I

Basic properties of the electromagnetic field

1.1 The electromagnetic field

1.1.1 Maxwell's equations

THE state of excitation which is established in space by the presence of electric charges is said to constitute an *electromagnetic field*. It is represented by two vectors, **E** and **B**, called the *electric vector* and the *magnetic induction* respectively.*

To describe the effect of the field on material objects, it is necessary to introduce a second set of vectors, viz. *the electric current density* **j**, *the electric displacement* **D**, and *the magnetic vector* **H**.

The space and time derivatives of the five vectors are related by *Maxwell's equations*, which hold at every point in whose neighbourhood the physical properties of the medium are continuous:†

$$\operatorname{curl} \mathbf{H} - \frac{1}{c}\dot{\mathbf{D}} = \frac{4\pi}{c}\mathbf{j}, \tag{1}$$

$$\operatorname{curl} \mathbf{E} + \frac{1}{c}\dot{\mathbf{B}} = 0, \tag{2}$$

the dot denoting differentiation with respect to time.

* In elementary considerations **E** and **H** are, for historical reasons, usually regarded as the basic field vectors, and **D** and **B** as describing the influence of matter. In general theory, however, the present interpretation is compulsory for reasons connected with the electrodynamics of moving media.

The four Maxwell equations (1)–(4) can be divided into two sets of equations, one consisting of two homogeneous equations (right-hand side zero), containing **E** and **B**, the other of two nonhomogeneous equations (right-hand side different from zero), containing **D** and **H**. If a coordinate transformation of space and time (relativistic Lorentz transformation) is carried out, the equations of each group transform together, the equations remaining unaltered in form if $\mathbf{j}/c$ and ρ are transformed as a four-vector, and each of the pairs **E**, **B** and **D**, **H** as a six-vector (antisymmetric tensor of the second order). Since the nonhomogeneous set contains charges and currents (which represent the influence of matter), one has to attribute the corresponding pair (**D**, **H**) to the influence of matter. It is, however, customary to refer to **H** and not to **B** as the *magnetic field vector*; we shall conform to this terminology when there is no risk of confusion.

† The so-called Gaussian system of units is used here, i.e. the electrical quantities (**E**, **D**, **j** and ρ) are measured in electrostatic units, and the magnetic quantities (**H** and **B**) in electromagnetic units. The constant c in (1) and (2) relates the unit of charge in the two systems; it is the velocity of light in the vacuum and is approximately equal to 3×10^{10} cm/s. (A more accurate value is given in §1.2.)

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

They are supplemented by two scalar relations:

$$\operatorname{div}\mathbf{D} = 4\pi\rho, \tag{3}$$

$$\operatorname{div}\mathbf{B} = 0. \tag{4}$$

Eq. (3) may be regarded as a defining equation for the electric charge density ρ and (4) may be said to imply that no free magnetic poles exist.

From (1) it follows (since div curl $\equiv 0$) that

$$\operatorname{div}\mathbf{j} = -\frac{1}{4\pi}\operatorname{div}\dot{\mathbf{D}},$$

or, using (3),

$$\frac{\partial\rho}{\partial t} + \operatorname{div}\mathbf{j} = 0. \tag{5}$$

By analogy with a similar relation encountered in hydrodynamics, (5) is called the *equation of continuity*. It expresses the fact that the charge is conserved in the neighbourhood of any point. For if one integrates (5) over any region of space, one obtains, with the help of Gauss' theorem,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\rho\,\mathrm{d}V + \int\mathbf{j}\cdot\mathbf{n}\,\mathrm{d}S = 0, \tag{6}$$

the second integral being taken over the surface bounding the region and the first throughout the volume, **n** denoting the unit outward normal. This equation implies that the total charge

$$e = \int\rho\,\mathrm{d}V \tag{7}$$

contained within the domain can only increase on account of the flow of electric current

$$\mathcal{J} = \int\mathbf{j}\cdot\mathbf{n}\,\mathrm{d}S. \tag{8}$$

If all the field quantities are independent of time, and if, moreover, there are no currents ($\mathbf{j} = 0$), the field is said to be *static*. If all the field quantities are time independent, but currents are present ($\mathbf{j} \neq 0$), one speaks of a *stationary field*. In optical fields the field vectors are very rapidly varying functions of time, but the sources of the field are usually such that, when averages over any macroscopic time interval are considered rather than the instantaneous values, the properties of the field are found to be independent of the instant of time at which the average is taken. The word *stationary* is often used in a wider sense to describe a field of this type. An example is a field constituted by the steady flux of radiation (say from a distant star) through an optical system.

1.1.2 Material equations

The Maxwell equations (1)–(4) connect the five basic quantities **E**, **H**, **B**, **D** and **j**. To allow a unique determination of the field vectors from a given distribution of currents

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

and charges, these equations must be supplemented by relations which describe the behaviour of substances under the influence of the field. These relations are known as *material equations*[*] (or *constitutive relations*). In general they are rather complicated; but if the field is time-harmonic (see §1.4.3), and if the bodies are at rest, or in very slow motion relative to each other, and if the material is *isotropic* (i.e. when its physical properties at each point are independent of direction), they take usually the relatively simple form[†]

$$\mathbf{j} = \sigma \mathbf{E}, \tag{9}$$

$$\mathbf{D} = \varepsilon \mathbf{E}, \tag{10}$$

$$\mathbf{B} = \mu \mathbf{H}. \tag{11}$$

Here σ is called the *specific conductivity*, ε is known as the *dielectric constant* (or permittivity) and μ is called the *magnetic permeability*.

Eq. (9) is the differential form of Ohm's law. Substances for which $\sigma \neq 0$ (or more precisely is not negligibly small; the precise meaning of this cannot, however, be discussed here) are called *conductors*. Metals are very good conductors, but there are other classes of good conducting materials such as ionic solutions in liquids and also in solids. In metals the conductivity decreases with increasing temperature. However, in other classes of materials, known as *semiconductors* (e.g. germanium), conductivity increases with temperature over a wide range.

Substances for which σ is negligibly small are called *insulators* or *dielectrics*. Their electric and magnetic properties are then completely determined by ε and μ. For most substances the magnetic permeability μ is practically unity. If this is not the case, i.e. if μ differs appreciably from unity, we say that the substance is *magnetic*. In particular, if $\mu > 1$, the substance is said to be *paramagnetic* (e.g. platinum, oxygen, nitrogen dioxide), while if $\mu < 1$ it is said to be *diamagnetic* (e.g. bismuth, copper, hydrogen, water).

If the fields are exceptionally strong, such as are obtained, for example, by focusing light that is generated by a laser, the right-hand sides of the material equations may have to be supplemented by terms involving components of the field vectors in powers higher than the first.[‡]

In many cases the quantities σ, ε and μ will be independent of the field strengths; in other cases, however, the behaviour of the material cannot be described in such a simple way. Thus, for example, in a gas of free ions the current, which is determined

[*] There is an alternative way of describing the behaviour of matter. Instead of the quantities $\varepsilon = D/E$, $\mu = B/H$ one considers the differences $\mathbf{D} - \mathbf{E}$ and $\mathbf{B} - \mathbf{H}$; these have a simpler physical significance and will be discussed in Chapter II.

[†] The more general relations, applicable also to moving bodies, are studied in the theory of relativity. We shall only need the following result from the more general theory: that in the case of moving charges there is, in addition to the conduction current $\sigma\mathbf{E}$, a convection current $\rho\mathbf{v}$, where $\mathbf{v}$ is the velocity of the moving charges and ρ the charge density (cf. p. 9).

[‡] Nonlinear relationship between the displacement vector **D** and the electric field **E** was first demonstrated in this way by P. A. Franken, A. E. Hill, C. W. Peters and G. Weinrich, *Phys. Rev. Lett.*, **7** (1961), 118.

For systematic treatments of nonlinear effects see N. Bloembergen, *Nonlinear Optics* (New York, W. A. Benjamin, Inc., 1965), P. N. Butcher and D. Cotter, *The Elements of Nonlinear Optics* (Cambridge, Cambridge University Press, 1990) or R. W. Boyd, *Nonlinear Optics* (Boston, Academic Press, 1992).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

by the mean speed of the ions, depends, at any moment, not on the instantaneous value of **E**, but on all its previous values. Again, in so-called *ferromagnetic* substances (substances which are very highly magnetic, e.g. iron, cobalt and nickel) the value of the magnetic induction **B** is determined by the past history of the field **H** rather than by its instantaneous value. The substance is then said to exhibit *hysteresis*. A similar history-dependence will be found for the electric displacement in certain dielectric materials. Fortunately hysteretic effects are rarely significant for the high-frequency field encountered in optics.

In the main part of this book we shall study the propagation in substances which light can penetrate without appreciable weakening (e.g. air, glass). Such substances are said to be *transparent* and must be electrical nonconductors ($\sigma = 0$), since conduction implies the evolution of Joule heat (see §1.1.4) and therefore loss of electromagnetic energy. Optical properties of conducting media will be discussed in Chapter XIV.

1.1.3 Boundary conditions at a surface of discontinuity

Maxwell's equations were only stated for regions of space throughout which the physical properties of the medium (characterized by ε and μ) are continuous. In optics one often deals with situations in which the properties change abruptly across one or more surfaces. The vectors **E**, **H**, **B** and **D** may then be expected also to become discontinuous, while ρ and **j** will degenerate into corresponding surface quantities. We shall derive relations describing the transition across such a discontinuity surface.

Let us replace the sharp discontinuity surface T by a thin transition layer within which ε and μ vary rapidly but continuously from their values near T on one side to their value near T on the other. Within this layer we construct a small near-cylinder, bounded by a stockade of normals to T; roofed and floored by small areas δA_1 and δA_2 on each side of T, at constant distance from it, measured along their common normal (Fig. 1.1). Since **B** and its derivatives may be assumed to be continuous throughout this cylinder, we may apply Gauss' theorem to the integral of div **B** taken throughout the volume of the cylinder and obtain, from (4),

$$\int \operatorname{div} \mathbf{B}\, \mathrm{d}V = \int \mathbf{B} \cdot \mathbf{n}\, \mathrm{d}S = 0; \tag{12}$$

the second integral is taken over the surface of the cylinder, and **n** is the unit outward normal.

Since the areas δA_1 and δA_2 are assumed to be small, **B** may be considered to have constants values $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ on δA_1 and δA_2, and (12) may then be replaced by

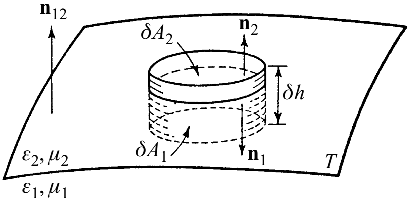

Fig. 1.1 Derivation of boundary conditions for the normal components of **B** and **D**.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\mathbf{B}^{(1)} \cdot \mathbf{n}_1 \delta A_1 + \mathbf{B}^{(2)} \cdot \mathbf{n}_2 \delta A_2 + \text{contribution from walls} = 0. \tag{13}$$

If the height δh of the cylinder decreases towards zero, the transition layer shrinks into the surface and the contribution from the walls of the cylinder tends to zero, provided that there is no surface flux of magnetic induction. Such flux never occurs, and consequently in the limit,

$$(\mathbf{B}^{(1)} \cdot \mathbf{n}_1 + \mathbf{B}^{(2)} \cdot \mathbf{n}_2)\delta A = 0, \tag{14}$$

δA being the area in which the cylinder intersects T. If $\mathbf{n}_{12}$ is the unit normal pointing from the first into the second medium, then $\mathbf{n}_1 = -\mathbf{n}_{12}$, $\mathbf{n}_2 = \mathbf{n}_{12}$ and (14) gives

$$\mathbf{n}_{12} \cdot (\mathbf{B}^{(2)} - \mathbf{B}^{(1)}) = 0, \tag{15}$$

i.e. *the normal component of the magnetic induction is continuous across the surface of discontinuity.*

The electric displacement $\mathbf{D}$ may be treated in a similar way, but there will be an additional term if charges are present. In place of (12) we now have from (3)

$$\int \operatorname{div} \mathbf{D} \, \mathrm{d}V = \int \mathbf{D} \cdot \mathbf{n} \, \mathrm{d}S = 4\pi \int \rho \, \mathrm{d}V. \tag{16}$$

As the areas δA_1 and δA_2 shrink together, the total charge remains finite, so that the volume density becomes infinite. Instead of the volume charge density ρ the concept of *surface charge density* $\hat{\rho}$ must then be used. It is defined by*

$$\lim_{\delta h \to 0} \int \rho \, \mathrm{d}V = \int \hat{\rho} \, \mathrm{d}A. \tag{17}$$

