FOUNDATIONS OF GEOMETRICAL OPTICS

3.1 APPROXIMATION FOR VERY SHORT WAVELENGTHS

The electromagnetic field associated with the propagation of visible light is characterized by very rapid oscillations (frequencies of the order of $10^{14}\,\text{sec}^{-1}$) or, what amounts to the same thing, by the smallness of the wavelength (of order 10^{-5} cm). It may therefore be expected that a good first approximation to the propagation laws in such cases may be obtained by a complete neglect of the finiteness of the wavelength. It is found that for many optical problems such a procedure is entirely adequate; in fact, phenomena which can be attributed to departures from this approximate theory (so-called diffraction phenomena, studied in Chapter VIII) can only be demonstrated by means of carefully conducted experiments.

The branch of optics which is characterized by the neglect of the wavelength, i.e. that corresponding to the limiting case $\lambda_0 \rightarrow 0$, is known as *geometrical optics*,* since in this approximation the optical laws may be formulated in the language of geometry. The energy may then be regarded as being transported along certain curves (light rays). A physical model of a pencil of rays may be obtained by allowing the light from a source of negligible extension to pass through a very small opening in an opaque screen. The light which reaches the space behind the screen will fill a region the boundary of which (the edge of the pencil) will, at first sight, appear to be sharp. A more careful examination will reveal, however, that the light intensity near the boundary varies rapidly but continuously from darkness in the shadow to lightness in the illuminated region, and that the variation is not monotonic but is of an oscillatory character, manifested by the appearance of bright and dark bands, called diffraction fringes. The region in which this rapid variation takes place is only of the order of magnitude of the wavelength. Hence, as long as this magnitude is neglected in comparison with the dimensions of the opening, we may speak of a sharply bounded pencil of rays.† On reducing the size of the opening down to the dimensions of the wavelength phenomena appear which need more refined study. If, however, one considers only the limiting case of negligible wavelengths, no restriction on the size of the opening is imposed, and we may say that an opening of vanishingly small dimensions defines an infinitely thin pencil—the light ray. It will be shown that the variation in the cross-section of a pencil of rays is a measure of the variation of the intensity of the light. Moreover it will be seen that it is possible to associate a state of polarization with each ray, and to study its variation along the ray.

---

* The historical development of geometrical optics is described by M. Herzberger, *Strahlenoptik* (Berlin, Springer, 1931), 179; *Z. Instrumentenkunde*, **52** (1932), 429–435, 485–493, 534–542, C. Carathéodory, *Geometrische Optik* (Berlin, Springer, 1937) and E. Mach, *The Principles of Physical Optics, A Historical and Philosophical Treatment* (First German edition 1913, English translation: London, Methuen, 1926; reprinted by Dover Publications, New York, 1953).

† That the boundary becomes sharp in the limit as $\lambda_0 \rightarrow 0$ was first shown by G. Kirchhoff, *Vorlesungen ü. Math. Phys.*, **2** (*Mathematische Optik*), (Leipzig, Teubner, 1891), p. 33. See also B. B. Baker and E. T. Copson, *The Mathematical Theory of Huygens' Principle* (Oxford, Clarendon Press, 2nd edition, 1950), p. 79, and A. Sommerfeld, *Optics* (New York, Academic Press, 1954), § 35.

Further it will be seen that for small wavelengths the field has the same general character as that of a plane wave, and, moreover, that within the approximation of geometrical optics the laws of refraction and reflection established for plane waves incident upon a plane boundary remain valid under more general conditions. Hence if a light ray falls on a sharp boundary (e.g. the surface of a lens) it is split into a reflected ray and a transmitted ray, and the changes in the state of polarization as well as the reflectivity and transmissivity may be calculated from the corresponding formulae for plane waves.

The preceding remarks imply that, when the wavelength is small enough, the sum total of optical phenomena may be deduced from geometrical considerations, by determining the paths of the light rays and calculating the associated intensity and polarization. We shall now formulate the appropriate laws by considering the implications of MAXWELL's equations when $\lambda_0 \to 0$.*

3.1.1 Derivation of the eikonal equation

We consider a general time-harmonic field

$$\left.\begin{aligned} \boldsymbol{E}(\boldsymbol{r}, t) &= \boldsymbol{E}_0(\boldsymbol{r})e^{-i\omega t}, \\ \boldsymbol{H}(\boldsymbol{r}, t) &= \boldsymbol{H}_0(\boldsymbol{r})e^{-i\omega t}, \end{aligned}\right\} \tag{1}$$

in a non-conducting isotropic medium. $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ denote complex vector functions of positions, and, as explained in § 1.4.3, the real parts of the expressions on the right-hand side of (1) are understood to represent the fields.

The complex vectors $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ will satisfy MAXWELL's equations in their time-free form, obtained on substituting (1) into (1)–(4) of § 1.1. In regions free of currents and charges ($\boldsymbol{j} = \rho = 0$), these equations are

$$\operatorname{curl} \boldsymbol{H}_0 + ik_0\varepsilon \boldsymbol{E}_0 = 0, \tag{2}$$

$$\operatorname{curl} \boldsymbol{E}_0 - ik_0\mu \boldsymbol{H}_0 = 0, \tag{3}$$

$$\operatorname{div} \varepsilon \boldsymbol{E}_0 = 0, \tag{4}$$

$$\operatorname{div} \mu \boldsymbol{H}_0 = 0. \tag{5}$$

Here the material relations $\boldsymbol{D} = \varepsilon \boldsymbol{E}$, $\boldsymbol{B} = \mu \boldsymbol{H}$ have been used and, as before, $k_0 = \omega/c = 2\pi/\lambda_0$, λ_0 being the vacuum wavelength.

We have seen that a homogeneous plane wave in a medium of refractive index $n = \sqrt{\varepsilon\mu}$, propagated in the direction specified by the unit vector $\boldsymbol{s}$, is represented by

$$\boldsymbol{E}_0 = \boldsymbol{e}e^{ik_0 n(\boldsymbol{s}\cdot\boldsymbol{r})}, \qquad \boldsymbol{H}_0 = \boldsymbol{h}e^{ik_0 n(\boldsymbol{s}\cdot\boldsymbol{r})}, \tag{6}$$

where $\boldsymbol{e}$ and $\boldsymbol{h}$ are constant, generally complex vectors. For a (monochromatic) electric dipole field in the vacuum we found (cf. § 2.2) that

$$\boldsymbol{E}_0 = \boldsymbol{e}e^{ik_0 r}, \qquad \boldsymbol{H}_0 = \boldsymbol{h}e^{ik_0 r}, \tag{7}$$

---

* It was first shown by A. SOMMERFELD and J. RUNGE, *Ann. d. Physik*, **35** (1911), 289, using a suggestion of P. DEBYE, that the basic equation of geometrical optics (the eikonal equation (15b)) may be derived from the (scalar) wave equation in the limiting case $\lambda_0 \to 0$. Generalizations which take into account the vectorial character of the electromagnetic field are due to W. IGNATOWSKY, *Trans. State Opt. Institute (Petrograd)*, **1** (1919), III; V. A. FOCK, *ibid.*, **3** (1924), 3; S. M. RYTOV, *Compt. Rend. (Doklady) Acad. Sci. URSS*, **18** (1938), 263; N. ARLEY, *Det. Kgl. Danske Videns Selsk.*, **22** (1945), No. 8; F. G. FRIEDLANDER, *Proc. Cambr. Phil. Soc.*, **43** (1947), 284; K. SUCHY, *Ann. d. Physik.*, **11** (1952), 113, *ibid.*, **12** (1953), 423, and *ibid.*, **13** (1953), 178; R. S. INGARDEN and A. KRZYWICKI, *Acta Phys. Polonica*, **14** (1955), 255.

suitable normalization of the dipole moment, independent of k_0.

These examples suggest that in regions which are many wavelengths away from the sources we may represent more general types of fields in the form

$$E_0 = e(r)e^{ik_0\mathscr{S}(r)}, \qquad H_0 = h(r)e^{ik_0\mathscr{S}(r)}, \tag{8}$$

where $\mathscr{S}(r)$, "the optical path", is a real scalar function of position, and $e(r)$ and $h(r)$ are vector functions of position, which may in general be complex.* With (8) as trial solution, MAXWELL's equations lead to a set of relations between e, h and $\mathscr{S}$. It will be shown that for large k_0 (small λ_0) these relations demand that $\mathscr{S}$ should satisfy a certain differential equation, which is independent of the amplitude vectors e and h.

From (8), using well-known vector identities,

$$\text{curl } H_0 = (\text{curl } h + ik_0 \text{ grad } \mathscr{S} \wedge h)e^{ik_0\mathscr{S}}, \tag{9}$$

$$\text{div } \mu H_0 = (\mu \text{ div } h + h \,.\, \text{grad } \mu + ik_0\mu h \,.\, \text{grad } \mathscr{S})e^{ik_0\mathscr{S}}, \tag{10}$$

with similar expressions for curl E_0 and div εE_0. Hence (2)–(5) become

$$\text{grad } \mathscr{S} \wedge h + \varepsilon e = -\frac{1}{ik_0} \text{curl } h, \tag{11}$$

$$\text{grad } \mathscr{S} \wedge e - \mu h = -\frac{1}{ik_0} \text{curl } e, \tag{12}$$

$$e \,.\, \text{grad } \mathscr{S} = -\frac{1}{ik_0}(e \,.\, \text{grad } \log \varepsilon + \text{div } e), \tag{13}$$

$$h \,.\, \text{grad } \mathscr{S} = -\frac{1}{ik_0}(h \,.\, \text{grad } \log \mu + \text{div } h). \tag{14}$$

We are interested in the solution for very large values of k_0. Hence as long as the multiplicative factors of $1/ik_0$ on the right-hand side are not exceptionally large they may be neglected, and the equations then reduce to

$$\text{grad } \mathscr{S} \wedge h + \varepsilon e = 0, \tag{11a}$$

$$\text{grad } \mathscr{S} \wedge e - \mu h = 0, \tag{12a}$$

$$e \,.\, \text{grad } \mathscr{S} = 0, \tag{13a}$$

$$h \,.\, \text{grad } \mathscr{S} = 0. \tag{14a}$$

We can confine our attention to equations (11a) and (12a) alone, since (13a) and (14a) follow from them on scalar multiplication with grad $\mathscr{S}$. Now (11a) and (12a) may be regarded as a set of six simultaneous homogeneous linear scalar equations for the Cartesian components e_x, h_x, . . . of e and h. These simultaneous equations have non-trivial solutions only if a consistency condition (the vanishing of the associated

---

* Complex e and h are necessary, if all possible states of polarization are to be included. According to § 1.4 (75) real e and h correspond to fields which are linearly polarized.

