus

$$\boldsymbol{\xi} \cdot \mathbf{x} = p. \tag{10}$$

In two dimensions this locus is a straight line, in three dimensions it is a plane. In general the locus is an $(N-1)$-dimensional hyperplane in an N-dimensional space. The three-dimensional case is illustrated in Fig. 4.47. The plane to which the end points of the vectors $\mathbf{x}$ are constrained is perpendicular to the vector $\boldsymbol{\xi}$ and the distance of the plane from the origin, measured along the normal to the plane, is

* J. Radon, Ber. König. Säch. Akad. Wiss. (Leipzig), *Math. Phys. Klasse* **69** (1917), 262. An English translation of this paper is included in S. R. Deans, *The Radon Transform and some of its Applications* (New York, J. Wiley, 1983), pp. 204–217.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

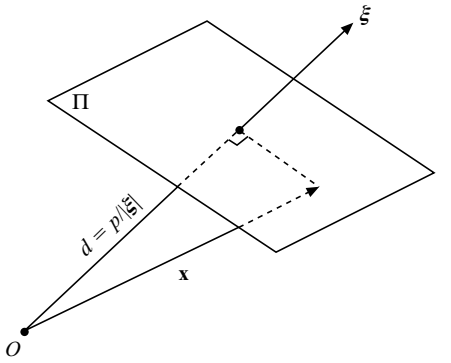

Fig. 4.47 Illustrating the geometrical interpretation of (10) in three dimensions. The locus of $\mathbf{x}$ is the plane Π, at distance $d = p/|\boldsymbol{\xi}|$ from the origin O, perpendicular to the vector $\boldsymbol{\xi}$.

$$d = \frac{p}{|\boldsymbol{\xi}|}. \tag{11}$$

There is a simple property of the Radon transform which we will find useful. Suppose that β is a real parameter different from zero. Then

$$F(\beta\boldsymbol{\xi},\, \beta p) = \int f(\mathbf{x})\delta(\beta p - \beta\boldsymbol{\xi}\cdot\mathbf{x})\, \mathrm{d}^N x. \tag{12}$$

If we use a well-known property of the Dirac delta function [Appendix IV, Eq. (10)] and the definition (9) we obtain the formula

$$F(\beta\boldsymbol{\xi},\, \beta p) = \frac{1}{|\beta|} F(\boldsymbol{\xi},\, p). \tag{13}$$

Eq. (13) is known as the *scaling law* for Radon transforms.

Because of the simple property (13) we may restrict $\boldsymbol{\xi}$, without loss of generality, to be an N-dimensional real unit vector $\mathbf{n} \equiv (n_1,\, n_2,\, \ldots\, n_N)$,

$$\mathbf{n}^2 \equiv n_1^2 + n_2^2 + \cdots + n_N^2 = 1, \tag{14}$$

and employ, in place of (9), the following definition of the Radon transform of $f(\mathbf{x})$:

$$F(\mathbf{n},\, p) = \int f(\mathbf{x})\delta(p - \mathbf{n}\cdot\mathbf{x})\, \mathrm{d}^N x. \tag{15}$$

This is a standard form of the Radon transform.

The function $F(\mathbf{n},\, p)$ is said to represent the *projection* of $f(\mathbf{x})$ onto the direction specified by the unit vector $\mathbf{n}$. It is the set of the integrals of $f(\mathbf{x})$, labeled by p, taken over all the hyperplanes defined by the equation

$$\mathbf{n}\cdot\mathbf{x} = p, \qquad (\mathbf{n}^2 = 1), \tag{16}$$

with the unit vector $\mathbf{n}$ being fixed.

Returning to the reconstruction problem discussed in §4.11.3, it is clear that the integral which appears on the left-hand side in (8) represents the Radon transform of the absorption coefficient $\alpha(\mathbf{r})$ in two dimensions, when the path of integration in (8)

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

is taken along a line perpendicular to the unit vector $\mathbf{n}$ at distance p from the origin. According to (8) the value of the ray integral may be obtained from measurements of the intensity $I(P)$ and from knowledge of the initial intensity $I(P_0)$. Hence the reconstruction problem of computerized tomography is to determine the function $f(\mathbf{x})$, often representing the absorption coefficient $\alpha(\mathbf{r})$, from knowledge of a set of Radon transforms of $f(\mathbf{x})$, i.e. from a set of projections (ray integrals) of $f(\mathbf{x})$. We will now show how this reconstruction can be carried out.