We shall also need later the concept of *surface current density* $\hat{\mathbf{j}}$, defined in a similar way:

$$\lim_{\delta h \to 0} \int \mathbf{j} \, \mathrm{d}V = \int \hat{\mathbf{j}} \, \mathrm{d}A. \tag{18}$$

If the area δA and the height δh are taken sufficiently small, (16) gives

$$\mathbf{D}^{(1)} \cdot \mathbf{n}_1 \, \delta A_1 + \mathbf{D}^{(2)} \cdot \mathbf{n}_2 \, \delta A_2 + \text{contribution from walls} = 4\pi\hat{\rho} \, \delta A.$$

The contribution from the walls tends to zero with δh, and we therefore obtain in the limit as $\delta h \to 0$,

$$\mathbf{n}_{12} \cdot (\mathbf{D}^{(2)} - \mathbf{D}^{(1)}) = 4\pi\hat{\rho}, \tag{19}$$

i.e. *in the presence of a layer of surface charge density $\hat{\rho}$ on the surface, the normal component of the electric displacement changes abruptly across the surface, by an amount equal to $4\pi\hat{\rho}$.*

* For later purposes we note a representation of the surface charge density and the surface current density in terms of the Dirac delta function (see Appendix IV). If the equation of the surface of discontinuity is $F(x, y, z) = 0$, then

$$\rho = \hat{\rho} |\operatorname{grad} F| \delta(F), \tag{17a}$$

$$\mathbf{j} = \hat{\mathbf{j}} |\operatorname{grad} F| \delta(F). \tag{18a}$$

These relations can immediately be verified by substituting from (17a) and (18a) into (17) and (18) and using the relation $\mathrm{d}F = |\operatorname{grad} F| \mathrm{d}h$ and the sifting property of the delta function.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Next, we examine the behaviour of the tangential components. Let us replace the sharp discontinuity surface by a continuous transition layer. We also replace the cylinder of Fig. 1.1 by a 'rectangular' area with sides parallel and perpendicular to T (Fig. 1.2).

Let $\mathbf{b}$ be the unit vector perpendicular to the plane of the rectangle. Then it follows from (2) and from Stokes' theorem that

$$\int \operatorname{curl} \mathbf{E} \cdot \mathbf{b}\, \mathrm{d}S = \int \mathbf{E} \cdot \mathrm{d}\mathbf{r} = -\frac{1}{c}\int \dot{\mathbf{B}} \cdot \mathbf{b}\, \mathrm{d}S, \tag{20}$$

the first and third integrals being taken throughout the area of the rectangle, and the second along its boundary. If the lengths P_1Q_1 ($= \delta s_1$), and P_2Q_2 ($= \delta s_2$) are small, $\mathbf{E}$ may be replaced by constant values $\mathbf{E}^{(1)}$ and $\mathbf{E}^{(2)}$ along each of these segments. Similarly $\dot{\mathbf{B}}$ may be replaced by a constant value. Eq. (20) then gives

$$\mathbf{E}^{(1)} \cdot \mathbf{t}_1\, \delta s_1 + \mathbf{E}^{(2)} \cdot \mathbf{t}_2\, \delta s_2 + \text{contribution from ends} = -\frac{1}{c}\dot{\mathbf{B}} \cdot \mathbf{b}\, \delta s \delta h, \tag{21}$$

where δs is the line element in which the rectangle intersects the surface. If now the height of the rectangle is gradually decreased, the contribution from the ends P_1P_2 and Q_1Q_2 will tend to zero, provided that $\mathbf{E}$ does not in the limit acquire sufficiently sharp singularities; this possibility will be excluded. Assuming also that $\dot{\mathbf{B}}$ remains finite, we obtain in the limit as $\delta h \to 0$,

$$(\mathbf{E}^{(1)} \cdot \mathbf{t}_1 + \mathbf{E}^{(2)} \cdot \mathbf{t}_2)\delta s = 0. \tag{22}$$

If $\mathbf{t}$ is the unit tangent along the surface, then (see Fig. 1.2) $\mathbf{t}_1 = -\mathbf{t} = -\mathbf{b} \times \mathbf{n}_{12}$, $\mathbf{t}_2 = \mathbf{t} = \mathbf{b} \times \mathbf{n}_{12}$, and (22) gives

$$\mathbf{b} \cdot [\mathbf{n}_{12} \times (\mathbf{E}^{(2)} - \mathbf{E}^{(1)})] = 0.$$

Since the orientation of the rectangle and consequently that of the unit vector $\mathbf{b}$ is arbitrary, it follows that

$$\mathbf{n}_{12} \times (\mathbf{E}^{(2)} - \mathbf{E}^{(1)}) = 0, \tag{23}$$

i.e. *the tangential component of the electric vector is continuous across the surface.*

Finally consider the behaviour of the tangential component of the magnetic vector. The analysis is similar, but there is an additional term if currents are present. In place of (21) we now have

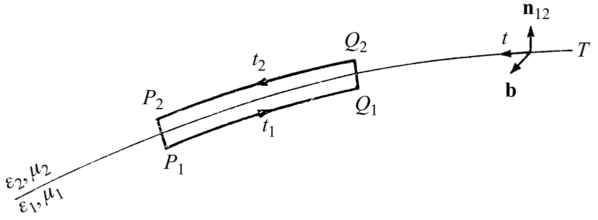

Fig. 1.2 Derivation of boundary conditions for the tangential components of **E** and **H**.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\mathbf{H}^{(1)} \cdot \mathbf{t}_1 \, \delta s_1 + \mathbf{H}^{(2)} \cdot \mathbf{t}_2 \, \delta s_2 + \text{contribution from ends} = \frac{1}{c} \dot{\mathbf{D}} \cdot \mathbf{b} \, \delta s \, \delta h + \frac{4\pi}{c} \hat{\mathbf{j}} \cdot \mathbf{b} \, \delta s. \tag{24}$$

On proceeding to the limit $\delta h \to 0$ as before, we obtain

$$\mathbf{n}_{12} \times (\mathbf{H}^{(2)} - \mathbf{H}^{(1)}) = \frac{4\pi}{c} \hat{\mathbf{j}}. \tag{25}$$

From (25) is follows that *in the presence of a surface current of density* $\hat{\mathbf{j}}$, *the tangential component (considered as a vector quantity) of the magnetic vector changes abruptly, its discontinuity being* $(4\pi/c)\hat{\mathbf{j}} \times \mathbf{n}_{12}$.

Apart from discontinuities due to the abrupt changes in the physical properties of the medium, the field vectors may also be discontinuous because of the presence of a source which begins to radiate at a particular instant of time $t = t_0$. The disturbance then spreads into the surrounding space, and at any later instant $t_1 > t_0$ will have filled a well-defined region. Across the (moving) boundary of this region, the field vectors will change abruptly from finite values on the boundary to the value zero outside it.

The various cases of discontinuity may be covered by rewriting Maxwell's equations in an integral form.* The general discontinuity conditions may also be written in the form of simple difference equations; a derivation of these equations is given in Appendix VI.

1.1.4 The energy law of the electromagnetic field

Electromagnetic theory interprets the light intensity as the energy flux of the field. It is therefore necessary to recall the energy law of Maxwell's theory.

From (1) and (2) it follows that

$$\mathbf{E} \cdot \operatorname{curl} \mathbf{H} - \mathbf{H} \cdot \operatorname{curl} \mathbf{E} = \frac{4\pi}{c} \mathbf{j} \cdot \mathbf{E} + \frac{1}{c} \mathbf{E} \cdot \dot{\mathbf{D}} + \frac{1}{c} \mathbf{H} \cdot \dot{\mathbf{B}}. \tag{26}$$

Also, by a well-known vector identity, the term on the left may be expressed as the divergence of the vector product of **H** and **E**:

$$\mathbf{E} \cdot \operatorname{curl} \mathbf{H} - \mathbf{H} \cdot \operatorname{curl} \mathbf{E} = -\operatorname{div} (\mathbf{E} \times \mathbf{H}). \tag{27}$$

From (26) and (27) we have that

$$\frac{1}{c} (\mathbf{E} \cdot \dot{\mathbf{D}} + \mathbf{H} \cdot \dot{\mathbf{B}}) + \frac{4\pi}{c} \mathbf{j} \cdot \mathbf{E} + \operatorname{div} (\mathbf{E} \times \mathbf{H}) = 0. \tag{28}$$

When we multiply this equation by $c/4\pi$, integrate throughout an arbitrary volume and apply Gauss's theorem, this gives

$$\frac{1}{4\pi} \int (\mathbf{E} \cdot \dot{\mathbf{D}} + \mathbf{H} \cdot \dot{\mathbf{B}}) \mathrm{d}V + \int \mathbf{j} \cdot \mathbf{E} \, \mathrm{d}V + \frac{c}{4\pi} \int (\mathbf{E} \times \mathbf{H}) \cdot \mathbf{n} \, \mathrm{d}S = 0, \tag{29}$$

where the last integral is taken over the boundary of the volume, **n** being the unit outward normal.

The relation (29) is a direct consequence of Maxwell's equations and is therefore

* See, for example, A. Sommerfeld, *Electrodynamics* (New York, Academic Press, 1952), p. 11; or J. A. Stratton, *Electromagnetic Theory* (New York, McGraw-Hill, 1941), p. 6.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

valid whether or not the material equations (9)–(11) hold. It represents, as will be seen, the *energy law* of an electromagnetic field. We shall discuss it here only for the case where the material equations (9)–(11) are satisfied. Generalizations to anisotropic media, where the material equations are of a more complicated form, will be considered later (Chapter XV).

We have, on using the material equations,

$$\left.\begin{aligned}\frac{1}{4\pi}(\mathbf{E}\cdot\dot{\mathbf{D}}) &= \frac{1}{4\pi}\mathbf{E}\cdot\frac{\partial}{\partial t}(\varepsilon\mathbf{E}) = \frac{1}{8\pi}\frac{\partial}{\partial t}(\varepsilon\mathbf{E}^2) = \frac{1}{8\pi}\frac{\partial}{\partial t}(\mathbf{E}\cdot\mathbf{D}),\\ \frac{1}{4\pi}(\mathbf{H}\cdot\dot{\mathbf{B}}) &= \frac{1}{4\pi}\mathbf{H}\cdot\frac{\partial}{\partial t}(\mu\mathbf{H}) = \frac{1}{8\pi}\frac{\partial}{\partial t}(\mu\mathbf{H}^2) = \frac{1}{8\pi}\frac{\partial}{\partial t}(\mathbf{H}\cdot\mathbf{B}).\end{aligned}\right\} \tag{30}$$

Setting

$$w_e = \frac{1}{8\pi}\mathbf{E}\cdot\mathbf{D}, \qquad w_m = \frac{1}{8\pi}\mathbf{H}\cdot\mathbf{B}, \tag{31}$$

and

$$W = \int (w_e + w_m)\,\mathrm{d}V, \tag{32}$$

(29) becomes,

$$\frac{\mathrm{d}W}{\mathrm{d}t} + \int \mathbf{j}\cdot\mathbf{E}\,\mathrm{d}V + \frac{c}{4\pi}\int(\mathbf{E}\times\mathbf{H})\cdot\mathbf{n}\,\mathrm{d}S = 0. \tag{33}$$

We shall show that W represents the total energy contained within the volume, so that w_e may be identified with the *electric energy density* and w_m with the *magnetic energy density* of the field.*

To justify the interpretation of W as the total energy we have to show that, for a closed system (i.e. one in which the field on the boundary surface may be neglected), the change in W as defined above is due to the work done by the field on the material charged bodies which are embedded in it. It suffices to do this for slow motion of the material bodies, which themselves may be assumed to be so small that they can be regarded as point charges e_k $(k = 1, 2, \ldots)$. Let the velocity of the charge e_k be $\mathbf{v}_k$ $(|\mathbf{v}_k| \ll c)$.