r being the distance from the dipole. Here e and h are no longer constant vectors, but at distances sufficiently far away from the dipole ($r \gg \lambda_0$) these vectors are. With determinant) is satisfied. This condition may be obtained simply by eliminating $\boldsymbol{e}$ or $\boldsymbol{h}$ between (11a) and (12a). Substituting for $\boldsymbol{h}$ from (12a), (11a) becomes

$$\frac{1}{\mu}[(\boldsymbol{e} \cdot \operatorname{grad} \mathscr{S}) \operatorname{grad} \mathscr{S} - \boldsymbol{e}(\operatorname{grad} \mathscr{S})^2] + \varepsilon \boldsymbol{e} = 0.$$

The first term vanishes on account of (13a), and the equation then reduces, since $\boldsymbol{e}$ does not vanish everywhere, to

$$(\operatorname{grad} \mathscr{S})^2 = n^2, \tag{15a}$$

or, written explicitly,

$$\left(\frac{\partial \mathscr{S}}{\partial x}\right)^2 + \left(\frac{\partial \mathscr{S}}{\partial y}\right)^2 + \left(\frac{\partial \mathscr{S}}{\partial z}\right)^2 = n^2(x, y, z), \tag{15b}$$

where as before $n = \sqrt{\varepsilon\mu}$ denotes the refractive index. The function $\mathscr{S}$ is often called the *eikonal** and (15b) is known as the *eikonal equation*; it is the basic equation of geometrical optics.† The surfaces

$$\mathscr{S}(\boldsymbol{r}) = \text{constant}$$

may be called the *geometrical wave surfaces* or the *geometrical wave-fronts*.‡

The eikonal equation was derived here by using the first-order MAXWELL's equations, but it may also be derived from the second-order wave equations for the electric or magnetic field vectors. To show this one substitutes from (1) and (8) into the wave equation § 1.2 (5) and obtains, after a straightforward calculation,

$$\mathbf{K}(\boldsymbol{e}, \mathscr{S}, n) + \frac{1}{ik_0}\mathbf{L}(\boldsymbol{e}, \mathscr{S}, n, \mu) + \frac{1}{(ik_0)^2}\mathbf{M}(\boldsymbol{e}, \varepsilon, \mu) = 0, \tag{16}$$

where

$$\mathbf{K}(\boldsymbol{e}, \mathscr{S}, n) = \{n^2 - (\operatorname{grad} \mathscr{S})^2\}\boldsymbol{e},$$

$$\mathbf{L}(\boldsymbol{e}, \mathscr{S}, n, \mu) = \{\operatorname{grad} \mathscr{S} \cdot \operatorname{grad} \log \mu - \nabla^2 \mathscr{S}\}\boldsymbol{e} - 2\{\boldsymbol{e} \cdot \operatorname{grad} \log n\} \operatorname{grad} \mathscr{S} \\ - 2\{\operatorname{grad} \mathscr{S} \cdot \operatorname{grad}\}\boldsymbol{e},$$

$$\mathbf{M}(\boldsymbol{e}, \varepsilon, \mu) = \operatorname{curl} \boldsymbol{e} \wedge \operatorname{grad} \log \mu - \nabla^2 \boldsymbol{e} - \operatorname{grad} (\boldsymbol{e} \cdot \operatorname{grad} \log \varepsilon).$$

The corresponding equation involving $\boldsymbol{h}$, obtained on substitution into the wave equation (6) in § 1.2 for $\boldsymbol{H}$ (or more simply by using the fact that MAXWELL's equations remain unchanged when $\boldsymbol{E}$ and $\boldsymbol{H}$ and simultaneously ε and $-\mu$ are interchanged), is

$$\mathbf{K}(\boldsymbol{h}, \mathscr{S}, n) + \frac{1}{(ik_0)}\mathbf{L}(\boldsymbol{h}, \mathscr{S}, n, \varepsilon) + \frac{1}{(ik_0)^2}\mathbf{M}(\boldsymbol{h}, \mu, \varepsilon) = 0. \tag{17}$$

---

* The term *eikonal* (from Greek εικῶν = image) was introduced in 1895 by H. BRUNS to describe certain related functions (cf. p. 133), but has come to be used in a wider sense.

† The eikonal equation may also be regarded as the equation of the characteristics of the wave equations (5) and (6), in § 1.2, for $\boldsymbol{E}$ and $\boldsymbol{H}$, and describes the propagation of discontinuities of the solutions of these equations. In geometrical optics we are, however, not concerned with the propagation of discontinuities but with time-harmonic (or nearly time-harmonic) solutions. The formal equivalence of the two interpretations is demonstrated in Appendix VI.

The eikonal equation will also be recognized as the HAMILTON–JACOBI equation of the variational problem $\delta\int n\,ds = 0$, the optical counterpart of which goes back to FERMAT (cf. § 3.3.2 and ...

It may be shown that in many cases of importance the spatial parts E_0 and H_0 of the field vectors may be developed into asymptotic series of the form*

$$E_0 = e^{ik_0\mathscr{S}} \sum_{m\geqslant 0} \frac{e^{(m)}}{(ik_0)^m}, \qquad H_0 = e^{ik_0\mathscr{S}} \sum_{m\geqslant 0} \frac{h^{(m)}}{(ik_0)^m}, \tag{18}$$

where $e^{(m)}$ and $h^{(m)}$ are functions of position, and $\mathscr{S}$ is the same function as before.† Geometrical optics corresponds to the leading terms of these expansions.

3.1.2 The light rays and the intensity law of geometrical optics

From (8), and from (54) and (55) in § 1.4, it follows that the time averages of the electric and magnetic energy densities $\langle w_e \rangle$ and $\langle w_m \rangle$ are given by

$$\langle w_e \rangle = \frac{\varepsilon}{16\pi} \mathbf{e} \cdot \mathbf{e}^\star, \qquad \langle w_m \rangle = \frac{\mu}{16\pi} \mathbf{h} \cdot \mathbf{h}^\star. \tag{19}$$

Substitution for $\mathbf{e}^\star$ from (11a) and for $\mathbf{h}$ from (12a) gives

$$\langle w_e \rangle = \langle w_m \rangle = \frac{1}{16\pi} [\mathbf{e}, \mathbf{h}^\star, \operatorname{grad} \mathscr{S}], \tag{20}$$

the square bracket denoting the scalar triple product. Hence, *within the accuracy of geometrical optics, the time-averaged electric and magnetic energy densities are equal.*

The time average of the Poynting vector is obtained from (8) and § 1.4 (52):

$$\langle \mathbf{S} \rangle = \frac{c}{8\pi} \mathscr{R}(\mathbf{e} \wedge \mathbf{h}^\star).$$

Using (12a), we obtain

$$\langle \mathbf{S} \rangle = \frac{c}{8\pi\mu} \{(\mathbf{e} \cdot \mathbf{e}^\star) \operatorname{grad} \mathscr{S} - (\mathbf{e} \cdot \operatorname{grad} \mathscr{S})\mathbf{e}^\star\}.$$

The last term vanishes on account of (13a) so that we have, if use is made of the expression for $\langle w_e \rangle$ and of MAXWELL's relation $\varepsilon\mu = n^2$,

$$\langle \mathbf{S} \rangle = \frac{2c}{n^2} \langle w_e \rangle \operatorname{grad} \mathscr{S}. \tag{21}$$

---

* We assume here that only one geometrical wave-front passes through each point. In some cases, for example when reflection takes place at obstacles present in the medium, several wave-fronts may pass through each point. The resulting field is then represented by the addition of series of the above type.

† The theory of such asymptotic expansions has its origin chiefly in the work of R. K. LUNEBURG, *Propagation of Electromagnetic Waves* (mimeographed lecture notes, New York University, 1947–1948). See also M. KLINE, *Comm. Pure and Appl. Math.*, **4** (1951), 225; *ibid.*, **8** (1955), 595 and W. BRAUNBECK, *Z. Naturforsch.*, **6** (1951), 672. A comprehensive account of the theory is given in M. KLINE and I. W. KAY, *Electromagnetic Theory and Geometrical Optics* [New York, Interscience Publishers, (1965)].

For sufficiently large K_0 the second and third terms may in general be neglected; then $K = 0$, giving again the eikonal equation. It will be seen later that the terms in the first power of $1/(iK_0)$ in (16) and (17) also prossess a physical interpretation

Since $\langle w_e \rangle = \langle w_m \rangle$, the term $2\langle w_e \rangle$ represents the time average $\langle w \rangle$ of the total energy density (i.e. $\langle w \rangle = \langle w_e \rangle + \langle w_m \rangle$). Also, on account of the eikonal equation, (grad $\mathscr{S}$)/n is a unit vector ($\boldsymbol{s}$ say),

$$\boldsymbol{s} = \frac{\operatorname{grad} \mathscr{S}}{n} = \frac{\operatorname{grad} \mathscr{S}}{|\operatorname{grad} \mathscr{S}|}, \tag{22}$$

and (21) shows that $\boldsymbol{s}$ is in the direction of the average Poynting vector. If, as before, we set $c/n = v$, (21) becomes

$$\langle \boldsymbol{S} \rangle = v \langle w \rangle \boldsymbol{s}. \tag{23}$$

Hence *the average Poynting vector is in the direction of the normal to the geometrical wave-front, and its magnitude is equal to the product of the average energy density and the velocity $v = c/n$.* This result is analogous to the relation (9) in § 1.4 for plane waves,

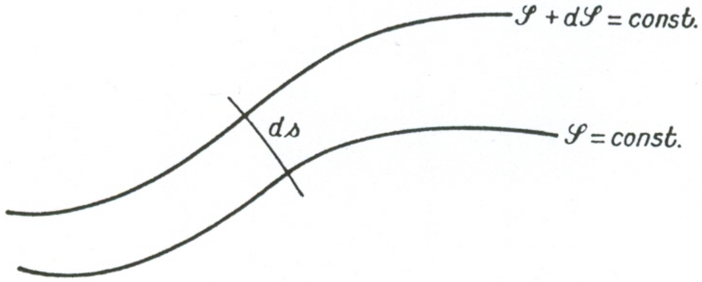

Fig. 3.1. Illustrating the meaning of the relation $n\boldsymbol{s} = \operatorname{grad} \mathscr{S}$.

and shows that *within the accuracy of geometrical optics the average energy density is propagated with the velocity $v = c/n$.*

The *geometrical light rays* may now be defined as the orthogonal trajectories to the geometrical wave-fronts $\mathscr{S}$ = constants. We shall regard them as oriented curves whose direction coincides everywhere with the direction of the average Poynting vector.* If $\boldsymbol{r}(s)$ denotes the position vector of a point P on a ray, considered as a function of the length of arc s of the ray, then $d\boldsymbol{r}/ds = \boldsymbol{s}$, and the equation of the ray may be written as

$$n \frac{d\boldsymbol{r}}{ds} = \operatorname{grad} \mathscr{S}. \tag{24}$$

From (13a) and (14a) it is seen that *the electric and magnetic vectors are at every point orthogonal to the ray.*

The meaning of (24) may be made clearer from the following remarks. Consider two neighbouring wave-fronts $\mathscr{S}$ = constant and $\mathscr{S} + d\mathscr{S}$ = constant (Fig. 3.1). Then

$$\frac{d\mathscr{S}}{ds} = \frac{d\boldsymbol{r}}{ds} \cdot \operatorname{grad} \mathscr{S} = n. \tag{25}$$

Hence the distance ds between points on the opposite ends of a normal cutting the two wave-fronts is inversely proportional to the refractive index, i.e. directly proportional to v.