Let us first represent the Dirac delta function which appears in the integrand on the right-hand side of (15) as a Fourier integral [see Appendix IV, Eq. (23)]

$$\delta(p - \mathbf{n}\cdot\mathbf{x}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}u(p-\mathbf{n}\cdot\mathbf{x})}\,\mathrm{d}u. \tag{17}$$

On substituting from (17) into (15) and interchanging the order of integrations we find that

$$F(\mathbf{n},\ p) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}(u\mathbf{n})\mathrm{e}^{-\mathrm{i}up}\,\mathrm{d}u, \tag{18}$$

where $\tilde{f}(\mathbf{K})$ is the N-dimensional Fourier transform of $f(\mathbf{x})$, viz.,

$$\tilde{f}(\mathbf{K}) = \int f(\mathbf{x})\mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{x}}\,\mathrm{d}^N x. \tag{19}$$

On taking the Fourier inverse of (18) we find that

$$\tilde{f}(u\mathbf{n}) = \int_{-\infty}^{\infty} F(\mathbf{n},\ p)\mathrm{e}^{\mathrm{i}up}\,\mathrm{d}p. \tag{20}$$

This is an important formula. It shows that the *N-dimensional Fourier transform* $\tilde{f}(\mathbf{K})$, *with the argument* $\mathbf{K} = u\mathbf{n}$, *of the unknown function* $f(\mathbf{x})$ *is just the one-dimensional Fourier transform, taken with respect to the scalar parameter p, of the Radon transform* $F(\mathbf{n},\ p)$ *of* $f(\mathbf{x})$. This result holds irrespective of the dimensionality N of the $\mathbf{x}$-space.

One may now carry out the reconstruction of the function $f(\mathbf{x})$ from the knowledge of the Radon transform $F(\mathbf{n},\ p)$, which represents the projection of the function $f(\mathbf{x})$ onto the direction $\mathbf{n}$. Hence from knowledge of all the projections of $f(\mathbf{x})$, which is equivalent to knowledge of the Radon transform for all values of its arguments, one can calculate all the Fourier components $\tilde{f}(\mathbf{K})$ by the use of (20). From knowledge of $\tilde{f}(\mathbf{K})$ one can then synthesize the unknown function $f(\mathbf{x})$, using the inverse of (19).*

4.11.5 Reconstruction of cross-sections and the projection-slice theorem of computerized tomography

As already mentioned, in diagnostic medicine one frequently wishes to determine the variation of the absorption coefficient in a cross-sectional plane through the human body. One can often obtain such information by the use of computerized tomography with X-rays. The appropriate formulae to carry out the reconstruction can readily be

* For fuller discussions of the Radon transform and some of its uses see, for example, S. R. Deans, *loc. cit.* or H. H. Barrett, in *Progress in Optics*, Vol. XXI, E. Wolf, ed. (Amsterdam, Elsevier, 1984), p. 217.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

obtained by specializing the results which we just derived to the situation where the dimension of the domains of interest (planar cross-sections) is $N = 2$. We will now briefly consider this important special case.

As already noted there are several types of arrangements (e.g. those shown in Fig. 4.46) which are commonly used. We will only consider parallel projections. The transmitted intensity is then measured on lines perpendicular to the ray directions, as shown in Fig. 4.48. It will be convenient to denote the two-dimensional vector which specifies points in the object by $\boldsymbol{\rho}$ (corresponding to $\mathbf{x}$ in the general case) and denote its Cartesian components by x and y (rather than by x_1 and x_2), as indicated in Fig. 4.48. We will refer to the function $f(\boldsymbol{\rho})$ which is to be determined as the *object function*. It may be the absorption coefficient (denoted by α in §4.11.2) or some other quantity of interest.