The force exerted by a field $(\mathbf{E}, \mathbf{B})$ on a charge e moving with velocity $\mathbf{v}$ is given by the so-called *Lorentz law*,

$$\mathbf{F} = e\left(\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B}\right), \tag{34}$$

which is based on experience. It follows that if all the charges e_k are displaced by $\delta\mathbf{x}_k$ $(k = 1, 2, \ldots)$ in time δt, the total work done is

* In the general case the densities are defined by the expressions

$$w_e = \frac{1}{4\pi}\int \mathbf{E}\cdot\mathrm{d}\mathbf{D}, \qquad w_m = \frac{1}{4\pi}\int \mathbf{H}\cdot\mathrm{d}\mathbf{B}.$$

When the relationship between **E** and **D** and between **H** and **B** is linear, as here assumed, these expressions reduce to (31).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\delta A = \sum_k \mathbf{F}_k \cdot \delta \mathbf{x}_k = \sum_k e_k \left(\mathbf{E}_k + \frac{1}{c} \mathbf{v}_k \times \mathbf{B} \right) \cdot \delta \mathbf{x}_k$$
$$= \sum_k e_k \mathbf{E}_k \cdot \delta \mathbf{x}_k = \sum_k e_k \mathbf{E}_k \cdot \mathbf{v}_k \, \delta t,$$

since $\delta \mathbf{x}_k = \mathbf{v}_k \, \delta t$. If the number of charged particles is large, we can consider the distribution to be continuous. We introduce the charge density ρ (i.e. total charge per unit volume) and the last equation becomes

$$\delta A = \delta t \int \rho \mathbf{v} \cdot \mathbf{E} \, \mathrm{d}V, \tag{35}$$

the integration being carried throughout an arbitrary volume. Now the velocity $\mathbf{v}$ does not appear explicitly in Maxwell's equations, but it may be introduced by using an experimental result found by Röntgen*, according to which a *convection current* (i.e. a set of moving charges) has the same electromagnetic effect as a *conduction current* in a wire. Hence the current density $\mathbf{j}$ appearing in Maxwell's equations can be split into two parts

$$\mathbf{j} = \mathbf{j}_{\mathrm{c}} + \mathbf{j}_v, \tag{36}$$

where

$$\mathbf{j}_{\mathrm{c}} = \sigma \mathbf{E}$$

is the conduction current density, and

$$\mathbf{j}_v = \rho \mathbf{v}$$

represents the convection current density. (35) may therefore be written as

$$\delta A = \delta t \int \mathbf{j}_v \cdot \mathbf{E} \, \mathrm{d}V. \tag{37}$$

Let us now define a vector $\mathbf{S}$ and a scalar Q by the relations

$$\mathbf{S} = \frac{c}{4\pi} (\mathbf{E} \times \mathbf{H}), \tag{38}$$

$$Q = \int \mathbf{j}_{\mathrm{c}} \cdot \mathbf{E} \, \mathrm{d}V = \int \sigma \mathbf{E}^2 \, \mathrm{d}V. \tag{39}$$

Then by (35) and (36)

$$\int \mathbf{j} \cdot \mathbf{E} \, \mathrm{d}V = Q + \int \mathbf{j}_v \cdot \mathbf{E} \, \mathrm{d}V$$
$$= Q + \frac{\delta A}{\delta t}, \tag{40}$$

where the second function is not, of course, a total derivative of a space-time function. Eq. (33) now takes the form

* W. C. Röntgen, *Ann. d. Physik*, **35** (1888), 264; **40** (1890), 93.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\frac{\mathrm{d}W}{\mathrm{d}t} = -\frac{\delta A}{\delta t} - Q - \int \mathbf{S}\cdot\mathbf{n}\,\mathrm{d}S. \tag{41}$$

For a nonconductor ($\sigma = 0$) we have that $Q = 0$. Assume also that the boundary surface is so far away that we can neglect the field on it, due to the electromagnetic processes inside; then $\int \mathbf{S}\cdot\mathbf{n}\,\mathrm{d}S = 0$, and integration of (41) gives

$$W + A = \text{constant}. \tag{42}$$

Hence, for an isolated system, the increase of W per unit time is due to the work done on the system during this time. This result justifies our definition of electromagnetic energy by means of (32).

The term Q represents the resistive dissipation of energy (called *Joule's heat*) in a conductor ($\sigma \neq 0$). According to (41) there is a further decrease in energy if the field extends to the boundary surface. The surface integral must therefore represent the flow of energy across this boundary surface. The vector $\mathbf{S}$ is known as the *Poynting vector* and represents the amount of energy which crosses per second a unit area normal to the directions of $\mathbf{E}$ and $\mathbf{H}$.

It should be noted that the interpretation of $\mathbf{S}$ as energy flow (more precisely as the density of the flow) is an abstraction which introduces a certain degree of arbitrariness. For the quantity which is physically significant is, according to (41), not $\mathbf{S}$ itself, but the integral of $\mathbf{S}\cdot\mathbf{n}$ taken over a closed surface. Clearly, from the value of the integral, no unambiguous conclusion can be drawn about the detailed distribution of $\mathbf{S}$, and alternative definitions of the energy flux density are therefore possible. One can always add to $\mathbf{S}$ the curl of an arbitrary vector, since such a term will not contribute to the surface integral as can be seen from Gauss' theorem and the identity $\operatorname{div}\operatorname{curl} \equiv 0$.* However, when the definition has been applied cautiously, in particular for averages over small but finite regions of space or time, no contradictions with experiments have been found. We shall therefore accept the above definition in terms of the Poynting vector of the density of the energy flow.

Finally we note that in a nonconducting medium ($\sigma = 0$) where no mechanical work is done ($A = 0$), the energy law may be written in the form of a hydrodynamical continuity equation for noncompressible fluids:

$$\frac{\partial w}{\partial t} + \operatorname{div}\mathbf{S} = 0, \qquad (w = w_e + w_m). \tag{43}$$

A description of propagation of light in terms of a hydrodynamical model is often helpful, particularly in the domain of geometrical optics and in connection with scalar diffraction fields, as it gives a picture of the energy transport in a simple and graphic manner. In optics, the (averaged) Poynting vector is the chief quantity of interest. The magnitude of the Poynting vector is a measure of the light intensity, and its direction represents the direction of propagation of the light.

* According to modern theories of fields the arbitrariness is even greater, allowing for alternative expressions for both the energy density and the energy flux, but consistent with the change of the Lagrangian density of the field by the addition of a four-divergence. For a discussion of this subject see, for example, G. Wentzel, *Quantum Theory of Fields* (New York, Interscience Publishers, 1949), especially §2 or J. D. Jackson, *Classical Electrodynamics* (New York, J. Wiley and Sons, 2nd ed. 1975), Sec. 12.10, especially p. 602.

1.2 The wave equation and the velocity of light

Maxwell's equations relate the field vectors by means of simultaneous differential equations. On elimination we obtain differential equations which each of the vectors must separately satisfy. We shall confine our attention to that part of the field which contains no charges or currents, i.e. where $\mathbf{j} = 0$ and $\rho = 0$.

We substitute for $\mathbf{B}$ from the material equation §1.1 (11) into the second Maxwell equation §1.1 (2), divide both sides by μ and apply the operator curl. This gives

$$\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl}\mathbf{E}\right) + \frac{1}{c}\operatorname{curl}\dot{\mathbf{H}} = 0. \tag{1}$$

Next we differentiate the first Maxwell equation §1.1 (1) with respect to time, use the material equation §1.1 (10) for $\mathbf{D}$, and eliminate $\operatorname{curl}\dot{\mathbf{H}}$ between the resulting equation and (1); this gives

$$\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl}\mathbf{E}\right) + \frac{\varepsilon}{c^2}\ddot{\mathbf{E}} = 0. \tag{2}$$

If we use the identities $\operatorname{curl} u\mathbf{v} = u\operatorname{curl}\mathbf{v} + (\operatorname{grad} u)\times\mathbf{v}$ and $\operatorname{curl}\operatorname{curl} = \operatorname{grad}\operatorname{div} - \nabla^2$, (2) becomes

$$\nabla^2\mathbf{E} - \frac{\varepsilon\mu}{c^2}\ddot{\mathbf{E}} + (\operatorname{grad}\ln\mu)\times\operatorname{curl}\mathbf{E} - \operatorname{grad}\operatorname{div}\mathbf{E} = 0. \tag{3}$$

Also from §1.1 (3), using again the material equation for $\mathbf{D}$ and applying the identity $\operatorname{div} u\mathbf{v} = u\operatorname{div}\mathbf{v} + \mathbf{v}\cdot\operatorname{grad} u$ we find

$$\varepsilon\operatorname{div}\mathbf{E} + \mathbf{E}\cdot\operatorname{grad}\varepsilon = 0. \tag{4}$$

Hence (3) may be written in the form

$$\nabla^2\mathbf{E} - \frac{\varepsilon\mu}{c^2}\ddot{\mathbf{E}} + (\operatorname{grad}\ln\mu)\times\operatorname{curl}\mathbf{E} + \operatorname{grad}(\mathbf{E}\cdot\operatorname{grad}\ln\varepsilon) = 0. \tag{5}$$

In a similar way we obtain an equation for $\mathbf{H}$ alone:

$$\nabla^2\mathbf{H} - \frac{\varepsilon\mu}{c^2}\ddot{\mathbf{H}} + (\operatorname{grad}\ln\varepsilon)\times\operatorname{curl}\mathbf{H} + \operatorname{grad}(\mathbf{H}\cdot\operatorname{grad}\ln\mu) = 0. \tag{6}$$

In particular, if the medium is homogeneous, $\operatorname{grad}\log\varepsilon = \operatorname{grad}\ln\mu = 0$, and (5) and (6) reduce to

$$\nabla^2\mathbf{E} - \frac{\varepsilon\mu}{c^2}\ddot{\mathbf{E}} = 0, \qquad \nabla^2\mathbf{H} - \frac{\varepsilon\mu}{c^2}\ddot{\mathbf{H}} = 0. \tag{7}$$

These are standard equations of wave motion and suggest the existence of electromagnetic waves propagated with a velocity*

$$v = c/\sqrt{\varepsilon\mu}. \tag{8}$$

* The concept of a velocity of an electromagnetic wave has actually an unambiguous meaning only in connection with waves of very simple kind, e.g. plane waves. That v does not represent the velocity of propagation of an arbitrary solution of (7) is obvious if we bear in mind that these equations also admit standing waves as solutions.

In this introductory section it is assumed that the reader is familiar with the concept of a plane wave, and we regard v as the velocity with which such a wave advances. The mathematical representation of a plane wave will be discussed in §1.3 and §1.4.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

The constant c was first determined by R. Kohlrausch and W. Weber in 1856 from the ratio of the values of the capacity of a condenser measured in electrostatic and electromagnetic units, and it was found to be identical with the velocity of light in free space. Using this result, Maxwell developed his electromagnetic theory of light, predicting the existence of electromagnetic waves; the correctness of this prediction was confirmed by the celebrated experiments of H. Hertz (see Historical introduction).

As in all wave theories of light, the elementary process which produces the optical impression is regarded as being a harmonic wave in space-time (studied in its simplest form in §1.3 and §1.4). If its frequency is in the range from $4 \times 10^{14}\ \mathrm{s}^{-1}$ to $7.5 \times 10^{14}\ \mathrm{s}^{-1}$ (approximately) it gives rise to the psychological impression of a definite colour. (The opposite, however, is not true: coloured light of a certain subjective quality may be a composition of harmonic waves of very different frequency distributions.) The actual connection between colour and frequency is very involved and will not be studied in this book.*

The first determination of the velocity of light† was made by Römer in 1675 from observations of the eclipses of the first satellite of Jupiter and later in a different way (from aberration of fixed stars) by Bradley (1728).

The first measurements of the velocity of light from terrestrial sources were carried out by Fizeau in 1849. It is necessary to employ a modulator, which marks off a portion of the beam‡ and for this purpose Fizeau used a rotation wheel. Later methods employed rotating mirrors or electronic shutters. The rotating mirror method was suggested by Wheatstone in 1834 and was used by Foucault in 1860. It was later systematically developed over a period of many years by Michelson. The average value based on about 200 measurements by Michelson gave c as 299,796 km/s. An optical shutter method employing a Kerr cell was developed by Karolus and Mittelstaedt (1928), Anderson (1937) and Hüttel (1940). The values of c obtained from these measurements are in excellent agreement with those based on indirect methods, such as determinations from the ratio of an electric charge measured in electrostatic and electromagnetic units; for example Rosa and Dorsey (1907) in this way found c as 299,784 km/s. Measurements of the velocity of electromagnetic waves on wires carried out by Mercier (1923) gave the value of c equal to 299,782 km/s. The value adopted by the Fifteenth General Conference of Weights and Measures§ is

$$c = 299{,}792.458\ \mathrm{km/s}. \tag{9}$$

The close agreement between the values of c obtained from measurements of very different kinds (and in some cases using radiation whose frequencies differ by a factor of hundreds of thousands from those used in the optical measurements) gives a striking confirmation of Maxwell's theory.

The dielectric constant ε is usually greater than unity, and μ is practically equal to

* The sensitivity of the human eye to different colours is, however, briefly discussed in §4.8.1.