* This definition of light rays is appropriate for isotropic media only.

P115

... square brackets the optical length of the ray which joins points P_1 and P_2 we have

$$[P_1P_2] = \int_{P_1}^{P_2} n ds = \mathscr{S}(P_2) - \mathscr{S}(P_1). \tag{26}$$

Since, as we have seen, the average energy density is propagated with the velocity $v = c/n$ along the ray,

$$n ds = \frac{c}{v} ds = c dt,$$

where dt is the time needed for the energy to travel the distance ds along the ray; hence

$$[P_1P_2] = c\int_{P_1}^{P_2} dt, \tag{27}$$

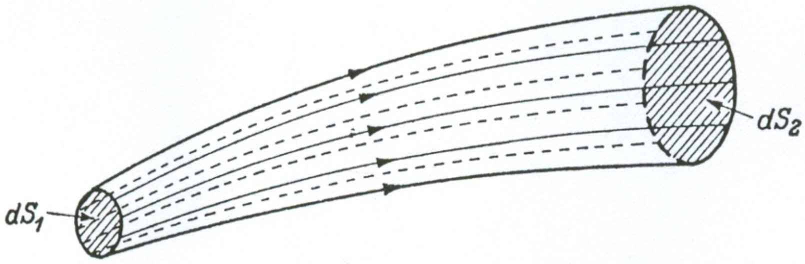

Fig. 3.2. Illustrating the intensity law of geometrical optics.

i.e. the optical length $[P_1P_2]$ is equal to the product of the vacuum velocity of light and the time needed for light to travel from P_1 to P_2.

The intensity of light I was defined as the absolute value of the time average of the Poynting vector. We therefore have from (23),

$$I = |\langle \mathbf{S} \rangle| = v\langle w \rangle, \tag{28}$$

and the conservation law § 1.4 (57) gives

$$\operatorname{div} (I\mathbf{s}) = 0. \tag{29}$$

To see the implications of this relation we take a narrow tube formed by all the rays proceeding from an element dS_1 of a wave-front $\mathscr{S}(\mathbf{r}) = a_1$ (a_1 being a constant), and denote by dS_2 the corresponding element in which these rays intersect another wave-front $\mathscr{S}(\mathbf{r}) = a_2$ (Fig. 3.2). Integrating (29) throughout the tube and applying Gauss' theorem we obtain

$$\int I\mathbf{s}.\boldsymbol{\nu}\, dS = 0, \tag{30}$$

$\boldsymbol{\nu}$ denoting the outward normal to the tube. Now

$$\begin{aligned} \mathbf{s}\,.\,\boldsymbol{\nu} &= \quad 1 \text{ on } dS_2, \\ &= -1 \text{ on } dS_1, \\ &= \quad 0 \text{ elsewhere,} \end{aligned}$$

so that (30) reduces to

$$I_1 dS_1 = I_2 dS_2, \tag{31}$$

The integral $\int_C n ds$ taken along a curve C is know as the opitical length of the curve. Denoting by square brackets the optical length of the ray which joins points P_1 and P_2 we have

I_1 and I_2 denoting the intensity on dS_1 and on dS_2 respectively. Hence *IdS remain constant along a tube of rays*. This result expresses *the intensity law of geometrica optics*.

We shall see later that in a homogeneous medium the rays are straight lines. Th intensity law may then be expressed in a somewhat different form. Assume first tha dS_1, and consequently also dS_2, are bounded by segments of lines of curvature (se Fig. 3.3). If R_1 and R_1' are the principal radii of curvature (cf. § 4.6.1) of the segment A_1B_1 and B_1C_1, then

$$A_1B_1 = R_1 d\theta, \qquad B_1C_1 = R_1' d\phi,$$

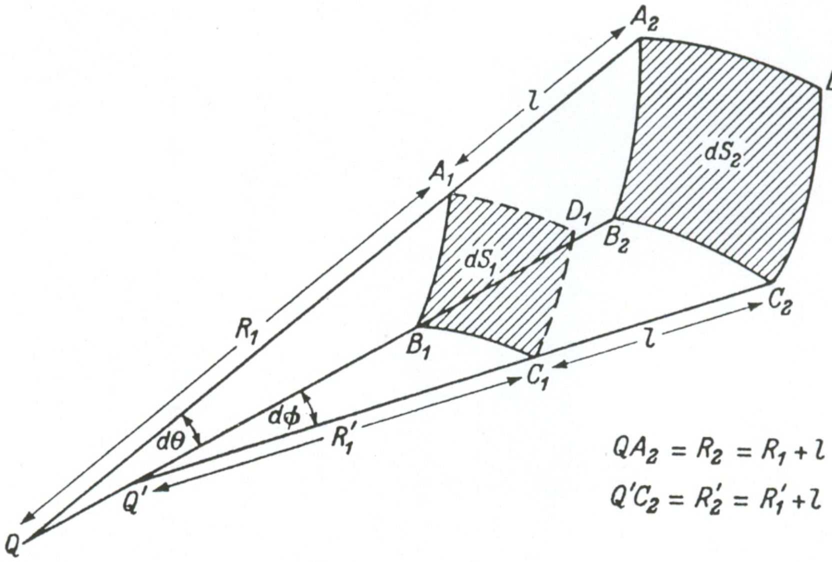

Fig. 3.3. Illustrating the intensity law of geometrical optics for rectilinear rays.

where $d\theta$ and $d\phi$ are the angles which A_1B_1 and B_1C_1 subtend at the respective centres of curvature Q and Q'. Hence

$$dS_1 = A_1B_1 \,.\, B_1C_1 = R_1R_1' d\theta d\phi; \tag{32}$$

similarly for an element dS_2 in which the bundle of rays through dS_1 meets another wave-front of the family,

$$dS_2 = A_2B_2 \,.\, B_2C_2 = R_2R_2' d\theta d\phi. \tag{33}$$

If l is the distance between dS_1 and dS_2 measured along the rays, then

$$R_2 = R_1 + l, \qquad R_2' = R_1' + l,$$

and it follows that

$$\frac{I_2}{I_1} = \frac{dS_1}{dS_2} = \frac{R_1R_1'}{R_2R_2'} = \frac{R_1R_1'}{(R_1+l)(R_1'+l)}. \tag{34}$$

If the areas dS_1 and dS_2 are bounded by arbitrary curves, (34) still holds. This can immediately be seen if we regard them as made up of a number of elements bounded by lines of curvature, and sum their contributions.

If $R_1 \ll l$, $R_1' \ll l$, (34) reduces to

$$\frac{I_2}{I_1} = \frac{R_1R_1'}{l^2}. \tag{35}$$

This formula is sometimes used in connection with the scattering of radiation.

The reciprocal $1/RR'$ of the product of the two principal radii of curvature is called the *Gaussian curvature* (or second curvature) of the surface. (34) shows that *the* *intensity at any point of a rectilinear ray is proportional to the Gaussian curvature of the wave-front which passes through that point.* In particular if all the (rectilinear) rays have a point in common, the wave-fronts are spheres centred at that point; then $R_1 = R_1'$, $R_2 = R_2'$ and we obtain (dropping the suffixes) the *inverse square law*

$$I = \frac{\text{constant}}{R^2}. \tag{36}$$

Returning to the general case of an arbitrary pencil of rays (curved or straight), we can write down an explicit expression in terms of the $\mathscr{S}$ function for the variation of the intensity along each ray. Substituting for $\mathbf{s}$ from (22) into (29), and using the identities div $u\mathbf{v} = u$ div $\mathbf{v} + \mathbf{v}$. grad u, and div grad $= \nabla^2$, we obtain

$$\frac{I}{n}\nabla^2\mathscr{S} + \operatorname{grad}\mathscr{S} \cdot \operatorname{grad}\frac{I}{n} = 0.$$

This may also be written as

$$\nabla^2\mathscr{S} + \operatorname{grad}\mathscr{S} \cdot \operatorname{grad}\log\frac{I}{n} = 0. \tag{37}$$

Let us now introduce the operator

$$\frac{\partial}{\partial\tau} = \operatorname{grad}\mathscr{S} \cdot \operatorname{grad}, \tag{38}$$

where τ is a parameter which specifies position along the ray. Then (37) may be written as

$$\frac{\partial}{\partial\tau}\log\frac{I}{n} = -\nabla^2\mathscr{S}$$

whence, on integration,

$$I = n e^{-\int^{\tau}\nabla^2\mathscr{S}d\tau}$$

But by (38), (15), and (25),

$$d\tau = \frac{d\mathscr{S}}{(\operatorname{grad}\mathscr{S})^2} = \frac{1}{n^2}d\mathscr{S} = \frac{1}{n}ds, \tag{39}$$

so that we finally obtain the following expressions for the ratio of the intensities at any two points of a ray:

$$\frac{I_2}{I_1} = \frac{n_2}{n_1}e^{-\int_{\mathscr{S}_1}^{\mathscr{S}_2}\frac{\nabla^2\mathscr{S}}{n^2}d\mathscr{S}} = \frac{n_2}{n_1}e^{-\int_{s_1}^{s_2}\frac{\nabla^2\mathscr{S}}{n}ds}, \tag{40}$$

the integrals being taken along the ray*.

3.1.3 Propagation of the amplitude vectors

We have seen that, when the wavelength is sufficiently small, the transport of energy may be represented by means of a simple hydrodynamical model which may be completely described in terms of the real scalar function $\mathscr{S}$, this function being a solution of the eikonal equation (15). According to traditional terminology, one understands by geometrical optics this approximate picture of energy propagation, using the concept of rays and wave-fronts. In other words polarization properties

---

* It has been shown by M. Kline, *Comm. Pure and Appl. Maths.*, **14** (1961), 473 that the intensity ratio (40) may be expressed in terms of an integral which involves the principal radii of curvature of the associated wavefronts. Kline's formula is a natural generalization, to inhomogeneous media, of the formula (34). See also M. Kline and I. W. Kay, *Electromagnetic Theory and Geometrical Optics* (New York, Interscience Publishers, 1965), p. 184.

are excluded. The reason for this restriction is undoubtedly due to the fact that the simple laws of geometrical optics concerning rays and wave-fronts were known from experiments long before the electromagnetic theory of light was established. It is however possible, and from our point of view quite natural, to extend the meaning of geometrical optics to embrace also certain geometrical laws relating to the propagation of the "amplitude vectors" $\boldsymbol{e}$ and $\boldsymbol{h}$. These laws may be easily deduced from the wave equations (16)–(17).