The projection of the object function onto the $\mathbf{n}$ direction ($\mathbf{n}^2 = 1$) which is perpendicular to the ray is given by the two-dimensional Radon transform

$$F(\mathbf{n},\, p) = \int f(\boldsymbol{\rho})\delta(p - \mathbf{n}\cdot\boldsymbol{\rho})\,\mathrm{d}^2\rho. \tag{21}$$

Let us introduce the two-dimensional Fourier transform $\tilde{f}(\boldsymbol{\kappa})$ of $f(\boldsymbol{\rho})$, ($\boldsymbol{\kappa}$ denoting a two-dimensional vector which is the analogue of the N-dimensional vector $\mathbf{K}$ in Fourier space),

$$\tilde{f}(\boldsymbol{\kappa}) = \int f(\boldsymbol{\rho})\mathrm{e}^{\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{\rho}}\,\mathrm{d}^2\rho, \tag{22}$$

and the one-dimensional Fourier transform of the Radon transform $F(\mathbf{n},\, p)$ with respect to the p variable, viz.

$$\tilde{F}(\mathbf{n},\, u) = \int_{-\infty}^{\infty} F(\mathbf{n},\, p)\mathrm{e}^{\mathrm{i}up}\,\mathrm{d}p. \tag{23}$$

Then Eq. (20), specialized to two dimensions, may be written in the compact form

$$\tilde{f}(u\mathbf{n}) = \tilde{F}(\mathbf{n},\, u). \tag{24}$$

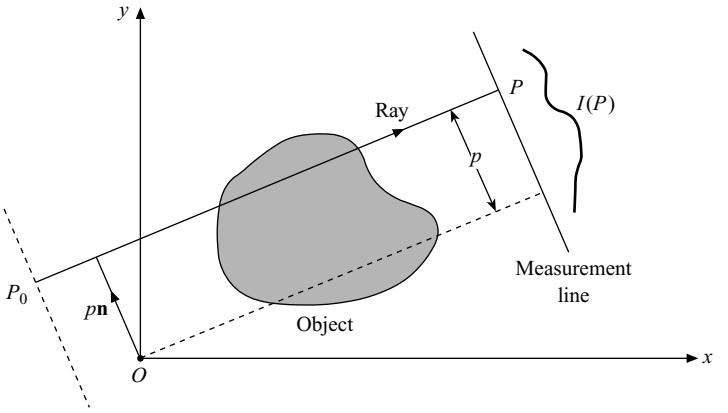

Fig. 4.48 Geometry and notation relating to the use of computerized tomography for determining a two-dimensional object function $f(\boldsymbol{\rho})$ from parallel projections.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

The function $\tilde{f}(u\mathbf{n})$ is often referred to as the slice along the $\mathbf{n}$ direction in the Fourier domain (the $\boldsymbol{\kappa}$-plane) of the object function $f(\boldsymbol{\rho})$. It is the set of values of $\tilde{f}(u\mathbf{n})$ for all u values, with the direction $\mathbf{n}$ being kept fixed. Formula (24) may then be said to imply that *the one-dimensional Fourier transform* $\tilde{F}(\mathbf{n}, u)$ *of the projection* $F(\mathbf{n}, p)$ *on* $\mathbf{n}$ *of the object function* $f(\boldsymbol{\rho})$ *is equal to the slice* $\tilde{f}(\boldsymbol{\kappa})$ *of the object, taken along the line* $\boldsymbol{\kappa} = u\mathbf{n}$ *through the origin in the Fourier domain*. This statement is often referred to as the *projection-slice theorem* or the Fourier slice theorem of computerized tomography.*

Making use of (24) we may express the object function $f(\boldsymbol{\rho})$ in terms of its parallel projections in a more explicit form, which is frequently used in practice. To do so we choose a fixed rectangular coordinate system with axes Ox_0, Oy_0 in the x,y-plane (the cross-section through the object) and denote by ϕ the angle which the unit normal $\mathbf{n}$ makes with the positive x_0-axis (see Fig. 4.49). Then

$$\mathbf{n} = \hat{\mathbf{x}}_0 \cos\phi + \hat{\mathbf{y}}_0 \sin\phi, \tag{25}$$

where $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{y}}_0$ are unit vectors along the x_0 and y_0 directions respectively, and

$$\tilde{\mathrm{f}}(u\mathbf{n}) \equiv \tilde{f}(u\cos\phi,\ u\sin\phi). \tag{26}$$