† For a description of the methods used for determination of the velocity of light, see for example, E. Bergstrand, *Encyclopedia of Physics*, ed. S. Flügge, Vol. 24 (Berlin, Springer, 1956), p. 1.

Detailed analysis of the results obtained by different methods is also given by R. T. Birge in *Rep. Progr. Phys.* (London, The Physical Society), **8** (1941), 90.

‡ Such determinations give essentially the group velocity (see §1.3.4). The difference between the group velocity and the phase velocity in air at standard temperature and pressure is about 1 part in 50,000.

§ Conférence Générale des Poids et Mesures, XV, Paris, 1975, Comptes Rendus des Séances (Paris, Bureau International des Poids et Mesures, 1976).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

unity for transparent substances, so that the velocity v is then according to (8) smaller than the vacuum velocity c. This conclusion was first demonstrated experimentally for propagation of light in water in 1850 by Foucault and Fizeau.

The value of v is not as a rule determined directly, but only relative to c, with the help of the law of refraction. According to this law, if a plane electromagnetic wave falls on to a plane boundary between two homogeneous media, the sine of the angle θ_1 between the normal to the incident wave and the normal to the surface bears a constant ratio to the sine of the angle θ_2 between the normal of the refracted wave and the surface normal (Fig. 1.3), this constant ratio being equal to the ratio of the velocities v_1 and v_2 of propagation in the two media:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}. \tag{10}$$

This result will be derived in §1.5. Here we only note that it is equivalent to the assumption that the wave-front, though it has a kink at the boundary, is continuous, so that the line of intersection between the incident wave and the boundary travels at the same speed (v', say) as the line of intersection between the refracted wave and the boundary. We then have

$$v_1 = v' \sin\theta_1, \qquad v_2 = v' \sin\theta_2, \tag{11}$$

from which, on elimination of v', (10) follows. This argument, in a slightly more elaborate form, is often given as an illustration of Huygens' construction (§3.3).

The value of the constant ratio in (10) is usually denoted by n_{12} and is called the *refractive index*, for refraction from the first into the second medium. We also define an '*absolute refractive index*' n of a medium; it is the refractive index for refraction from vacuum into that medium,

$$n = \frac{c}{v}. \tag{12}$$

If n_1 and n_2 are the absolute refractive indices of two media, the (relative) refractive index n_{12} for refraction from the first into the second medium then is

$$n_{12} = \frac{n_2}{n_1} = \frac{v_1}{v_2}. \tag{13}$$

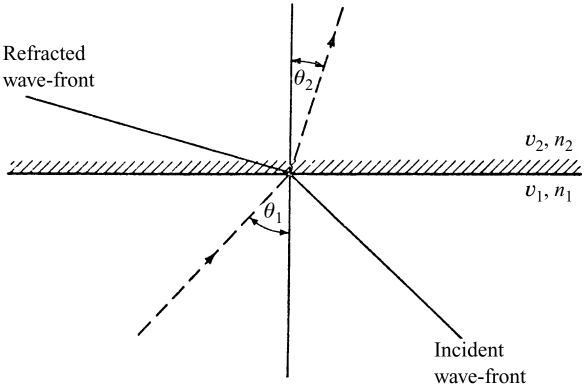

Fig. 1.3 Illustrating the refraction of a plane wave.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Table 1.1. *Refractive indices and static dielectric constants of certain gases*

	n (yellow light)	$\sqrt{\varepsilon}$
Air	1.000294	1.000295
Hydrogen H_2	1.000138	1.000132
Carbon dioxide CO_2	1.000449	1.000473
Carbon monoxide CO	1.000340	1.000345

Table 1.2. *Refractive indices and static dielectric constants of certain liquids*

	n (yellow light)	$\sqrt{\varepsilon}$
Methyl alcohol CH_3OH	1.34	5.7
Ethyl alcohol C_2H_5OH	1.36	5.0
Water H_2O	1.33	9.0

Comparison of (12) and (8) gives Maxwell's formula:

$$n = \sqrt{\varepsilon\mu}. \tag{14}$$

Since for all substances with which we shall be concerned, μ is effectively unity (nonmagnetic substances), the refractive index should then be equal to the square root of the dielectric constant, which has been assumed to be a constant of the material. On the other hand, well-known experiments on prismatic colours, first carried out by Newton, show that the index of refraction depends on the colour, i.e. on the frequency of the light. If we are to retain Maxwell's formula, it must be supposed that ε is not a constant characteristic of the material, but is a function of the frequency of the field. The dependence of ε on frequency can only be treated by taking into account the atomic structure of matter, and will be briefly discussed in §2.3.

Maxwell's formula (with ε equal to the static dielectric constant) gives a good approximation for such substances as gases with a simple chemical structure which do not disperse light substantially, i.e. for those whose optical properties do not strongly depend on the colour of the light. Results of some early measurements for such gases, carried out by L. Boltzmann,* are given in Table 1.1. Eq. (14) also gives a good approximation for liquid hydrocarbons; for example benzene C_6H_6 has $n = 1.482$ for yellow light whilst $\sqrt{\varepsilon} = 1.489$. On the other hand, there is a strong deviation from the formula for many solid bodies (e.g. glasses), and for some liquids, as illustrated in Table 1.2.

1.3 Scalar waves

In a homogeneous medium in regions free of currents and charges, each rectangular component $V(\mathbf{r}, t)$ of the field vectors satisfies, according to §1.2 (7), the homogeneous wave equation

* L. Boltzmann, *Wien. Ber.*, **69** (1874), 795; *Pogg. Ann.*, **155** (1875), 403; *Wiss. Abh. Physik-techn. Reichsanst.*, **1**, Nr. 26, 537.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\nabla^2 V - \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2} = 0. \tag{1}$$

We shall now briefly examine the simplest solution of this equation.

1.3.1 Plane waves

Let $\mathbf{r}(x, y, z)$ be a position vector of a point P in space and $\mathbf{s}(s_x, s_y, s_z)$ a unit vector in a fixed direction. Any solution of (1) of the form

$$V = V(\mathbf{r}\cdot\mathbf{s}, t) \tag{2}$$

is said to represent a *plane wave*, since at each instant of time V is constant over each of the planes

$$\mathbf{r}\cdot\mathbf{s} = \text{constant}$$

which are perpendicular to the unit vector $\mathbf{s}$.

It will be convenient to choose a new set of Cartesian axes $O\xi$, $O\eta$, $O\zeta$ with $O\zeta$ in the direction of $\mathbf{s}$. Then (see Fig. 1.4)

$$\mathbf{r}\cdot\mathbf{s} = \zeta, \tag{3}$$

and one has

$$\frac{\partial}{\partial x} = s_x\frac{\partial}{\partial \zeta}, \qquad \frac{\partial}{\partial y} = s_y\frac{\partial}{\partial \zeta}, \qquad \frac{\partial}{\partial z} = s_z\frac{\partial}{\partial \zeta}.$$

From these relations one easily finds that

$$\nabla^2 V = \frac{\partial^2 V}{\partial \zeta^2}, \tag{4}$$

so that (1) becomes

$$\frac{\partial^2 V}{\partial \zeta^2} - \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2} = 0. \tag{5}$$

If we set

$$\zeta - vt = p, \qquad \zeta + vt = q, \tag{6}$$

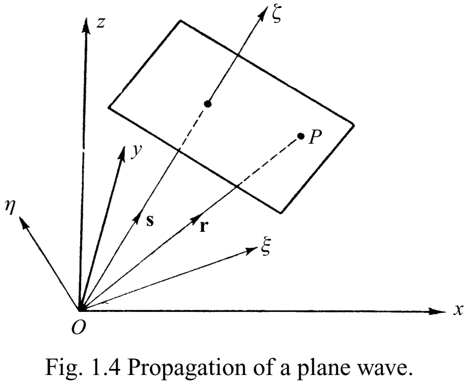

Fig. 1.4 Propagation of a plane wave.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

(5) takes the form

$$\frac{\partial^2 V}{\partial p \partial q} = 0. \tag{7}$$

The general solution of this equation is

$$\begin{aligned} V &= V_1(p) + V_2(q) \\ &= V_1(\mathbf{r} \cdot \mathbf{s} - vt) + V_2(\mathbf{r} \cdot \mathbf{s} + vt), \end{aligned} \tag{8}$$

where V_1 and V_2 are arbitrary functions.

We see that the argument of V_1 is unchanged when (ζ, t) is replaced by $(\zeta + v\tau, t + \tau)$, where τ is arbitrary. Hence V_1 represents a disturbance which is propagated with velocity v in the positive ζ direction. Similarly $V_2(\zeta + vt)$ represents a disturbance which is propagated with velocity v in the negative ζ direction.

1.3.2 Spherical waves

Next we consider solutions representing spherical waves, i.e. solutions of the form

$$V = V(r, t), \tag{9}$$

where $r = |r| = \sqrt{x^2 + y^2 + z^2}$.

Using the relations $\partial/\partial x = (\partial r/\partial x)(\partial/\partial r) = (x/r)(\partial/\partial r)$, etc., one finds after a straightforward calculation that

$$\nabla^2 V = \frac{1}{r}\frac{\partial^2}{\partial r^2}(rV), \tag{10}$$

so that the wave equation (1) now becomes

$$\frac{\partial^2}{\partial r^2}(rV) - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}(rV) = 0. \tag{11}$$

Now this equation is identical with (5), if ζ is replaced in the latter by r and V by rV. Hence the solution of (11) can immediately be written down from (8):

$$V = \frac{V_1(r - vt)}{r} + \frac{V_2(r + vt)}{r}, \tag{12}$$

V_1 and V_2 being again arbitrary functions. The first term on the right-hand side of (12) represents a spherical wave diverging from the origin, the second a spherical wave converging towards the origin, the velocity of propagation being v in both cases.

1.3.3 Harmonic waves. The phase velocity

At a point $\mathbf{r}_0$ in space the wave disturbance is a function of time only:

$$V(\mathbf{r}_0, t) = F(t). \tag{13}$$

As will be evident from our earlier remarks about colour, the case when F is periodic is of particular interest. Accordingly we consider the case when F has the form

$$F(t) = a\cos(\omega t + \delta). \tag{14}$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Here a (>0) is called the *amplitude*, and the argument $\omega t + \delta$ of the cosine term is called the *phase*. The quantity

$$\nu = \frac{\omega}{2\pi} = \frac{1}{T} \tag{15}$$

is called the *frequency* and represents the number of vibrations per second. ω is called the *angular* frequency and gives the number of vibrations in 2π seconds. Since F remains unchanged when t is replaced by $t + T$, T is the *period* of the vibrations. Wave functions (i.e. solutions of the wave equation) of the form (14) are said to be *harmonic* with respect to time.

Let us first consider a wave function which represents a *harmonic plane wave* propagated in the direction specified by a unit vector $\mathbf{s}$. According to §1.3.1 it is obtained on replacing t by $t - \mathbf{r}\cdot\mathbf{s}/v$ in (14):

$$V(\mathbf{r}, t) = a\cos\left[\omega\left(t - \frac{\mathbf{r}\cdot\mathbf{s}}{v}\right) + \delta\right]. \tag{16}$$

Eq. (16) remains unchanged when $\mathbf{r}\cdot\mathbf{s}$ is replaced by $\mathbf{r}\cdot\mathbf{s} + \lambda$, where

$$\lambda = v\frac{2\pi}{\omega} = vT. \tag{17}$$

The length λ is called the *wavelength*. It is also useful to define a *reduced wavelength* λ_0 as

$$\lambda_0 = cT = n\lambda; \tag{18}$$

this is the wavelength which corresponds to a harmonic wave of the same frequency propagated *in vacuo*. In spectroscopy one uses also the concept of a *wave number*[*] κ, which is defined as the number of wavelengths *in vacuo*, per unit of length (cm):

$$\kappa = \frac{1}{\lambda_0} = \frac{\nu}{c}. \tag{19}$$

It is also convenient to define vectors $\mathbf{k}_0$ and $\mathbf{k}$ in the direction $\mathbf{s}$ of propagation, whose lengths are respectively

$$k_0 = 2\pi\kappa = \frac{2\pi}{\lambda_0} = \frac{\omega}{c}, \tag{20}$$

and

$$k = nk_0 = \frac{2\pi}{\lambda} = \frac{n\omega}{c} = \frac{\omega}{v}. \tag{21}$$

The vector $\mathbf{k} = k\mathbf{s}$ is called the *wave vector* or the *propagation vector* in the medium, $\mathbf{k}_0 = k_0\mathbf{s}$ being the corresponding vector in the vacuum.