Since $\mathscr{S}$ satisfies the eikonal equation, it follows that $\mathbf{K} = 0$, and we see that when k_0 is sufficiently large (λ_0 small enough), only the $\mathbf{L}$-terms need to be retained in (16) and (17). Hence, in the present approximation, the amplitude vectors and the eikonal are connected by the relations $\mathbf{L} = 0$. If we use again the operator $\partial/\partial\tau$ introduced by (38), the equations $\mathbf{L} = 0$ become

$$\frac{\partial \boldsymbol{e}}{\partial \tau} + \frac{1}{2}\left(\nabla^2\mathscr{S} - \frac{\partial \log \mu}{\partial \tau}\right)\boldsymbol{e} + (\boldsymbol{e} \,.\, \text{grad} \log n) \,\text{grad}\, \mathscr{S} = 0, \tag{41}$$

$$\frac{\partial \boldsymbol{h}}{\partial \tau} + \frac{1}{2}\left(\nabla^2\mathscr{S} - \frac{\partial \log \varepsilon}{\partial \tau}\right)\boldsymbol{h} + (\boldsymbol{h} \,.\, \text{grad} \log n) \,\text{grad}\, \mathscr{S} = 0. \tag{42}$$

These are the required *transport equations* for the variation of $\boldsymbol{e}$ and $\boldsymbol{h}$ along each ray. The implications of these equations can best be understood by examining separately the variation of the magnitude and of the direction of these vectors.

We multiply (41) scalarly by $\boldsymbol{e}^\star$ and add to the resulting equation the corresponding equation obtained by taking the complex conjugate. This gives

$$\frac{\partial}{\partial \tau}(\boldsymbol{e} \,.\, \boldsymbol{e}^\star) + \left(\nabla^2\mathscr{S} - \frac{\partial \log \mu}{\partial \tau}\right)\boldsymbol{e} \,.\, \boldsymbol{e}^\star = 0. \tag{43}$$

On account of the identity $\text{div}\, u\mathbf{v} = u \,\text{div}\, \mathbf{v} + \mathbf{v} \,.\, \text{grad}\, u$, the second and third term may be combined as follows:

$$\nabla^2\mathscr{S} - \frac{\partial \log \mu}{\partial \tau'} = \nabla^2\mathscr{S} - \text{grad}\,\mathscr{S} \,.\, \text{grad} \log \mu = \mu \,\text{div}\left(\frac{1}{\mu}\,\text{grad}\,\mathscr{S}\right). \tag{44}$$

Integrating (43) along a ray, the following expression for the ratio of $\boldsymbol{e} \,.\, \boldsymbol{e}^\star$ at any two points of the ray is obtained:*

$$\frac{(\boldsymbol{e} \,.\, \boldsymbol{e}^\star)_2}{(\boldsymbol{e} \,.\, \boldsymbol{e}^\star)_1} = e^{-\int_{\tau_1}^{\tau_2} \mu \,\text{div}\left(\frac{1}{\mu}\text{grad}\,\mathscr{S}\right) d\tau} = e^{-\int_{s_1}^{s_2} \sqrt{\frac{\mu}{\varepsilon}} \,\text{div}\left(\frac{1}{\mu}\text{grad}\,\mathscr{S}\right) ds} \tag{45}$$

---

* This relation may also be written in the alternative form

$$\left(\frac{\boldsymbol{e} \,.\, \boldsymbol{e}^\star}{\mu}\right)_2 = \left(\frac{\boldsymbol{e} \,.\, \boldsymbol{e}^\star}{\mu}\right)_1 e^{-\int_{s_1}^{s_2} \frac{\nabla^2\mathscr{S}}{n} ds} \tag{45a}$$

which follows when (43) is re-written in the form

$$\frac{\partial}{\partial \tau}\left[\log\left(\frac{\boldsymbol{e} \,.\, \boldsymbol{e}^\star}{\mu}\right)\right] = -\nabla^2\mathscr{S},$$

and the integral is taken along a ray. (45a) is in fact only another way of expressing the relation (40) for the variation of intensity, and follows from it when the relation

$$I = \frac{2c}{n}\langle w_e \rangle = \frac{c\varepsilon}{8\pi n}(\boldsymbol{e} \,.\, \boldsymbol{e}^\star)$$

and the MAXWELL formula $\varepsilon\mu = n^2$ are used.

p119

Similarly

$$\frac{(\boldsymbol{h}\,.\,\boldsymbol{h}^\star)_2}{(\boldsymbol{h}\,.\,\boldsymbol{h}^\star)_1} = e^{-\int_{s_1}^{s_2}\sqrt{\frac{\varepsilon}{\mu}}\,\mathrm{div}\left(\frac{1}{\varepsilon}\,\mathrm{grad}\,\mathscr{S}\right)ds}. \tag{46}$$

Next consider the variation of the complex unit vectors

$$\boldsymbol{u} = \frac{\boldsymbol{e}}{\sqrt{\boldsymbol{e}\,.\,\boldsymbol{e}^\star}}, \qquad \boldsymbol{v} = \frac{\boldsymbol{h}}{\sqrt{\boldsymbol{h}\,.\,\boldsymbol{h}^\star}}, \tag{47}$$

along each ray. Substitution into (41) gives

$$\frac{\partial \boldsymbol{u}}{\partial \tau} + \frac{1}{2}\left[\frac{\partial \log(\boldsymbol{e}\,.\,\boldsymbol{e}^\star)}{\partial \tau} + \nabla^2\mathscr{S} - \frac{\partial \log \mu}{\partial \tau}\right]\boldsymbol{u} + (\boldsymbol{u}\,.\,\mathrm{grad}\log n)\,\mathrm{grad}\,\mathscr{S} = 0.$$

The second, third and fourth terms vanish on account of (43), and it follows that

$$\frac{d\boldsymbol{u}}{d\tau} \equiv n\frac{d\boldsymbol{u}}{ds} = -\,(\boldsymbol{u}\,.\,\mathrm{grad}\log n)\,\mathrm{grad}\,\mathscr{S}, \tag{48}$$

and similarly

$$\frac{d\boldsymbol{v}}{d\tau} \equiv n\frac{d\boldsymbol{v}}{ds} = -\,(\boldsymbol{v}\,.\,\mathrm{grad}\log n)\,\mathrm{grad}\,\mathscr{S}. \tag{49}$$

This is the required law for the variation of $\boldsymbol{u}$ and $\boldsymbol{v}$ along each ray.* In particular, *for a homogeneous medium* (n = constant) (48) and (49) reduce to $d\boldsymbol{u}/ds = d\boldsymbol{v}/ds = 0$ *so that* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *then remain constant along each ray.*

Finally we note that for a time-harmonic homogeneous plane wave in a homogeneous medium, $\mathscr{S} = n\boldsymbol{s}.\boldsymbol{r}$ and $\boldsymbol{e}$, $\boldsymbol{h}$, ε and μ are all constants, and consequently $\boldsymbol{K} = \boldsymbol{L} = \boldsymbol{M} \equiv 0$ in (16). Such a wave (whatever its frequency) therefore obeys rigorously the laws of geometrical optics.

3.1.4 Generalizations and the limits of validity of geometrical optics

The considerations of the preceding sections apply to a strictly monochromatic field. Such a field, which may be regarded as a typical FOURIER component of an arbitrary field, is produced by a harmonic oscillator, or by a set of such oscillators of the same frequency.

In optics one usually deals with a source which emits light within a narrow, but nevertheless finite, frequency range. The source may then be regarded as arising from a large number of harmonic oscillators whose frequencies fall within this range. To obtain the intensity at a typical field point P one has to sum the individual fields produced by each oscillator (element of the source):

$$\boldsymbol{E} = \sum_n \boldsymbol{E}_n, \qquad \boldsymbol{H} = \sum_n \boldsymbol{H}_n. \tag{50}$$

---

* The relations (48) and (49) have an interesting interpretation in terms of non-Euclidean geometry. If we consider the associated non-Euclidean space whose line element is given by

$$ds' = nds = n\sqrt{dx^2 + dy^2 + dz^2},$$

then the geometrical light rays correspond to geodesics in this space, and (48) and (49) may be shown to imply that each of the two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is transferred parallel to itself (in the sense of Levi-Civita parallelism) along each ray. Cf. E. BORTOLOTTI, *Rend. R. Acc. Naz. Linc.*, 6a, **4** (1926), 552; R. K. LUNEBURG, *Mathematical Theory of Optics* (mimeographed lecture notes, Brown University, Providence, R.I., 1944, p. 55–59; printed version published by University of California Press, Berkeley and Los Angeles, 1964, p. 51–55); M. KLINE and I. W. KAY, *Electromagnetic Theory and Geometrical Optics* (New York, Interscience Publishers, 1965), p. 180–183; S. M. RYTOV, *Compt. Rend. (Doklady) Acad. Sci. URSS*, **18** (1938), 263.

The intensity is then given by (using real representation)

$$I(P) = |\langle \boldsymbol{S} \rangle| = \frac{c}{4\pi} |\langle \boldsymbol{E} \wedge \boldsymbol{H} \rangle| = \frac{c}{4\pi} |\sum_{n,m} \langle \boldsymbol{E}_n \wedge \boldsymbol{H}_m \rangle|$$

$$= \frac{c}{4\pi} |\sum_n \langle \boldsymbol{E}_n \wedge \boldsymbol{H}_n \rangle + \sum_{n \neq m} \langle \boldsymbol{E}_n \wedge \boldsymbol{H}_m \rangle|. \tag{51}$$

In many optical problems it is usually permissible to assume that the second sum in (51) vanishes (the fields are then said to be *incoherent*), so that

$$I(P) = \frac{c}{4\pi} |\sum_n \langle \boldsymbol{E}_n \wedge \boldsymbol{H}_n \rangle| = |\sum_n \langle \boldsymbol{S}_n \rangle|, \tag{52}$$

$\boldsymbol{S}_n$ denoting the Poynting vector due to the nth element of the source. It is not possible to discuss at this stage the conditions under which the neglect of the second

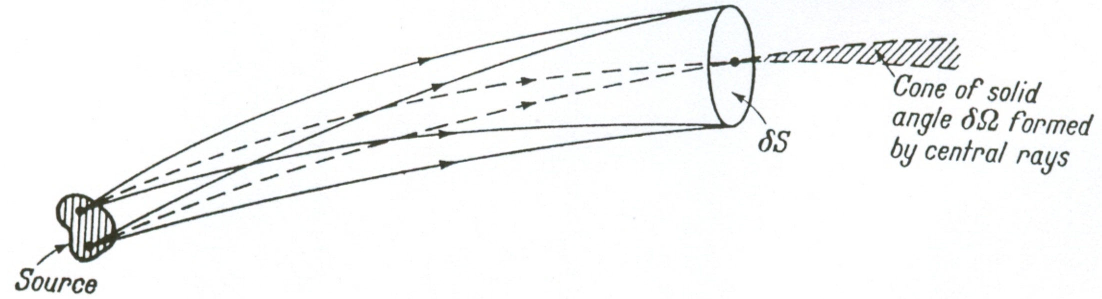

Fig. 3.4. Illustrating the intensity law of geometrical optics for an extended incoherent source.

term in (51) is justified, but this point will be considered fully later, in connection with partial coherence (Chapter X).