The Fourier representation of $f(\boldsymbol{\rho})$ (the Fourier inverse) of (22) may be written in the form

$$f(\boldsymbol{\rho}) = \frac{1}{(2\pi)^2}\int_0^{2\pi} \mathrm{d}\phi \int_0^{\infty} \tilde{f}(u\cos\phi,\ u\sin\phi)\mathrm{e}^{-\mathrm{i}u\mathbf{n}\cdot\boldsymbol{\rho}} u\,\mathrm{d}u. \tag{27}$$

Using the relations $\cos(\phi+\pi) = -\cos\phi$, $\sin(\phi+\pi) = -\sin\phi$ and substituting for $\tilde{f}$ on the right-hand side from the basic relation (24), (27) becomes

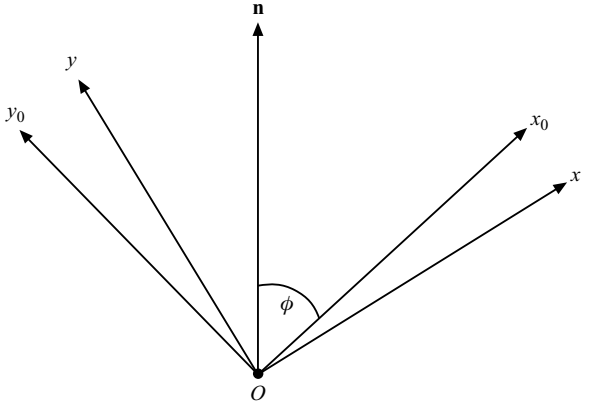

Fig. 4.49 Illustrating the notation relating to reconstruction of planar cross-sections of the object function $f(\boldsymbol{\rho})$.

* A. C. Kak in *Array Signal Processing*, S. Haykin, ed. (Englewood Cliffs, NJ, Prentice-Hall, 1985), pp. 372–374.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

$$f(\boldsymbol{\rho}) = \frac{1}{(2\pi)^2}\int_0^{\pi} \mathrm{d}\phi \int_{-\infty}^{\infty} \tilde{F}(\mathbf{n},\, u)\mathrm{e}^{-\mathrm{i}u\mathbf{n}\cdot\boldsymbol{\rho}}|u|\,\mathrm{d}u, \tag{28}$$

where $\mathbf{n}$ is given by (25).

Formula (28) is customarily used to determine the unknown object function $f(\boldsymbol{\rho})$ throughout the $\boldsymbol{\rho}$-plane from the one-dimensional Fourier transforms $\tilde{F}(\mathbf{n}, u)$ of the projections $F(\mathbf{n}, p)$ of $f(\boldsymbol{\rho})$. This formula is the basis of the so-called *back-projection algorithm for parallel projections*.

The credit for originating the field of computerized tomography goes largely to A. Cormack and G. Hounsfield. In 1963 Cormack* described a mathematical algorithm for reconstruction of objects from line integrals and also published results obtained with a scanner which used a computer to reconstruct images of laboratory test objects. The first tomographic scanner for medical use was built under the direction of Hounsfield† around 1970 and the first patient, suspected to have a brain tumor, was diagnosed with it in 1971. The reconstruction confirmed the presence of the tumor. In the reconstruction 180 projections were taken, each consisting of 160 ray integrals, providing $160 \times 180 = 28{,}800$ data. Since then the subject of computerized tomography has developed into a large technological field. Many types of scanners have been introduced and numerous schemes for processing the large volume of data have been proposed. When this set of data is processed, images with resolution of the order of 1 mm^2 may be obtained with X-rays in the 20–150 keV energy range. An example of tomographic reconstruction is given in Fig. 4.50. In spite of the large volume of data