Instead of the constant δ one also uses the concept of *path length* l, which is the distance through which a wave-front recedes when the phase increases by δ:

$$l = \frac{v}{\omega}\delta = \frac{\lambda}{2\pi}\delta = \frac{\lambda_0}{2\pi n}\delta. \tag{22}$$

[*] We shall refer to κ as the ‘spectroscopic wave number’ and reserve the term ‘wave number’ for k_0 or k, defined by (20) and (21), as customary in optics.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Let us now consider time-harmonic waves of more complicated form. A general time-harmonic, real, scalar wave of frequency ω may be defined as a real solution of the wave equation, of the form

$$V(\mathbf{r},\ t) = a(\mathbf{r})\cos[\omega t - g(\mathbf{r})], \tag{23}$$

a (> 0) and g being real scalar functions of positions. The surfaces

$$g(\mathbf{r}) = \text{constant} \tag{24}$$

are called *cophasal surfaces* or *wave surfaces*. In contrast with the previous case, the surfaces of constant amplitude of the wave (23) do not, in general, coincide with the surfaces of constant phase. Such a wave is said to be *inhomogeneous*.

Calculations with harmonic waves are simplified by the use of exponential instead of trigonometric functions. Eq. (23) may be written as

$$V(\mathbf{r},\ t) = \mathcal{R}\{U(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}\}, \tag{25}$$

where

$$U(\mathbf{r}) = a(\mathbf{r})\mathrm{e}^{\mathrm{i}\,g(\mathbf{r})}, \tag{26}$$

and $\mathcal{R}$ denotes the real part. On substitution from (26) into the wave equation (1), one finds that U must satisfy the equation

$$\nabla^2 U + n^2 {k_0}^2 U = 0. \tag{27}$$

U is called the *complex amplitude*[*] of the wave. In particular, for a plane wave one has

$$g(\mathbf{r}) = \omega\left(\frac{\mathbf{r}\cdot\mathbf{s}}{v}\right) - \delta = k(\mathbf{r}\cdot\mathbf{s}) - \delta = \mathbf{k}\cdot\mathbf{r} - \delta. \tag{28}$$

If the operations on V are linear, one may drop the symbol $\mathcal{R}$ in (25) and operate directly with the complex function, the real part of the final expression being then understood to represent the physical quantity in question. However, when dealing with expressions which involve nonlinear operations such as squaring, etc. (e.g. in calculations of the electric or magnetic energy densities), one must in general take the real parts first and operate with these alone.[†]

Unlike a plane harmonic wave, the more general wave (25) is not periodic with respect to space. The phase $\omega t - g(\mathbf{r})$ is, however, seen to be the same for $(\mathbf{r},\ t)$ and $(\mathbf{r} + \mathrm{d}\mathbf{r},\ t + \mathrm{d}t)$, provided that

$$\omega\,\mathrm{d}t - (\operatorname{grad} g)\cdot\mathrm{d}\mathbf{r} = 0. \tag{29}$$

If we denote by $\mathbf{q}$ the unit vector in the direction of $\mathrm{d}\mathbf{r}$, and write $\mathrm{d}\mathbf{r} = \mathbf{q}\,\mathrm{d}s$, then (29) gives

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\omega}{\mathbf{q}\cdot\operatorname{grad} g}. \tag{30}$$

This expression will be numerically smallest when $\mathbf{q}$ is the normal to the cophasal surface, i.e. when $\mathbf{q} = \operatorname{grad} g/|\operatorname{grad} g|$, the value then being

[*] In the case of a plane wave, one often separates the constant factor $\mathrm{e}^{-\mathrm{i}\delta}$ and implies by complex amplitude only the variable part $a\mathrm{e}^{\mathrm{i}k\cdot r}$.

[†] This is not necessary when only a time average of a quadratic expression is required [see §1.4, (54)–(56)].

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$v^{(p)}(\mathbf{r}) = \frac{\omega}{|\text{grad } g|}. \tag{31}$$

$v^{(p)}(\mathbf{r})$ is called the *phase velocity* and is the speed with which each of the cophasal surfaces advances. For a plane electromagnetic wave one has from (28) that grad $g = \mathbf{k}$, and therefore

$$v^{(p)} = \frac{\omega}{\mathbf{k}} = \frac{c}{\sqrt{\varepsilon\mu}},$$

because of (21). For waves of more complicated form, the phase velocity $v^{(p)}$ will in general differ from $c/\sqrt{\varepsilon\mu}$ and will vary from point to point even in a homogeneous medium. However, it will be seen later (§3.1.2) that, when the frequency is sufficiently large, the phase velocity is *approximately* equal to $c/\sqrt{\varepsilon\mu}$, even for waves whose cophasal surfaces are not plane.

It must be noted that the expression for $\mathrm{d}s/\mathrm{d}t$ given by (30) is not the resolute of the phase velocity in the $\mathbf{q}$ direction, i.e. the phase velocity does not behave as a vector. On the other hand its reciprocal, i.e. the quantity

$$\frac{\mathrm{d}t}{\mathrm{d}s} = \frac{\mathbf{q} \cdot \text{grad } g}{\omega}, \tag{32}$$

is seen to be the component of the vector (grad $g)/\omega$ in the $\mathbf{q}$ direction. The vector (grad $g)/\omega$ is sometimes called *phase slowness*.

The phase velocity may in certain cases be greater than c. For plane waves this will be so when $n = \sqrt{\varepsilon\mu}$ is smaller than unity, as in the case of dispersing media in regions of the so-called anomalous dispersion* (see §2.3.4). Now according to the theory of relativity, signals can never travel faster than c. This implies that the phase velocity cannot correspond to a velocity with which a signal is propagated. It is, in fact, easy to see that the phase velocity cannot be determined experimentally and must therefore be considered to be void of any direct physical significance. For in order to measure this velocity, it would be necessary to affix a mark to the infinitely extended smooth wave and to measure the velocity of the mark. This would, however, mean the replacement of the infinite harmonic wave train by another function of space and time.

1.3.4 Wave packets. The group velocity

The monochromatic waves considered in the preceding section are idealizations never strictly realized in practice. It follows from Fourier's theorem that any wave $V(\mathbf{r}, t)$ (provided it satisfies certain very general conditions) may be regarded as a superposition of monochromatic waves of different frequencies:

$$V(\mathbf{r}, t) = \int_0^\infty a_\omega(\mathbf{r}) \cos[\omega t - g_\omega(\mathbf{r})] \, \mathrm{d}\omega. \tag{33}$$

* The problem of propagation of electromagnetic signals in dispersive media has been investigated in classic papers by A. Sommerfeld, *Ann. d. Physik*, **44** (1914), 177 and by L. Brillouin, *ibid.*, **44** (1914), 203. English translations of these papers are included in L. Brillouin, *Wave Propagation and Group Velocity* (New York, Academic Press, 1960), pp. 17, 43. A systematic treatment of propagation of transient electromagnetic fields through dielectric media which exhibit both dispersion and absorption is given in K. E. Oughstun and G. C. Sherman, *Electromagnetic Pulse Propagation in Causal Dielectrics* (Berlin and New York, Springer, 1994).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

It will again be convenient to use a complex representation, in which V is regarded as the real part of an associated complex wave:[*]

$$V(\mathbf{r},\, t) = \mathcal{R}\int_0^\infty a_\omega(\mathbf{r})\mathrm{e}^{-\mathrm{i}[\omega t - g_\omega(\mathbf{r})]}\,\mathrm{d}\omega. \tag{33a}$$

A wave may be said to be 'almost monochromatic,' if the Fourier amplitudes a_ω differ appreciably from zero only within a narrow range

$$\overline{\omega} - \tfrac{1}{2}\Delta\omega \leqslant \omega \leqslant \overline{\omega} + \tfrac{1}{2}\Delta\omega \qquad (\Delta\omega/\overline{\omega} \ll 1)$$

around a mean frequency $\overline{\omega}$. In such a case one usually speaks of a *wave group* or a *wave packet*.[†]

To illustrate some of the main properties of a wave group, consider first a wave formed by the superposition of two plane monochromatic waves of the same amplitudes and slightly different frequencies and wave numbers, propagated in the direction of the z-axis:

$$V(z,\, t) = a\mathrm{e}^{-\mathrm{i}(\omega t - kz)} + a\mathrm{e}^{-\mathrm{i}[(\omega+\delta\omega)t-(k+\delta k)z]}. \tag{34}$$

The symbol $\mathcal{R}$ is omitted here in accordance with the convention explained earlier. Eq. (34) may be written in the form

$$\begin{aligned} V(z,\, t) &= a[\mathrm{e}^{\frac{1}{2}\mathrm{i}(t\delta\omega + z\delta k)} + \mathrm{e}^{-\frac{1}{2}\mathrm{i}(t\delta\omega - z\delta k)}]\mathrm{e}^{-\mathrm{i}(\overline{\omega}t - \overline{k}z)} \\ &= 2a\cos[\tfrac{1}{2}(t\delta\omega - z\delta k)]\mathrm{e}^{-\mathrm{i}(\overline{\omega}t-\overline{k}z)}, \end{aligned} \tag{35}$$

where

$$\overline{\omega} = \omega + \tfrac{1}{2}\delta\omega, \qquad \overline{k} = k + \tfrac{1}{2}\delta k \tag{36}$$

are the mean frequency and the mean wave number respectively. Eq. (35) may be interpreted as representing a plane wave of frequency $\overline{\omega}$ and wavelength $2\pi/\overline{k}$ propagated in the z direction. The amplitude of this wave is, however, not constant, but varies with time and position, between the values $2a$ and 0 (Fig. 1.5), giving rise to the well-known phenomenon of beats. The successive maxima of the amplitude function are at intervals

$$\delta t = \frac{4\pi}{\delta\omega}\ (\text{with } z \text{ fixed}) \quad \text{or} \quad \delta z = \frac{4\pi}{\delta k}\ (\text{with } t \text{ fixed}) \tag{37}$$

from each other, whilst the maxima of the phase function are at intervals

$$\delta t = \frac{2\pi}{\overline{\omega}}\ (\text{with } z \text{ fixed}) \quad \text{or} \quad \delta z = \frac{2\pi}{\overline{k}}\ (\text{with } t \text{ fixed}). \tag{38}$$

Hence, since $\delta\omega/\overline{\omega}$ and $\delta k/\overline{k}$ are assumed to be small compared with unity, the amplitude will vary slowly in comparison with the other term.

From (35) it follows that the planes of constant amplitude and, in particular, the maxima of the amplitude are propagated with the velocity

[*] For a fuller discussion of the complex representation of real polychromatic waves see §10.2.

[†] Strictly speaking, in order that V should exhibit properties commonly attributed to a wave group, one should also assume that over the effective frequency range the phase function g_ω can be approximated by a linear function of ω.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

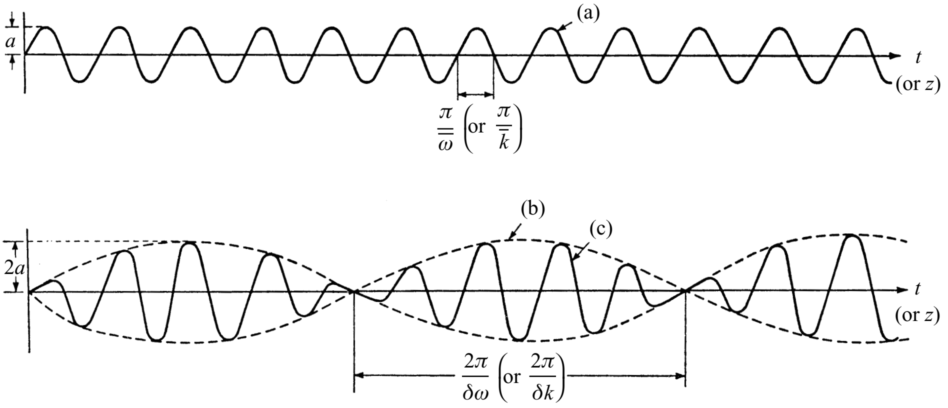
(a)
a
t
(or z)
$\frac{\pi}{\overline{\omega}}\left(\text{or } \frac{\pi}{\overline{k}}\right)$
(b)
(c)
2a
t
(or z)
$\frac{2\pi}{\delta\omega}\left(\text{or } \frac{2\pi}{\delta k}\right)$

Fig. 1.5 A simple wave group: (a) the wave $a\cos(\overline{\omega}t - \overline{k}z)$; (b) the wave $2a\cos[\frac{1}{2}(t\delta\omega - z\delta k)]$; (c) the wave group $2a\cos[\frac{1}{2}(t\delta\omega - z\delta k)]\cos(\overline{\omega}t - \overline{k}z)$. The ordinate represents one of the two independent variables (t or z) whilst the other is kept constant.

$$v^{(g)} = \frac{\delta\omega}{\delta k}, \tag{39}$$

whilst the planes of constant phase are propagated with the velocity

$$v^{(p)} = \frac{\overline{\omega}}{\overline{k}}. \tag{40}$$

$v^{(g)}$ is called the *group velocity* of the wave. Since V obeys the wave equation, the frequency ω and the wave number k are related; in a medium of refractive index n, one has (see (21))

$$k = n(\omega)\frac{\omega}{c}, \tag{41}$$

where n is the refractive index function. Eq. (41) expresses the *dispersion* of the wave. In a nondispersive medium, n is independent of ω; the phase velocity $v^{(p)}$ and the group velocity $v^{(g)}$ are then both equal to c/n. In a dispersive medium, however, the two velocities will, in general, be different.