Let δS be a small portion of a wave-front associated with one particular element of the source. Every element of the source sends through δS a tube of rays, and the central rays of these tubes fill a cone of solid angle $\delta\Omega$ (Fig. 3.4). If the semi-vertical angle of this cone is small enough, we may neglect the variation of $\boldsymbol{S}_n$ with direction, and (52) may then be replaced by

$$I(P) = \sum_n |\langle \boldsymbol{S}_n \rangle| = \sum_n I_n. \tag{53}$$

Now the number of elements (oscillators) may be regarded as being so large that no appreciable error is introduced by treating the distribution as continuous. The contribution due to each element is then infinitesimal, but the total effect is finite. The sum (integral) is then proportional to $\delta\Omega$:

$$I(P) = B\delta\Omega,$$

and the total (time-averaged) energy flux δF which crosses the element δS per unit time is given by

$$\delta F = B\delta\Omega\delta S. \tag{54}$$

This formula is of importance in photometry, and will be used later.

We must now briefly consider the limits of validity of geometrical optics. The eikonal equation was derived on the assumption that the terms on the right-hand sides of (11) and (12) may be neglected. If the dimensionless quantities ε, μ and

P121

grad $\mathscr{S}$| are assumed to be of order unity, we see that this neglect will be justified provided that the magnitudes of the changes in $\boldsymbol{e}$ and $\boldsymbol{h}$ are small compared with the magnitudes of $\boldsymbol{e}$ and $\boldsymbol{h}$ over domains whose linear dimensions are of the order of a wavelength. This condition is violated, for example, at boundaries of shadows, for across such boundaries the intensity (and therefore also $\boldsymbol{e}$ and $\boldsymbol{h}$) changes rapidly. In the neighbourhood of points where the intensity distribution has a very sharp maximum (e.g. at a focus, see § 8.8), geometrical optics likewise cannot be expected to describe correctly the behaviour of the field.

The transport equations (41) and (42) for the complex amplitude vector $\boldsymbol{e}$ and $\boldsymbol{h}$ were obtained on the assumption that $\mathscr{S}$ satisfies the eikonal equation, and that the terms $\lambda_0|\boldsymbol{M}(\boldsymbol{e}, \varepsilon, \mu)|$ and $\lambda_0|\boldsymbol{M}(\boldsymbol{h}, \mu, \varepsilon)|$ are small compared with $|\boldsymbol{L}(\boldsymbol{e}, \mathscr{S}, n, \mu)|$ and $|\boldsymbol{L}(\boldsymbol{h}, \mathscr{S}, n, \varepsilon)|$ respectively. This imposes certain additional restrictions on, not only the first, but also the second derivatives of $\boldsymbol{e}$ and $\boldsymbol{h}$. These conditions are rather complicated and will not be studied here.

It is, of course, possible to obtain improved approximations by retaining some of the higher-order terms in the expansions (18) for the field vectors.* In problems of instrumental optics, the practical advantage of such a procedure is, however, doubtful, since the closer the special regions are approached the more terms have to be included, and the expansions usually break down completely at points of particular interest (e.g. at a focus or at a caustic surface). A more powerful approach to the study of the intensity distribution in such regions is offered by methods which will be discussed in the chapters on diffraction.

Finally we stress that the simplicity of the geometrical optics model arises essentially from the fact that, in general, the field behaves *locally* as a plane wave. At optical wavelengths, the regions for which this simple geometrical model is inadequate are an exception rather than a rule; in fact for most optical problems geometrical optics furnishes at least a good starting point for more refined investigations.

3.2 GENERAL PROPERTIES OF RAYS

3.2.1 The differential equation of light rays

The light rays have been defined as the orthogonal trajectories to the geometrical wave-fronts $\mathscr{S}(x, y, z) =$ constant and we have seen that, if $\boldsymbol{r}$ is a position vector of a typical point on a ray and s the length of the ray measured from a fixed point on it, then

$$n\frac{d\boldsymbol{r}}{ds} = \operatorname{grad}\mathscr{S}. \tag{1}$$

* It has been suggested by J. B. Keller [*J. Appl. Phys.*, **28** (1957), 426; also *Calculus of Variations and its Application*, ed. L. M. Graves (New York, McGraw-Hill, 1958), 27] that the behaviour of the contributions represented by the higher-order terms may be studied by means of a model which is an extension of ordinary geometrical optics. In this theory the concept of a *diffracted ray* is introduced, which obeys a generalized Fermat's principle. With each such ray an appropriate field is associated and is assumed to satisfy the same propagation laws as the geometrical optics field. Some applications of the theory were described by J. B. Keller, *Trans. Inst. Radio Eng.*, A.P.—4 (1956), 312 and J. B. Keller, R. M. Lewis and B. D. Seckler, *J. Appl. Phys.*, **28** (1957), 570. See also M. Kline and I. W. Kay, *loc. cit.*

This equation specifies the rays by means of the function $\mathscr{S}$, but one can easily deriv from it a differential equation which specifies the rays directly in terms of the refractive index function $n(\boldsymbol{r})$.

Differentiating (1) with respect to s we obtain

$$\frac{d}{ds}\left(n\frac{d\boldsymbol{r}}{ds}\right) = \frac{d}{ds}(\operatorname{grad}\mathscr{S})$$

$$= \frac{d\boldsymbol{r}}{ds}\cdot\operatorname{grad}(\operatorname{grad}\mathscr{S})$$

$$= \frac{1}{n}\operatorname{grad}\mathscr{S}\cdot\operatorname{grad}(\operatorname{grad}\mathscr{S}) \qquad \text{(by (1))}$$

$$= \frac{1}{2n}\operatorname{grad}[(\operatorname{grad}\mathscr{S})^2]$$

$$= \frac{1}{2n}\operatorname{grad} n^2 \qquad \text{(by § 3.1 (15))}$$

i.e.

$$\frac{d}{ds}\left(n\frac{d\boldsymbol{r}}{ds}\right) = \operatorname{grad} n. \tag{2}$$

This is the vector form of the differential equations of the light rays. In particular, *in a homogeneous medium* $n =$ constant and (2) then reduces to

$$\frac{d^2\boldsymbol{r}}{ds^2} = 0,$$

whence

$$\boldsymbol{r} = s\boldsymbol{a} + \boldsymbol{b}, \tag{3}$$

$\boldsymbol{a}$ and $\boldsymbol{b}$ being constant vectors. (3) is a vector equation of a straight line in the direction of the vector $\boldsymbol{a}$, passing through the point $\boldsymbol{r} = \boldsymbol{b}$. Hence *in a homogeneous medium the light rays have the form of straight lines.*

As an example of some interest, let us consider rays in a medium which has spherical symmetry, i.e. where the refractive index depends only on the distance r from a fixed point O:

$$n = n(r). \tag{4}$$

This case is approximately realized by the earth's atmosphere, when the curvature of the earth is taken into account.

Consider the variation of the vector $\boldsymbol{r} \wedge [n(r)\boldsymbol{s}]$ along the ray. We have

$$\frac{d}{ds}(\boldsymbol{r}\wedge n\boldsymbol{s}) = \frac{d\boldsymbol{r}}{ds}\wedge n\boldsymbol{s} + \boldsymbol{r}\wedge\frac{d}{ds}(n\boldsymbol{s}). \tag{5}$$

Since $d\boldsymbol{r}/ds = \boldsymbol{s}$, the first term on the right vanishes. The second term may, on account of (2), be written as $\boldsymbol{r}\wedge\operatorname{grad} n$. Now from (4)

$$\operatorname{grad} n = \frac{\boldsymbol{r}}{r}\frac{dn}{dr},$$

so that the second term on the right-hand side of (5) also vanishes. Hence

$$\boldsymbol{r}\wedge n\boldsymbol{s} = \text{constant}. \tag{6}$$

This relation implies that all the rays are plane curves, situated in a plane through the origin, and that along each ray

$$nr \sin\phi = \text{constant}, \tag{7}$$

where ϕ is the angle between the position vector $\boldsymbol{r}$ and the tangent at the point $\boldsymbol{r}$ on the ray (see Fig. 3.5). Since $r \sin\phi$ represents the perpendicular distance d from the origin to the tangent, (7) may also be written as

$$nd = \text{constant}. \tag{8}$$

This relation is sometimes called the *formula of Bouguer* and is the analogue of a well-known formula in dynamics, which expresses the conservation of angular momentum of a particle moving under the action of a central force.

To obtain an explicit expression for the rays in a spherically symmetrical medium, we recall from elementary geometry that, if (r,θ) are the polar coordinates of a

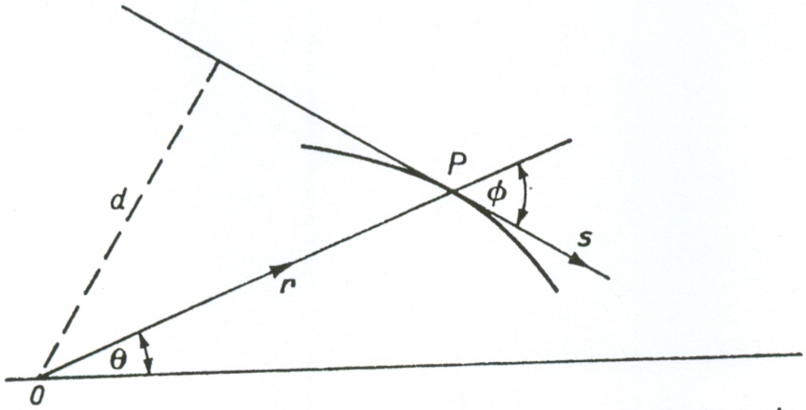

Fig. 3.5. Illustrating BOUGUER's formula nd = constant, for rays in a medium with spherical symmetry.

plane curve, then the angle ϕ between the radius vector to a point P on the curve and the tangent at P is given by*

$$\sin\phi = \frac{r(\theta)}{\sqrt{r^2(\theta) + \left(\dfrac{dr}{d\theta}\right)^2}}. \tag{9}$$

From (7) and (9)

$$\frac{dr}{d\theta} = \frac{r}{c}\sqrt{n^2r^2 - c^2}, \tag{10}$$

c being a constant. The equation of rays in a medium with spherical symmetry may therefore be written in the form

$$\theta = c\int^r \frac{dr}{r\sqrt{n^2r^2 - c^2}}. \tag{11}$$

Let us now return to the general case and consider the *curvature vector* of a ray, i.e. the vector

$$\boldsymbol{K} = \frac{d\boldsymbol{s}}{ds} = \frac{1}{\rho}\boldsymbol{\nu}, \tag{12}$$

---

* See, for example, R. COURANT, *Differential and Integral Calculus*, Vol. I (Glasgow, Blackie, 2nd edition, 1942), p. 265.

whose magnitude $1/\rho$ is the reciprocal of the radius of curvature; $\boldsymbol{\nu}$ is the unit principal normal at a typical point of the ray.