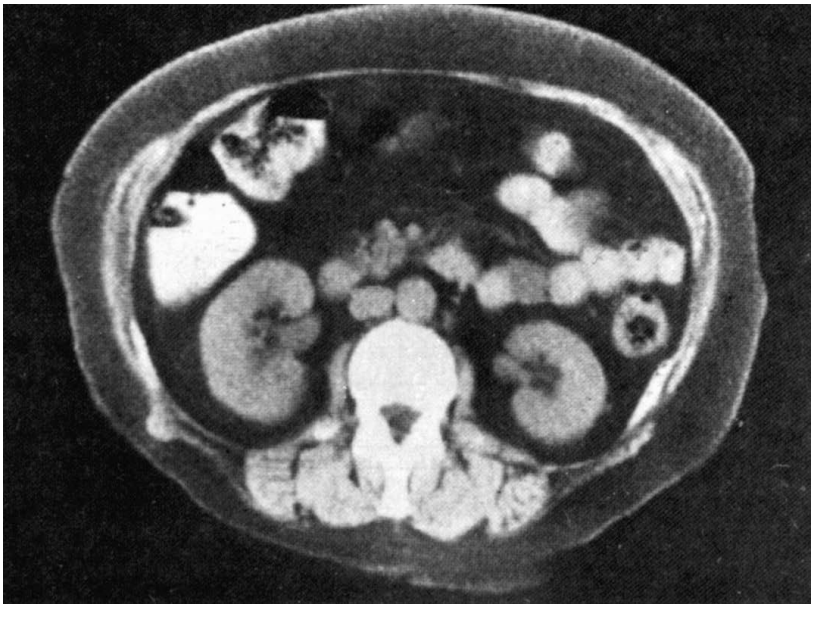

Fig. 4.50 CT scan taken through the kidneys. From the 1979 Nobel lecture by G. N. Hounsfield [*Nobel Lectures, Physiology or Medicine, 1971–1980*, J. Lindsten, ed. (World Scientific, Singapore, 1992), p. 569].

* A. Cormack, *J. Appl. Phys.*, **34** (1963), 2722. † G. N. Hounsfield, *British J. Radiol.*, **46** (1973), 1016.

https://doi.org/10.1017/9781108769914.007 Published online by Cambridge University Press

collected in such measurements, the total dosage of X-rays is typically lower than that used in conventional radiology.

As we have already mentioned, computerized tomography is used today in many different fields and also with different types of radiation and even with elementary particles. It has been applied in investigations to determine the geological composition under the Earth's surface required for oil exploration and for the detection of hazardous regions in the excavation of mines. It is also used for nondestructive testing of materials, in holographic interferometry for determining the three-dimensional distribution of the refractive index generated by temperature variations in gases, etc.

Medical applications abound. Computerized tomography with acoustical waves is used to detect cancerous tumors in women's breasts. In nuclear medicine radioactive isotopes are introduced inside the patient's body and the distribution of radioactivity inside various organs is then determined by tomographic methods. The reconstruction provides maps of the distribution of molecules in the organ under examination*.

Computerized tomographic reconstructions in electron microscopy were responsible for basic discoveries in molecular biology and in the determination of the structure of viruses. In the pioneering researches by A. Klug and his collaborators† a resolution of the order of 30 Å was achieved.

The importance of computerized tomography may be judged from the fact that several Nobel prizes have been awarded for its invention and for discoveries made with it: In 1979 the Nobel prize in physiology and medicine was awarded jointly to A. M. Cormack and G. N. Hounsfield and the 1982 Nobel prize in chemistry was awarded to A. Klug.

Finally we mention that there was a precursor to medical tomography in quite a different field. In 1956 Bracewell‡ obtained one-dimensional projections of the two-dimensional radio sky from which he determined the brightness distribution across a radio source.

* A historical account of the development of computerized tomography in the context of medical diagnostics is given in S. Webb, *From the Watching of Shadows* (Bristol, Adam Hilger, 1990).

† D. J. de Rosier and A. Klug, *Nature*, **217** (1968), 130; R. A. Crowther, D. J. de Rosier and A. Klug, *Proc. Roy. Soc. Lond.*, **A 317** (1970), 319; A. Klug and R. A. Crowther, *Nature*, **238** (1972), 435.

‡ R. N. Bracewell, *Austral. J. Phys.*, **9** (1956), 198. See also R. N. Bracewell and A. C. Riddle, *Astroph. J.*, **150** (1967), 427.