Since $\delta\omega$ is assumed to be small, $\delta\omega/\delta k$ may be replaced by the derivative $\mathrm{d}\omega/\mathrm{d}k$, so that the expression for the group velocity may be written as

$$v^{(g)} = \frac{\mathrm{d}\omega}{\mathrm{d}k}. \tag{42}$$

This relation is, in fact, valid under more general conditions, as we shall now show. Consider a one-dimensional wave group

$$V(z, t) = \int_{(\Delta\omega)} a_\omega \mathrm{e}^{-\mathrm{i}(\omega t - kz)}\,\mathrm{d}\omega, \tag{43}$$

where $\Delta\omega$ denotes the small interval around a mean frequency $\overline{\omega}$ ($\Delta\omega/\overline{\omega} \ll 1$) for which a_ω differs appreciably from zero. Let $\overline{k} = n(\overline{\omega})\overline{\omega}/c$ be the corresponding wave number. Then (43) may be expressed in the form

$$V(z, t) = A(z, t)\mathrm{e}^{-\mathrm{i}(\overline{\omega}t - \overline{k}z)}, \tag{44}$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

where

$$A(z,\ t) = \int_{(\Delta\omega)} a_\omega \mathrm{e}^{-\mathrm{i}[(\omega-\overline{\omega})t-(k-\overline{k})z]}\,\mathrm{d}\omega \sim \int_{(\Delta\omega)} a_\omega \mathrm{e}^{-\mathrm{i}\left[(\omega-\overline{\omega})\left\{t-\left(\frac{\mathrm{d}k}{\mathrm{d}\omega}\right)_{\overline{\omega}}z\right\}\right]}\,\mathrm{d}\omega, \tag{45}$$

if $\Delta\omega$ is sufficiently small. V may again be interpreted as a plane wave with variable amplitude, of frequency $\overline{\omega}$ and wave number $\overline{k}$, propagated in the z direction. The amplitude $A(z,\ t)$ is represented as a superposition of harmonic waves of frequencies $\omega - \overline{\omega}$. Since $\Delta\omega/\overline{\omega}$ is assumed to be small compared with unity, A will vary slowly in comparison with the other term. In general A is complex, so that there is a contribution (arg A) to the phase $\overline{\omega}t - \overline{k}z$. The surfaces

$$t = \left(\frac{\mathrm{d}k}{\mathrm{d}\omega}\right)_{\overline{\omega}} z \tag{46}$$

are seen to play a special role: on each such surface $A(z,\ t)$ is constant. Hence the velocity of advance of a definite value of A and also of the maximum of $|A|$ is given by the *group velocity*

$$v^{(g)} = \left(\frac{\mathrm{d}\omega}{\mathrm{d}k}\right)_{\overline{k}} \tag{47}$$

as before.

The following relations are seen to hold between the group velocity and the phase velocity:

$$v^{(g)} = \frac{\mathrm{d}}{\mathrm{d}k}(v^{(p)}k) = v^{(p)} + k\frac{\mathrm{d}v^{(p)}}{\mathrm{d}k} = v^{(p)} - \lambda\frac{\mathrm{d}v^{(p)}}{\mathrm{d}\lambda}, \tag{48}$$

all the quantities here referring to mean frequency $\overline{\omega}$.

Finally, let us consider the general three-dimensional wave group

$$V(\mathbf{r},\ t) = \mathcal{R}\int_{(\Delta\omega)} a_\omega(\mathbf{r})\mathrm{e}^{-\mathrm{i}[\omega t - g_\omega(\mathbf{r})]}\,\mathrm{d}\omega. \tag{49}$$

By analogy with (43), we separate a term corresponding to the mean frequency $\overline{\omega}$, and write

$$V(\mathbf{r},\ t) = A(\mathbf{r},\ t)\mathrm{e}^{-\mathrm{i}[\overline{\omega}t - g_{\overline{\omega}}(\mathbf{r})]}, \tag{50}$$

where

$$A(\mathbf{r},\ t) = \int_{(\Delta\omega)} a_\omega(\mathbf{r})\mathrm{e}^{-\mathrm{i}\{(\omega-\overline{\omega})t-[g_\omega(\mathbf{r})-g_{\overline{\omega}}(\mathbf{r})]\}}\,\mathrm{d}\omega \sim \int_{(\Delta\omega)} a_\omega(\mathbf{r})\mathrm{e}^{-\mathrm{i}\left\{(\omega-\overline{\omega})\left[t-\left(\frac{\partial g(\mathbf{r})}{\partial\omega}\right)_{\overline{\omega}}\right]\right\}}\,\mathrm{d}\omega, \tag{51}$$

if $\Delta\omega$ is sufficiently small. Eq. (50) represents a wave of frequency $\overline{\omega}$ whose (generally complex) amplitude $A(\mathbf{r},\ t)$ varies both in space and time, this variation again being slow in comparison with the other term. By analogy with (46) the surface

$$t = \left[\frac{\partial g(\mathbf{r})}{\partial\omega}\right]_{\overline{\omega}} \tag{52}$$

may be expected to play a special role. However, the amplitude function A is now not necessarily constant over each such surface, since the Fourier amplitudes a_ω are now

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

functions not only of the frequency but also of the position. The significance of (52) is seen by considering the absolute amplitude $M = |A|$. We have

$$M^2(\mathbf{r},\ t) = A(\mathbf{r},\ t)A^\star(\mathbf{r},\ t) = \int_{(\Delta\omega)}\int_{(\Delta\omega)} a_\omega(\mathbf{r})a_{\omega'}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\left\{(\omega-\omega')\left[t-\left(\frac{\partial g(\mathbf{r})}{\partial\omega}\right)_{\varpi}\right]\right\}}\,\mathrm{d}\omega\,\mathrm{d}\omega'. \tag{53}$$

Obviously the imaginary part of the double integral vanishes since M^2 is real. (Formally this may be verified by interchanging the independent variables ω and ω' and noting that the imaginary part of the integrand then changes sign.) Hence

$$M^2(\mathbf{r},\ t) = \int_{(\Delta\omega)}\int_{(\Delta\omega)} a_\omega(\mathbf{r})a_{\omega'}(\mathbf{r})\cos\left\{(\omega-\omega')\left[t-\left(\frac{\partial g(\mathbf{r})}{\partial\omega}\right)_{\varpi}\right]\right\}\mathrm{d}\omega\,\mathrm{d}\omega'. \tag{54}$$

Fixing our attention on any particular point $\mathbf{r} = \mathbf{r}_0$ and remembering that a_ω is either positive or zero, we see that $M^2(\mathbf{r}_0,\ t)$ attains its maximum when the argument of the cosine term is zero, i.e. when $t = (\partial g(\mathbf{r}_0)/\partial\omega)_{\varpi}$. Thus (52) represents the surfaces on which the absolute amplitude attains its maximum at time t in the sense just explained. It is therefore appropriate to define the group velocity of the general three-dimensional wave group as the velocity with which these surfaces advance. Considering a small displacement $\delta\mathbf{r} = \mathbf{q}\delta s$, where $\mathbf{q}$ is a unit vector in the direction normal to the surface, we have from (52) that the corresponding change δt is given by

$$\delta t = \delta s\left|\operatorname{grad}\left(\frac{\partial g(\mathbf{r})}{\partial\omega}\right)_{\varpi}\right|, \tag{55}$$

so that the group velocity of the general three-dimensional group is given by

$$v^{(g)} = \frac{1}{\left|\operatorname{grad}\left(\frac{\partial g}{\partial\omega}\right)_{\varpi}\right|}. \tag{56}$$

This expression should be compared with (31)

$$v^{(p)} = \frac{1}{\left|\operatorname{grad}\frac{g}{\omega}\right|} \tag{57}$$

for the phase velocity of a general harmonic wave. In the special case of a group of plane waves propagated in the z direction, $g_\omega = kz$, and (56) reduces to (47).

It is evident from the preceding discussion that the effective frequency range $\Delta\omega$ is an important parameter relating to a wave group; it is this quantity which substantially determines the rate of variation of the amplitude and the phase. If the medium is not strongly dispersive, a wave group will travel a considerable distance without appreciable variation. In such circumstances, the group velocity, which may be considered as the velocity of the propagation of the group as a whole, will also represent the velocity at which the energy is propagated.* This, however, is not true in general. In particular, in regions of anomalous dispersion (see §2.3.4) the group velocity may exceed the

*See, for example, F. Borgnis, *Z. Phys.*, **117** (1941), 642; L. J. F. Broer, *Appl. Sci. Res.*, **A2** (1951), 329.

speed of light in vacuum or become negative,[*] but there is no conflict with the special theory of relativity.[†]

1.4 Vector waves

1.4.1 The general electromagnetic plane wave

The simplest electromagnetic field is that of a plane wave; then each Cartesian component of the field vectors and consequently **E** and **H** are, according to §1.3.1, functions of the variable $u = \mathbf{r}\cdot\mathbf{s} - vt$ only:

$$\mathbf{E} = \mathbf{E}(\mathbf{r}\cdot\mathbf{s} - vt), \qquad \mathbf{H} = \mathbf{H}(\mathbf{r}\cdot\mathbf{s} - vt), \tag{1}$$

s denoting as before a unit vector in the direction of propagation.

Denoting by a dot differentiation with respect to t, and by a prime differentiation with respect to the variable u, we have

$$\left.\begin{aligned}\dot{\mathbf{E}} &= -v\mathbf{E}',\\ (\text{curl}\,\mathbf{E})_x &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = E'_z s_y - E'_y s_z = (\mathbf{s}\times\mathbf{E}')_x.\end{aligned}\right\} \tag{2}$$

Substituting these expressions into Maxwell's equations §1.1 (1), §1.1 (2) with $\mathbf{j} = 0$, and using the material equations §1.1 (10), §1.1 (11) we obtain

$$\left.\begin{aligned}\mathbf{s}\times\mathbf{H}' + \frac{\varepsilon v}{c}\mathbf{E}' &= 0,\\ \mathbf{s}\times\mathbf{E}' - \frac{\mu v}{c}\mathbf{H}' &= 0.\end{aligned}\right\} \tag{3}$$

If we set the additive constants of integration equal to zero (i.e. neglect a field constant in space) and set, as before, $v/c = 1/\sqrt{\varepsilon\mu}$, (3) gives, on integration,

$$\left.\begin{aligned}\mathbf{E} &= -\sqrt{\frac{\mu}{\varepsilon}}\,\mathbf{s}\times\mathbf{H},\\ \mathbf{H} &= \sqrt{\frac{\varepsilon}{\mu}}\,\mathbf{s}\times\mathbf{E}.\end{aligned}\right\} \tag{4}$$

Scalar multiplication with **s** gives

$$\mathbf{E}\cdot\mathbf{s} = \mathbf{H}\cdot\mathbf{s} = 0. \tag{5}$$

This relation expresses the 'transversality' of the field, i.e. it shows that the electric and magnetic field vectors lie in planes normal to the direction of propagation. From (4) and (5) it is seen that **E**, **H** and **s** form a right-handed orthogonal triad of vectors. We also have from (4) that

$$\sqrt{\mu}H = \sqrt{\varepsilon}E, \tag{6}$$

where $E = |\mathbf{E}|$, $H = |\mathbf{H}|$.

Let us now consider the amount of energy which crosses, in unit time, an element of area perpendicular to the direction of propagation. Imagine a cylinder whose axis is parallel to **s** and whose cross-sectional area is unity. The amount of energy which

[*] L. J. Wang, A. Kuzmich and A. Dogariu, *Nature*, **406** (2000), 277.