From (2) and (12) it follows that

$$n\boldsymbol{K} = \text{grad}\, n - \frac{dn}{ds}\,\boldsymbol{s}. \tag{13}$$

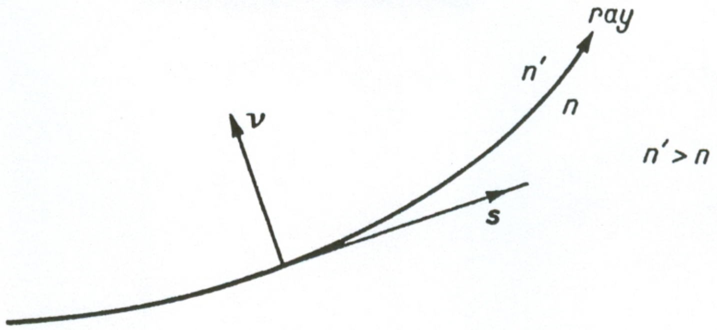

Fig. 3.6. Bending of a ray in a heterogeneous medium.

This relation shows that *the gradient of the refractive index lies in the osculating plane of the ray.*

If we multiply (13) scalarly by $\boldsymbol{K}$ and use (12) we find that

$$|\boldsymbol{K}| = \frac{1}{\rho} = \boldsymbol{\nu} \,.\, \text{grad} \log n. \tag{14}$$

Since ρ is always positive, this implies that as we proceed along the principal normal the refractive index increases i.e. *the ray bends towards the region of higher refractive index* (Fig. 3.6).

3.2.2 The laws of refraction and reflection

So far it has been assumed that the refractive index function n is continuous. We must now discuss the behaviour of rays when they cross a surface separating two homogeneous media of different refractive indices. It has been shown by SOMMERFELD

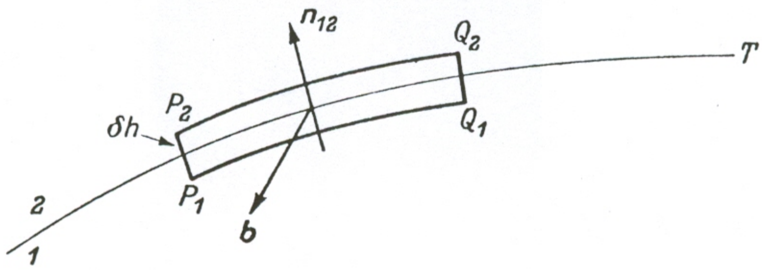

Fig. 3.7. Derivation of the laws of refraction and reflection.

and RUNGE (*loc. cit.*) that the behaviour can easily be determined by an argument similar to that used in deriving the conditions relating to the changes in the field vectors across a surface discontinuity (cf. § 1.1.3).

It follows from (1), on account of the identity curl grad $\equiv 0$, that the vector $n\boldsymbol{s} = n d\boldsymbol{r}/ds$, called sometimes the *ray vector*, satisfies the relation

$$\text{curl}\, n\boldsymbol{s} = 0. \tag{15}$$

As in § 1.1.3 we replace the discontinuity surface T by a transition layer throughout which ε, μ and n change rapidly but continuously from their values near T on one

If $\boldsymbol{b}$ denotes the unit normal to this area, then we have from (15), on integrating throughout the area and applying STOKES' theorem,

$$\int (\text{curl } n\boldsymbol{s}) \,.\, \boldsymbol{b}\, dS = \int n\boldsymbol{s} \,.\, d\boldsymbol{r} = 0, \tag{16}$$

the second integral being taken along the boundary curve $P_1Q_1Q_2P_2$. Proceeding to the limit as the height $\delta h \to 0$, in a strictly similar manner as in the derivation of § 1.1 (23), we obtain

$$\boldsymbol{n}_{12} \wedge (n_2\boldsymbol{s}_2 - n_1\boldsymbol{s}_1) = 0, \tag{17}$$

where $\boldsymbol{n}_{12}$ is the unit normal to the boundary surface pointing from the first into the second medium. (17) implies that *the tangential component of the ray vector $n\boldsymbol{s}$ is*

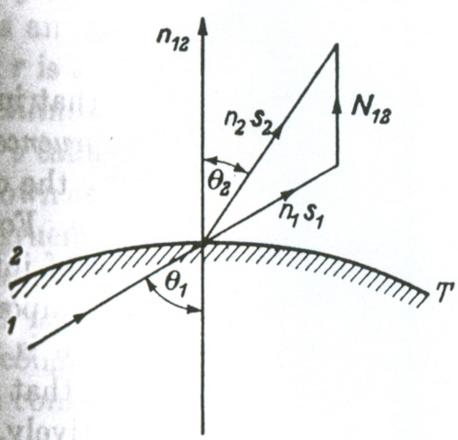

Fig. 3.8 (a). Illustrating the law of refraction.

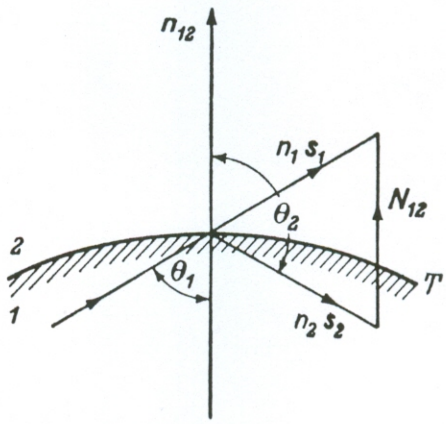

Fig. 3.8 (b). Illustrating the law of reflection.

continuous across the surface or, what amounts to the same thing, *the vector $\boldsymbol{N} = n_2\boldsymbol{s}_2 - n_1\boldsymbol{s}_1$ is normal to the surface.*

Let θ_1 and θ_2 be the angles which the incident ray and the refracted ray make with the normal $\boldsymbol{n}_{12}$ to the surface (see Fig. 3.8a). Then it follows from (17) that

$$n_2(\boldsymbol{n}_{12} \wedge \boldsymbol{s}_2) = n_1(\boldsymbol{n}_{12} \wedge \boldsymbol{s}_1), \tag{18}$$

so that

$$n_2 \sin \theta_2 = n_1 \sin \theta_1. \tag{19}$$

(18) implies that *the refracted ray lies in the same plane as the incident ray and the normal to the surface* (*the plane of incidence*) and (19) shows that *the ratio of the sine of the angle of refraction to the sine of the angle of incidence is equal to the ratio n_1/n_2 of the refractive indices.* These two results express *the law of refraction* (*Snell's law*). This law has already been derived in § 1.5 for the special case of plane waves. But whilst the earlier discussion concerned a plane wave of *arbitrary* wavelength falling upon a plane refracting surface, the present analysis applies to waves and refracting surfaces of more general form, provided that the wavelength is sufficiently small ($\lambda_0 \to 0$). This condition means, in practice, that the radii of curvature of the incident wave and of the boundary surface must be large compared to the wavelength of the incident light.

As in the case treated in § 1.5 we must expect that there will be another wave, the reflected wave, propagated back into the first medium. Setting $n_2 = n_1$ in (18) and (19) (see Fig. 3.8b) it follows that *the reflected ray lies in the plane of incidence* and that $\sin \theta_2 = \sin \theta_1$; hence

$$\theta_2 = \pi - \theta_1. \tag{20}$$

The last two results express *the law of reflection.*

3.2.3 Ray congruences and their focal properties

The relation (15), namely

$$\text{curl}\, n\boldsymbol{s} = 0, \tag{21}$$

characterizes all the ray systems which can be realized in an isotropic medium and distinguishes them from more general families of curves. In a homogeneous isotropic medium n is constant, and (21) then reduces to

$$\text{curl}\, \boldsymbol{s} = 0. \tag{22}$$

Rays in a heterogeneous isotropic medium can also be characterized by a relation independent of n. It may be obtained by applying to (21) the identity $\text{curl}\, n\boldsymbol{s} = n\, \text{curl}\, \boldsymbol{s} + (\text{grad}\, n) \wedge \boldsymbol{s}$ and taking the scalar product with $\boldsymbol{s}$. It then follows that a system of rays in any *isotropic* medium must satisfy the relation

$$\boldsymbol{s} \,.\, \text{curl}\, \boldsymbol{s} = 0. \tag{23}$$

A system of curves which fills a portion of space in such a way that in general a single curve passes through each point of the region is called a *congruence*. If there exists a family of surfaces which cut each of the curves orthogonally the congruence is said to be *normal*; if there is no such family, it is said to be *skew*. For ordinary geometrical optics (light propagation) only normal congruences are of interest, but in electron optics (see Appendix II) skew congruences also play an important part.

If each curve of the congruence is a straight line the congruence is said to be *rectilinear*; (23) and (22) are the necessary and sufficient conditions that the curves should represent a *normal* and a *normal rectilinear congruence* respectively.*

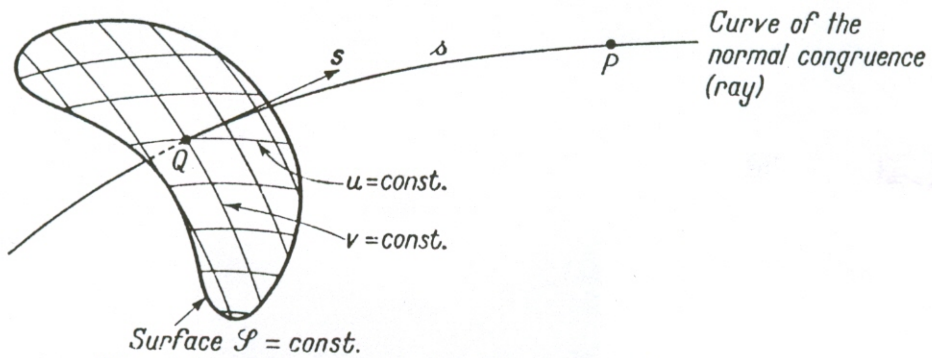

Fig 3.9. Notation relating to a normal congruence.

Let us choose a set of curvilinear coordinate lines u, v on one of the orthogonal surfaces $\mathscr{S}(x, y, z) = \text{constant}$. To every point $Q(u, v)$ of this surface there will then correspond one curve of the congruence, namely that curve which meets $\mathscr{S}$ in Q. Let $\boldsymbol{r}$ denote the position vector of a point P on the curve. $\boldsymbol{r}$ may then be regarded as a function of the coordinates (u, v) and of the length of arc s between Q and P, measured along the curve (Fig. 3.9).