[†] R. Y. Chiao and A. M. Steinberg in *Progress in Optics*, Vol. XXXVII, ed. E. Wolf (Amsterdam, Elsevier, 1997), p. 345 *et seq*.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

crosses the base of the cylinder in unit time is then equal to the energy which was contained in the portion of the cylinder of length v. Hence the energy flux is equal to vw, where w is the energy density. According to (6) and §1.1 (31), the energy density is given by

$$w = \frac{\varepsilon}{4\pi}E^2 = \frac{\mu}{4\pi}H^2. \tag{7}$$

The Poynting vector, on the other hand, is according to §1.1 (38) given by

$$\mathbf{S} = \frac{c}{4\pi}EH\mathbf{s} = \frac{c}{4\pi}\sqrt{\frac{\varepsilon}{\mu}}E^2\mathbf{s} = \frac{c}{4\pi}\sqrt{\frac{\mu}{\varepsilon}}H^2\mathbf{s}. \tag{8}$$

Comparison of (7) and (8) shows that

$$\mathbf{S} = \frac{c}{\sqrt{\varepsilon\mu}}w\mathbf{s} = vw\mathbf{s}. \tag{9}$$

We see that, in agreement with §1.1.4, the Poynting vector represents the flow of energy both with regard to its magnitude and direction of propagation.

1.4.2 The harmonic electromagnetic plane wave

Of particular interest is the case when the plane wave is time-harmonic, i.e. when each Cartesian component of **E** and **H** is of the form

$$a\cos(\tau + \delta) = \mathcal{R}\{a\mathrm{e}^{-\mathrm{i}(\tau+\delta)}\} \qquad (a > 0). \tag{10}$$

Here τ denotes the variable part of the phase factor, i.e.

$$\tau = \omega\left(t - \frac{\mathbf{r}\cdot\mathbf{s}}{v}\right) = \omega t - \mathbf{k}\cdot\mathbf{r}. \tag{11}$$

We choose the z-axis in the **s** direction. Then, since according to (5) the field is transversal, only the x- and y-components of **E** and **H** are different from zero. We shall now consider the nature of the curve which the end point of the electric vector describes at a typical point in space; this curve is the locus of the points whose coordinates (E_x, E_y) are

$$\left.\begin{aligned} E_x &= a_1\cos(\tau + \delta_1), \\ E_y &= a_2\cos(\tau + \delta_2). \end{aligned}\right\} \tag{12}$$

(a) Elliptic polarization

In order to eliminate τ between the first two equations of (12), we re-write them in the form

$$\left.\begin{aligned} \frac{E_x}{a_1} &= \cos\tau\cos\delta_1 - \sin\tau\sin\delta_1, \\ \frac{E_y}{a_2} &= \cos\tau\cos\delta_2 - \sin\tau\sin\delta_2. \end{aligned}\right\} \tag{13}$$

Hence

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\left.\begin{aligned}\frac{E_x}{a_1}\sin\delta_2-\frac{E_y}{a_2}\sin\delta_1&=\cos\tau\sin(\delta_2-\delta_1),\\\frac{E_x}{a_1}\cos\delta_2-\frac{E_y}{a_2}\cos\delta_1&=\sin\tau\sin(\delta_2-\delta_1).\end{aligned}\right\}\tag{14}$$

Squaring and adding gives

$$\left(\frac{E_x}{a_1}\right)^2+\left(\frac{E_y}{a_2}\right)^2-2\frac{E_x}{a_1}\frac{E_y}{a_2}\cos\delta=\sin^2\delta,\tag{15}$$

where

$$\delta=\delta_2-\delta_1.\tag{16}$$

Eq. (15) is the equation of a conic. It is an ellipse, since the associated determinant is not negative:

$$\begin{vmatrix}\dfrac{1}{a_1^2} & -\dfrac{\cos\delta}{a_1a_2}\\ -\dfrac{\cos\delta}{a_1a_2} & \dfrac{1}{a_2^2}\end{vmatrix}=\frac{1}{a_1^2a_2^2}(1-\cos^2\delta)=\frac{\sin^2\delta}{a_1^2a_2^2}\geqslant 0.$$

The ellipse is inscribed in a rectangle whose sides are parallel to the coordinate axes and whose lengths are $2a_1$ and $2a_2$ (Fig. 1.6). The ellipse touches the sides at the point $(\pm a_1, \pm a_2\cos\delta)$ and $(\pm a_1\cos\delta, \pm a_2)$.

The wave (12) is then said to be *elliptically polarized*. It is easily seen that the wave associated with the magnetic vector is also elliptically polarized. For by (5) and (12),

$$\left.\begin{aligned}H_x&=-\sqrt{\frac{\varepsilon}{\mu}}E_y=-\sqrt{\frac{\varepsilon}{\mu}}a_2\cos(\tau+\delta_2),\\H_y&=\sqrt{\frac{\varepsilon}{\mu}}E_x=\sqrt{\frac{\varepsilon}{\mu}}a_1\cos(\tau+\delta_1).\end{aligned}\right\}\tag{17}$$

The end point of the magnetic vector therefore describes an ellipse which is inscribed into a rectangle whose sides are parallel to the x and y directions and whose lengths are $2\sqrt{\varepsilon/\mu}a_2$, $2\sqrt{\varepsilon/\mu}a_1$.

In general the axes of the ellipse are not in the Ox and Oy directions. Let $O\xi$, $O\eta$ be a new set of axes along the axes of the ellipse and let ψ $(0\leqslant\psi<\pi)$ be the angle between Ox and the direction $O\xi$ of the major axis (Fig. 1.6). Then the components E_ξ and E_η are related to E_x and E_y by

$$\left.\begin{aligned}E_\xi&=E_x\cos\psi+E_y\sin\psi,\\E_\eta&=-E_x\sin\psi+E_y\cos\psi.\end{aligned}\right\}\tag{18}$$

If $2a$ and $2b$ $(a\geqslant b)$ are the lengths of the axes of the ellipse, the equation of the ellipse referred to $O\xi$, $O\eta$ is:

$$\left.\begin{aligned}E_\xi&=a\cos(\tau+\delta_0),\\E_\eta&=\pm b\sin(\tau+\delta_0).\end{aligned}\right\}\tag{19}$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

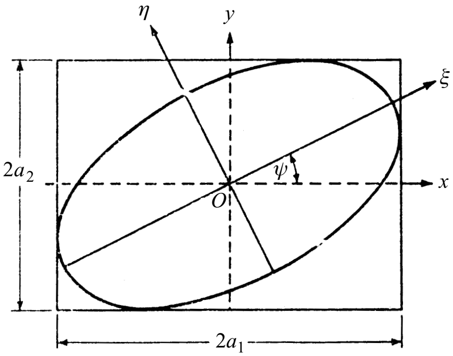

Fig. 1.6 Elliptically polarized wave. The vibrational ellipse for the electric vector.

The two signs distinguish the two possible senses in which the end point of the electric vector may describe the ellipse.

To determine a and b we compare (18) and (19) and use (13):

$$\begin{aligned} a(\cos\tau\cos\delta_0 - \sin\tau\sin\delta_0) &= a_1(\cos\tau\cos\delta_1 - \sin\tau\sin\delta_1)\cos\psi \\ &\quad + a_2(\cos\tau\cos\delta_2 - \sin\tau\sin\delta_2)\sin\psi; \\ \pm b(\sin\tau\cos\delta_0 + \cos\tau\sin\delta_0) &= -a_1(\cos\tau\cos\delta_1 - \sin\tau\sin\delta_1)\sin\psi \\ &\quad + a_2(\cos\tau\cos\delta_2 - \sin\tau\sin\delta_2)\cos\psi. \end{aligned}$$

Next we equate the coefficients of $\cos\tau$ and $\sin\tau$:

$$a\cos\delta_0 = a_1\cos\delta_1\cos\psi + a_2\cos\delta_2\sin\psi, \tag{20a}$$

$$a\sin\delta_0 = a_1\sin\delta_1\cos\psi + a_2\sin\delta_2\sin\psi, \tag{20b}$$

$$\pm b\cos\delta_0 = a_1\sin\delta_1\sin\psi - a_2\sin\delta_2\cos\psi, \tag{21a}$$

$$\pm b\sin\delta_0 = -a_1\cos\delta_1\sin\psi + a_2\cos\delta_2\cos\psi. \tag{21b}$$

On squaring and adding (20a) and (20b) and using (16) we obtain

$$\left.\begin{aligned} a^2 &= a_1^2\cos^2\psi + a_2^2\sin^2\psi + 2a_1a_2\cos\psi\sin\psi\cos\delta, \\ &\text{and similarly from (21a) and (21b)} \\ b^2 &= a_1^2\sin^2\psi + a_2^2\cos^2\psi - 2a_1a_2\cos\psi\sin\psi\cos\delta. \end{aligned}\right\} \tag{22}$$

Hence

$$a^2 + b^2 = a_1^2 + a_2^2. \tag{23}$$

Next we multiply (20a) by (21a), (20b) by (21b) and add. This gives

$$\mp ab = a_1a_2\sin\delta. \tag{24}$$

Further on dividing (21a) by (20a) and (21b) by (20b) we obtain

$$\pm\frac{b}{a} = \frac{a_1\sin\delta_1\sin\psi - a_2\sin\delta_2\cos\psi}{a_1\cos\delta_1\cos\psi + a_2\cos\delta_2\sin\psi} = \frac{-a_1\cos\delta_1\sin\psi + a_2\cos\delta_2\cos\psi}{a_1\sin\delta_1\cos\psi + a_2\sin\delta_2\sin\psi},$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

and these relations give the following equation for ψ:

$$(a_1^2 - a_2^2)\sin 2\psi = 2a_1a_2\cos\delta\cos 2\psi.$$

It will be convenient to introduce an auxiliary angle α ($0 \leqslant \alpha \leqslant \pi/2$), such that

$$\frac{a_2}{a_1} = \tan\alpha. \tag{25}$$

The preceding equation then becomes

$$\tan 2\psi = \frac{2a_1a_2}{a_1^2 - a_2^2}\cos\delta = \frac{2\tan\alpha}{1 - \tan^2\alpha}\cos\delta,$$

i.e.

$$\tan 2\psi = (\tan 2\alpha)\cos\delta. \tag{26}$$

Now from (23) and (24) we also have

$$\mp\frac{2ab}{a^2 + b^2} = \frac{2a_1a_2}{a_1^2 + a_2^2}\sin\delta = (\sin 2\alpha)\sin\delta. \tag{27}$$

Let χ ($-\pi/4 \leqslant \chi \leqslant \pi/4$) be another auxiliary angle, such that

$$\mp\frac{b}{a} = \tan\chi. \tag{28}$$

The numerical value of $\tan\chi$ represents the ratio of the axes of the ellipse and the sign of χ distinguishes the two senses in which the ellipse may be described. Eq. (27) may be written in the form

$$\sin 2\chi = (\sin 2\alpha)\sin\delta. \tag{29}$$

It will be useful to summarize the results. If a_1, a_2 and the phase difference δ are given, referred to an arbitrary set of axes, and if α ($0 \leqslant \alpha \leqslant \pi/2$) denotes an angle such that

$$\tan\alpha = \frac{a_2}{a_1}, \tag{30}$$

then the principal semiaxes a and b of the ellipse and the angle ψ ($0 \leqslant \psi < \pi$) which the major axis makes with Ox are specified by the formulae

$$a^2 + b^2 = a_1^2 + a_2^2, \tag{31a}$$

$$\tan 2\psi = (\tan 2\alpha)\cos\delta, \tag{31b}$$

$$\sin 2\chi = (\sin 2\alpha)\sin\delta, \tag{31c}$$

where χ ($-\pi/4 < \chi \leqslant \pi/4$) is an auxiliary angle which specifies the shape and orientation of the vibrational ellipse:

$$\tan\chi = \mp b/a. \tag{32}$$

Conversely, if the lengths a and b of the axes and the orientation of the ellipse are known (i.e. a, b and ψ given) these formulae enable the determination of the amplitudes a_1, a_2 and the phase difference δ. In Chapter XV, instruments will be described by means of which these quantities may be directly determined.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Before discussing some important special cases, we must say a few words about the terminology. We distinguish two cases of polarization, according to the sense in which the end point of the electric vector describes the ellipse. It seems natural to call the polarization right-handed or left-handed according to whether the rotation of **E** and the direction of propagation form a right-handed or left-handed screw. But the traditional terminology is just the opposite — being based on the apparent behaviour of **E** when 'viewed' face on by the observer. We shall conform throughout this book to this customary usage. Thus we say that the polarization is *right-handed* when to an observer looking in the direction from which the light is coming, the end point of the electric vector would appear to describe the ellipse in the clockwise sense. If we consider the values of (12) for two time instants separated by a quarter of a period, we see that in this case $\sin\delta > 0$, or by (29), $0 < \chi \leqslant \pi/4$. For *left-handed* polarization the opposite is the case, i.e. to an observer looking in the direction from which the light is propagated, the electric vector would appear to describe the ellipse anti-clockwise; in this case $\sin\delta < 0$, so that $-\pi/4 \leqslant \chi < 0$.