Consider two neighbouring curves of the congruence passing through the points (u, v) and $(u + du, v + dv)$ on $\mathscr{S}$, and let us examine whether there are points on these curves such that the distance between them is of the second or higher order (one says that the curves cut to first order at such points). Points with this property are called *foci* and must satisfy the equation

$$\boldsymbol{r}(u, v, s) = \boldsymbol{r}(u + du, v + dv, s + ds) \tag{24}$$

to the first order.

* For a more detailed discussion of congruences of curves see, for example, C. E. WEATHERBURN, *Differential Geometry of Three Dimensions* (Cambridge University Press), Vol. I (1927), Chapter X;

Expanding (24) we obtain

$$\mathbf{r}_u du + \mathbf{r}_v dv + \mathbf{s} ds = 0, \tag{25}$$

where $\mathbf{r}_u, \mathbf{r}_v$ are the partial derivatives with respect to u and v. Condition (25) implies that $\mathbf{r}_u, \mathbf{r}_v$ and $\mathbf{s}$ are coplanar. This is equivalent to saying that the scalar triple product of the three vectors vanishes, i.e.

$$[\mathbf{r}_u, \mathbf{r}_v, \mathbf{s}] = 0. \tag{26}$$

The number of foci on a given curve (u, v) depends on the number of values of s which satisfy (26). If $\mathbf{r}$ is a polynomial in s of degree m, then since $\mathbf{s} = d\mathbf{r}/ds$, it is seen that (26) is an equation of degree $3m - 1$ in s. In particular, if the congruence is rectilinear, $\mathbf{r}$ is a linear function of s ($m = 1$), showing that *there are two foci on each ray of a rectilinear congruence.*

If u and v take on all possible values, the foci will describe a surface, represented by (26), known as the *focal surface*; in optics it is called the *caustic surface*. Any curve of the congruence is tangent to the focal surface at each focus of the curve. The tangent plane at any point of the focal surface is known as the *focal plane.*

We shall mainly be concerned with rays in a homogeneous medium, i.e. with rectilinear congruences. Some further properties of such congruences will be discussed in § 4.6, in connection with astigmatic pencils of rays.

3.3 OTHER BASIC THEOREMS OF GEOMETRICAL OPTICS

With the help of the relations established in the preceding sections, we shall now derive a number of theorems concerning rays and wave-fronts.

3.3.1 Lagrange's integral invariant

Assume first that the refractive index n is a continuous function of position. Then as in § 3.2 (16) it follows on applying STOKES' theorem to the integral, taken over any open surface, of the normal component of curl $n\mathbf{s}$, that

$$\oint n\mathbf{s} \,.\, d\mathbf{r} = 0. \tag{1}$$

The integral extends over the closed boundary curve C of the surface. (1) is known as *Lagrange's integral invariant** and implies that *the integral*

$$\int_{P_1}^{P_2} n\mathbf{s} \,.\, d\mathbf{r} \tag{2}$$

taken between any two points P_1 and P_2 in the field, is independent of the path of integration.

* Sometimes called *Poincaré's invariant.* In fact it is only a special one-dimensional case of much more general integral invariants discussed by J. H. POINCARÉ in his *Les Méthodes Nouvelles de la Mécanique Céleste*, **3** (Paris, Gauthier-Villars, 1899). Cf. E. CARTAN, *Lecons sur les Invariants Integraux* (Paris, Hermann, 1922). See also our Appendix I, eq. (85).

With the help of the law of refraction it is easily seen that (1) also holds when t curve C intersects a surface which separates two homogeneous media of differe refractive indices. To show this, let C_1 and C_2 be the portions of C on each side of t refracting surface T (Fig. 3.10), and let the points of intersection of C with the surfa T be joined by another curve K in the surface. On taking (1) along each of the loo C_1K and C_2K and on adding the equations, we obtain

$$\int_{C_1} n_1\mathbf{s}_1 \cdot d\mathbf{r} + \int_{C_2} n_2\mathbf{s}_2 \cdot d\mathbf{r} + \int_K (n_2\mathbf{s}_2 - n_1\mathbf{s}_1) \cdot d\mathbf{r} = 0. \qquad ($$

The integral over K vanishes, since according to the law of refraction the vect $\mathbf{N} = n_1\mathbf{s}_1 - n_2\mathbf{s}_2$ is at each point of K perpendicular to the surface, and consequentl (3) reduces to (1).

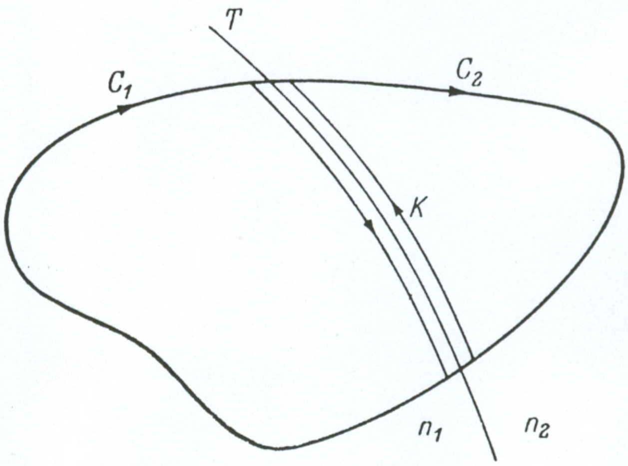

Fig. 3.10. Derivation of the LAGRANGE's integral invariant in the presence of a surface of discontinuity of the refractive index.

3.3.2 The principle of Fermat

The *principle of Fermat*, known also as the principle of the *shortest optical path** asserts that *the optical length*

$$\int_{P}^{P_2} n ds \qquad (4)$$

of an actual ray between any two points P_1 and P_2 is shorter than the optical length of any other curve which joins these points and which lies in a certain regular neighbourhood of it. By a regular neighbourhood we mean one that may be covered by rays in such a way that one (and only one) ray passes through each point of it. Such a covering is exhibited, for example, by rays from a point source P_1 in that domain around P_1 where the rays on account of refraction or reflection or on account of their curvature do not intersect each other.

Before proving this theorem it may be mentioned that it is possible to formulate FERMAT's principle in a form which is weaker but which has a wider range of validity. According to this formulation the actual ray is distinguished from other curves (no

* Since by § 3.1 (27)

$$\int_{P_1}^{P_2} n ds = c \int_{P_1}^{P_2} dt$$

it is also known as the *principle of least time.*

longer restricted to lie in a regular neighbourhood) by a *stationary value* of the integral.*

To prove FERMAT's principle, we take a pencil of rays and compare a segment P_1P_2 of a ray $\bar{C}$ with an arbitrary curve C joining P_1 and P_2 (Fig. 3.11). Let two neighbouring orthogonal trajectories (wave-fronts) of the pencil intersect C in Q_1 and Q_2 and $\bar{C}$ in $\bar{Q}_1$ and $\bar{Q}_2$. Further let Q_2' be the point of intersection of the trajectory $Q_2\bar{Q}_2$ with the ray $\bar{C}'$ which passes through Q_1.

Applying LAGRANGE's integral relation to the small triangle $Q_1Q_2Q_2'$, we have

$$(n\mathbf{s}\,.\,d\mathbf{r})_{Q_1Q_2} + (n\mathbf{s}\,.\,d\mathbf{r})_{Q_2Q_2'} - (nds)_{Q_1Q_2} = 0. \tag{5}$$

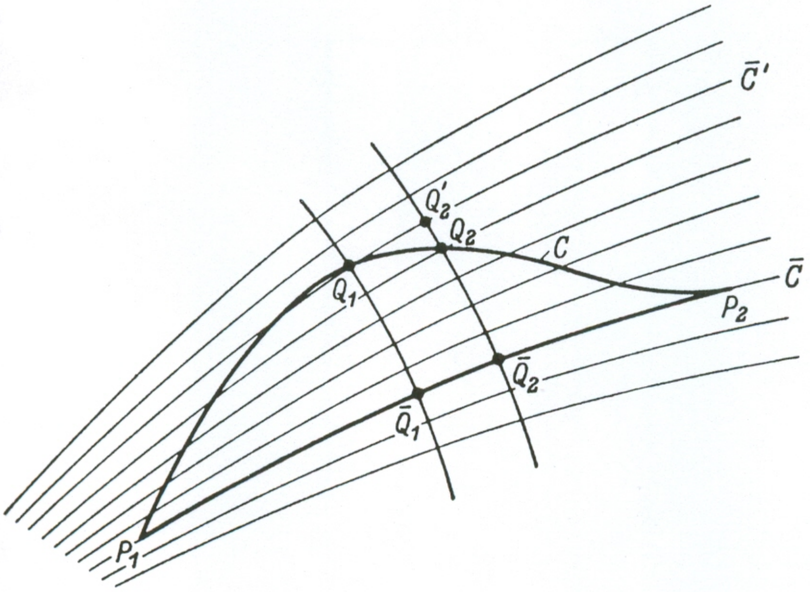

Fig. 3.11. Illustrating FERMAT's principle.

Now from the definition of the scalar product

$$(n\mathbf{s}\,.\,d\mathbf{r})_{Q_1Q_2} \leqslant (nds)_{Q_1Q_2}.$$

Further, $\mathbf{s}$ is orthogonal to $d\mathbf{r}$ on the wave-front, so that

$$(n\mathbf{s}\,.\,d\mathbf{r})_{Q_2Q_2'} = 0.$$

Also from § 3.1 (25), since Q_1, Q_2' and $\bar{Q}_1$, $\bar{Q}_2$ are corresponding points on the two wave-fronts,

$$(nds)_{Q_1Q_2'} = (nds)_{\bar{Q}_1\bar{Q}_2}.$$

On substituting from the last three relations into (5) we find that

$$(nds)_{\bar{Q}_1\bar{Q}_2} \leqslant (nds)_{Q_1Q_2}, \tag{6}$$

whence, on integration,

$$\int_{\bar{C}} nds \leqslant \int_C nds. \tag{7}$$

* To find the curves for which the integral has a stationary value we must apply in general the methods of the variational calculus, described in Appendix I. It is shown there that such curves satisfy the EULER Differential Equations AI (7). In the present case these are nothing but the equations § 3.2. (2) of the rays as shown in section 11 of Appendix I.

It has been stressed by C. CARATHÉODORY (*loc. cit.*) that the stationary value is never a true maximum. In the weaker formulation of FERMAT's principle it is therefore appropriate to speak of a stationary value rather than of an extremal value. The minimal formulation on the other hand corresponds to a "strong minimum" in the sense of JACOBI (Appendix I, section 10).

Moreover, the equality sign could only hold if the directions of $\boldsymbol{s}$ and $d\boldsymbol{r}$ were coincident at every point of C, i.e. if the comparison curve was an actual ray. This case is excluded by our assumption that not more than one ray passes through any point of the neighbourhood. Hence the optical length of the ray is smaller than the optical length of the comparison curve, which is FERMAT's principle.

It can easily be seen that, when the regularity condition is not fulfilled, the optical length of the ray may no longer be a minimum. Consider for example a field of rays

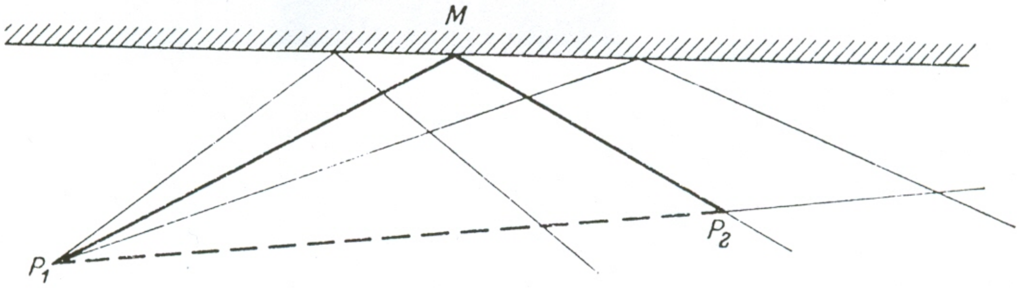

Fig. 3.12. Field of rays obtained by reflection of light from a point source on a plane mirror.

from a point source P_1 in a homogeneous medium, reflected by a plane mirror (Fig. 3.12). Two rays then pass through each point P_2; the optical length of the direct ray P_1P_2 is an absolute minimum but the reflected ray P_1MP_2 gives a minimum only relative to curves in a certain restricted neighbourhood of it. In general when rays from a point source P_1 are refracted or reflected at boundaries between homogeneous media, the regular neighbourhood will terminate on the envelope (caustic) formed by the rays. The point P_1' at which a ray from a point source at P_1 touches the envelope is called the *conjugate* of P_1 on the particular ray. For the optical length of a ray P_1P_2

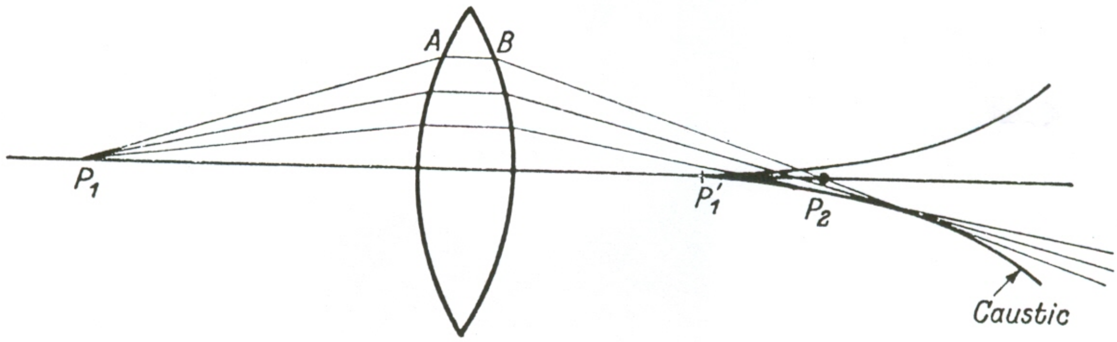

Fig. 3.13. Caustic formed by rays from an axial point source after passing through a lens.

to be a minimum, P_2 must lie between P_1 and P_1', i.e. P_1 and P_2 must lie on the same side of the caustic. For example, in the case of an uncorrected lens (Fig. 3.13) the central ray from P_1 has a minimal optical length only up to the tip (P_1') of the caustic (the Gaussian image of P_1). For any point P_2 which lies behind the envelope the optical length of the direct path $P_1P_1'P_2$ exceeds that of the broken path P_1ABP_2.

3.3.3 The theorem of Malus and Dupin and some related theorems

The light rays have been defined as the orthogonal trajectories of the wave surfaces $\mathscr{S}(x, y, z) =$ constant, $\mathscr{S}$ being a solution of the eikonal equation (15) in § 3.1. This is a natural way of introducing the light rays when the laws of geometrical optics are to be deduced from MAXWELL's equations. Historically, however, geometrical optics developed as the theory of light rays which were defined differently, namely as curves hich the line integral $\int nds$ has a stationary value. Formulated this way etrical optics may then be developed purely along the lines of calculus of ions.* P13

iational considerations are of considerable importance as they often reveal gies between different branches of physics. In particular there is a close analogy en geometrical optics and the mechanics of a moving particle; this was brought ery clearly by the celebrated investigations of Sir W. R. HAMILTON, whose ach became of great value in modern physics, especially in applications to DE LIE's wave mechanics. In order not to interrupt the optical considerations, an nt of the relevant parts of the calculus of variation and of the Hamiltonian gy are given in separate sections (Appendix I and II). Here we shall only show

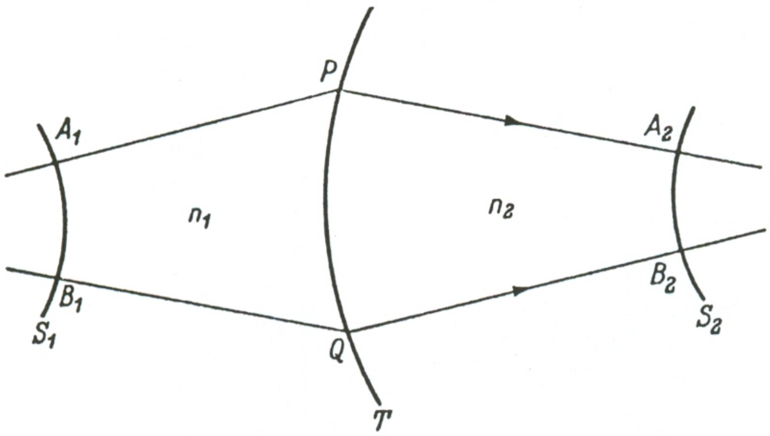

Fig. 3.14. Illustrating the theorem of MALUS and DUPIN.

several theorems, which played an important part in the development of etrical optics, may be derived from LAGRANGE's integral invariant.

nsider rays in a homogeneous medium: if they all have a point in common, hen they then proceed from a point source, they are said to form a *homocentric*. Such a pencil forms a normal congruence, since every ray of the pencil is cut gonally by spheres centred on the mutual point of intersection of the rays. In MALUS† showed that, if a homocentric pencil of rectilinear rays is refracted or ted at a surface, the resulting pencil (in general no longer homocentric) will again a normal congruence. Later DUPIN (1816), QUETELET (1825), and GERGONNE) generalized MALUS's result. These investigations lead to the following theorem, n sometimes as *the theorem of Malus and Dupin: A normal rectilinear congruence ns normal after any number of refractions or reflections.*‡

will be sufficient to establish the theorem for a single refraction. Consider a al rectilinear congruence of rays in a homogeneous medium of refractive index n_1 ssume that the rays undergo a refraction at a surface T which separates this um from another homogeneous medium of refractive index n_2 (Fig. 3.14).

S_1 be one of the orthogonal trajectories (wave-fronts) in the first region, and and P be the points of intersections of a typical ray in the first medium with

---

systematic treatment of this kind is given for example in C. CARATHÉODORY (*loc. cit.*).

. MALUS, Optique Dioptrique, *J. École polytechn.*, **7** (1808), 1–44, 84–129. Also his "Traité que", *Mém. présent. à l'Institut par divers savants*, **2** (1811), 214–302. References and an nt of the interesting history of the MALUS–DUPIN theorem can be found in the *Mathematical s of Sir William Rowan Hamilton*, **1** (Geometrical Optics), edited by A. W. CONWAY and SYNGE (Cambridge University Press, 1931), p. 463.

. LEVI-CIVITA, *Rend. R. Acc. Naz. Linc.*, **9** (1900) 237 established the converse theorem, y that in general two normal rectilinear congruences may be transformed into each other by e refraction or reflection.

S_1 and with T respectively, and let A_2 be any point on the refracted ray. If the point A_1 is displaced to another point B_1 on the wave-front, the point P will be displaced to another point Q on the refracting surface. Now take a point B_2, on the ray which is refracted at Q, such that the optical path from B_1 to B_2 is equal to the optical path from A_1 to A_2:

$$[A_1PA_2] = [B_1QB_2]. \tag{8}$$

As B_1 takes on all possible positions on S_1 the point B_2 describes a surface S_2. It will now be shown that the refracted ray QB_2 is perpendicular to this surface.

Applying LAGRANGE's integral invariant to the closed path $A_1PA_2B_2QB_1A_1$, it follows that

$$\int_{A_1PA_2} nds + \int_{A_2B_2} n\mathbf{s}\,.\,d\mathbf{r} + \int_{B_2QB_1} nds + \int_{B_1A_1} n\mathbf{s}\,.\,d\mathbf{r} = 0. \tag{9}$$

Now by (8),

$$\int_{A_1PA_2} nds + \int_{B_2QB_1} nds = 0. \tag{10}$$

Moreover, since on S_1 the unit vector $\mathbf{s}$ is everywhere orthogonal to S_1,

$$\int_{B_1A_1} n\mathbf{s}\,.\,d\mathbf{r} = 0, \tag{11}$$

so that (9) reduces to

$$\int_{A_2B_2} n\mathbf{s}\,.\,d\mathbf{r} = 0. \tag{12}$$

This relation must hold for every curve on S_2. This is only possible if $\mathbf{s}\,.\,d\mathbf{r} = 0$ for every linear element $d\mathbf{r}$ of S_2, i.e. if the refracted rays are orthogonal to the surface; in other words *if the refracted rays form a normal congruence.* The proof for reflection is strictly analogous.

Since $[A_1PA_2] = [B_1QB_2]$ it follows that *the optical path length between any two orthogonal surfaces (wave-fronts) is the same for all rays.* This result clearly remains valid when several successive refractions or reflections takes place and, as is immediately obvious from eq. (26) in § 3.1 it also applies to rays in a medium with continuously varying refractive index. This theorem is known as the *principle of equal optical path*; it implies that the orthogonal trajectories (geometrical wavefronts) of a normal congruence of rays, or of a set of normal congruences generated by successive refractions or reflections, are "optically parallel" to each other (cf. Appendix I).

A related theorem, first put forward by HUYGENS* asserts that *each element of a wave-front may be regarded as the centre of a secondary disturbance which gives rise to spherical wavelets*; and moreover that *the position of the wave-front at any later time is the envelope of all such wavelets.* This result, sometimes called *Huygens' construction,* is essentially a rule for the construction of a set of surfaces which are "optically parallel" to each other. If the medium is homogeneous, one can use in the construction wavelets of finite radius, in other cases one has to proceed in infinitesimal steps.

HUYGENS' theorem was later extended by FRESNEL and led to the formulation of the so-called Huygens-Fresnel principle, which is of great importance in the theory of diffraction (see § 8.2), and which may be regarded as the basic postulate of the wave theory of light.

---

* *Traité de la Lumière* (Leyden, 1690); English translation (*Treatise on Light*) by S. P. THOMPSON (London, Macmillan & Co., 1912).