For reasons connected with the historical development of optics, the direction of the magnetic vector is often called the *direction of polarization* and the plane containing the magnetic vector and the direction of propagation is known as the *plane of polarization*. This terminology is, however, not used by all writers; some define these quantities with respect to the electric rather than the magnetic vector. This lack of uniformity arises partly from the fact that there is no single physical entity which could be described without ambiguity as 'the light vector'. When particular attention is paid to the physical effect of the field vectors, there would actually be some grounds for regarding **E** as the light vector, for every action is a consequence of the motion of elementary charged particles (electrons, nuclei) set into motion by the electromagnetic field. The mechanical force **F** of the field on the particle is then given by Lorentz' law, §1.1 (34).

$$\mathbf{F} = e\left(\mathbf{E} + \frac{\mu}{c}\mathbf{v} \times \mathbf{H}\right),$$

e being the charge and **v** the velocity of the particle. Hence the electric vector is seen to act even when the particle is at rest. On the other hand, the magnetic vector plays a part only when the particle is in motion; however, since v/c is usually very small compared to unity this effect may often be neglected. Nevertheless the 'direction of polarization' and the 'plane of polarization' are usually associated with the magnetic vector. The reason for this nomenclature will become apparent in the next section when polarization on reflection is discussed.

To avoid confusion we shall, in accordance with more recent practice, not use the terms 'direction of polarization' and 'plane of polarization'; instead we shall speak of *direction of vibration* and *plane of vibration* to denote the direction of a field vector and the plane containing the field vector and the direction of propagation, the vector in question being specified in each case.

(b) Linear and circular polarization

Two special cases are of particular importance, namely when the polarization ellipse degenerates into a straight line or a circle.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

According to (12) the ellipse will reduce to a straight line when

$$\delta = \delta_2 - \delta_1 = m\pi \qquad (m = 0, \pm 1, \pm 2, \ldots). \tag{33}$$

Then

$$\frac{E_y}{E_x} = (-1)^m \frac{a_2}{a_1}, \tag{34}$$

and we say that **E** is *linearly polarized.** One of the coordinate axes, x say, may be chosen along this line. Then only one component (E_x) remains. Moreover, since the electric and magnetic vectors are orthogonal and lie in the plane perpendicular to the z direction, the component H_x then vanishes, so that **H** is linearly polarized in the y direction.

The other special case of importance is that of a *circularly polarized* wave, the ellipse then degenerating into a circle. Clearly a necessary condition for this is that the circumscribed rectangle shall become a square:

$$a_1 = a_2 = a. \tag{35}$$

Also, one of the **E** components must be zero when the other has an extreme value; this demands that

$$\delta = \delta_2 - \delta_1 = m\pi/2 \qquad (m = \pm 1, \pm 3, \pm 5 \ldots), \tag{36}$$

and (15) then reduces to the equation of the circle

$$E_x^2 + E_y^2 = a^2. \tag{37}$$

When the polarization is *right-handed* $\sin\delta > 0$, so that

$$\delta = \frac{\pi}{2} + 2m\pi \qquad (m = 0, \pm 1, \pm 2, \ldots), \tag{38}$$

$$\left.\begin{aligned} E_x &= a\cos(\tau + \delta_1), \\ E_y &= a\cos(\tau + \delta_1 + \pi/2) = -a\sin(\tau + \delta_1). \end{aligned}\right\} \tag{39}$$

For *left-handed polarization* $\sin\delta < 0$, so that

$$\delta = -\frac{\pi}{2} + 2m\pi \qquad (m = 0, \pm 1, \pm 2, \ldots), \tag{40}$$

$$\left.\begin{aligned} E_x &= a\cos(\tau + \delta_1), \\ E_y &= a\cos(\tau + \delta_1 - \pi/2) = a\sin(\tau + \delta_1). \end{aligned}\right\} \tag{41}$$

If, instead of the real representation, the complex one is used (i.e. if the exponential instead of the cosine function is written in (12)), then

$$\frac{E_y}{E_x} = \frac{a_2}{a_1}\mathrm{e}^{\mathrm{i}(\delta_1 - \delta_2)} = \frac{a_2}{a_1}\mathrm{e}^{-\mathrm{i}\delta}, \tag{42}$$

and one can immediately determine from the value of this ratio the nature of the polarization. One has for

* The less appropriate term *plane polarized* is also used.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

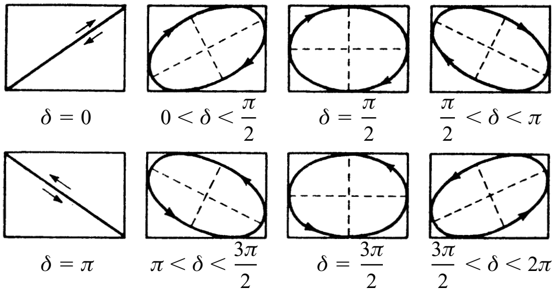
$\delta = 0$
$0 < \delta < \frac{\pi}{2}$
$\delta = \frac{\pi}{2}$
$\frac{\pi}{2} < \delta < \pi$
$\delta = \pi$
$\pi < \delta < \frac{3\pi}{2}$
$\delta = \frac{3\pi}{2}$
$\frac{3\pi}{2} < \delta < 2\pi$

Fig. 1.7 Elliptical polarization with various values of the phase difference δ.

(i) *Linearly polarized electric wave* ($\delta = m\pi$, $m = 0, \pm 1, \pm 2, \ldots$):

$$E_y/E_x = (-1)^m \frac{a_2}{a_1}.$$

(ii) *Right-handed circularly polarized electric wave** ($a_1 = a_2$, $\delta = \pi/2$):

$$E_y/E_x = \mathrm{e}^{-\mathrm{i}\pi/2} = -\mathrm{i}.$$

(iii) *Left-handed circularly polarized electric wave* ($a_1 = a_2$, $\delta = -\pi/2$):

$$E_y/E_x = \mathrm{e}^{\mathrm{i}\pi/2} = \mathrm{i}.$$

More generally it may be shown that for right-handed elliptical polarization, the ratio E_y/E_x has a negative imaginary part, whereas for left-handed elliptical polarization the imaginary part is positive.

Fig. 1.7 illustrates how the polarization ellipse changes with varying δ.

(c) Characterization of the state of polarization by Stokes parameters

To characterize the polarization ellipse three independent quantities are necessary, e.g. the amplitudes a_1 and a_2 and the phase difference δ, or the major and minor axes a, b and the angle χ which specifies the orientation of the ellipse. For practical puposes it is convenient to characterize the state of polarization by certain parameters which are all of the same physical dimensions, and which were introduced by G. G. Stokes in 1852, in his investigations relating to partially polarized light. We shall define them later (§10.9.3) in their full generality. We will also show there that for any given wave these parameters may be determined from simple experiments.

The *Stokes parameters* of a plane monochromatic wave are the four quantities

$$\left.\begin{aligned} s_0 &= a_1^2 + a_2^2, \\ s_1 &= a_1^2 - a_2^2, \\ s_2 &= 2a_1 a_2 \cos\delta, \\ s_3 &= 2a_1 a_2 \sin\delta. \end{aligned}\right\} \tag{43}$$

Only three of them are independent since they are related by the identity

* NB. This is right-handed according to the traditional, not the natural nomenclature.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$s_0^2 = s_1^2 + s_2^2 + s_3^2. \tag{44}$$

The parameter s_0 is evidently proportional to the intensity of the wave. The parameters s_1, s_2, and s_3 are related in a simple way to the angle ψ $(0 \leqslant \psi < \pi)$ which specifies the orientation of the ellipse and the angle χ $(-\pi/4 \leqslant \chi \leqslant \pi/4)$ which characterizes the ellipticity and the sense in which the ellipse is being described. In fact the following relations hold:

$$s_1 = s_0 \cos 2\chi \cos 2\psi, \tag{45a}$$

$$s_2 = s_0 \cos 2\chi \sin 2\psi, \tag{45b}$$

$$s_3 = s_0 \sin 2\chi. \tag{45c}$$

The relation (45c) follows from (25) and (29). To derive the other two relations we note that according to the equation preceding (26),

$$s_2 = s_1 \tan 2\psi. \tag{46}$$

The relation (45a) follows on substitution from (46) and from (45c) into (44). Finally, (45b) is obtained on substitution from (45a) into (46).

The relations (45) indicate a simple geometrical representation of all the different states of polarization: s_1, s_2 and s_3 may be regarded as the Cartesian coordinates of a point P on a sphere Σ of radius s_0, such that 2χ and 2ψ are the spherical angular coordinates of this point (see Fig. 1.8). Thus *to every possible state of polarization of a plane monochromatic wave of a given intensity* ($s_0 =$ *constant*), *there corresponds one point on* Σ *and vice versa*. Since χ is positive or negative according as the polarization is right-handed or left-handed, it follows from (45c) that right-handed polarization is represented by points on Σ which lie above the equatorial plane (x, y-plane), and left-

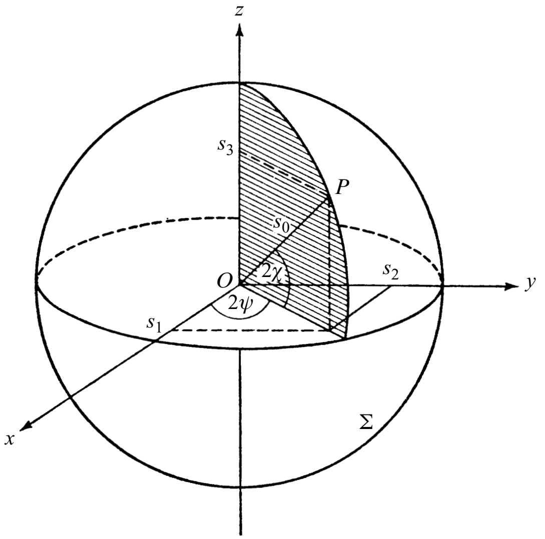

Fig. 1.8 Poincaré's representation of the state of polarization of a monochromatic wave. (The Poincaré sphere.)

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

handed polarization by points on Σ which lie below this plane. Further, for linearly polarized light the phase difference δ is zero or an integral multiple of π; according to (43) the Stokes parameter s_3 is then zero, so that linear polarization is represented by points in the equatorial plane. For circular polarization $a_1 = a_2$ and $\delta = \pi/2$ or $-\pi/2$ according as the polarization is right- or left-handed; hence right-handed circular polarization is represented by the north pole ($s_1 = s_2 = 0,\ s_3 = s_0$) and left-handed circular polarization by the south pole ($s_1 = s_2 = 0,\ s_3 = -s_0$). This geometrical representation of different states of polarization by points on a sphere is due to Poincaré,* and is very useful in crystal optics for determining the effect of crystalline media on the state of polarization of light traversing it.† The sphere Σ is called *the Poincaré sphere*.

1.4.3 Harmonic vector waves of arbitrary form

The main results of the preceding sections can easily be extended to time-harmonic waves of more complicated form.

A general time-harmonic real vector wave $\mathbf{V}(\mathbf{r},\ t)$ is a solution of the vector wave equation, whose Cartesian components $V_x,\ V_y,\ V_z$ are represented by expressions of the form §1.3 (23):

$$\left.\begin{aligned} V_x(\mathbf{r},\ t) &= a_1(\mathbf{r})\cos[\omega t - g_1(\mathbf{r})],\\ V_y(\mathbf{r},\ t) &= a_2(\mathbf{r})\cos[\omega t - g_2(\mathbf{r})],\\ V_z(\mathbf{r},\ t) &= a_3(\mathbf{r})\cos[\omega t - g_3(\mathbf{r})], \end{aligned}\right\} \tag{47}$$

where a_s and g_s ($s = 1,\ 2,\ 3$) are real functions of position. For the plane harmonic wave considered in the preceding section the a's were constant, and $g_s(\mathbf{r}) = \mathbf{k}\cdot\mathbf{r} - \delta_s$.

It will be convenient to write (47) in the form

$$V_x(\mathbf{r},\ t) = p_x(\mathbf{r})\cos\omega t + q_x(\mathbf{r})\sin\omega t \quad \text{etc.,} \tag{48}$$

where

$$\left.\begin{aligned} p_x(\mathbf{r}) &= a_1(\mathbf{r})\cos g_1(\mathbf{r}),\\ q_x(\mathbf{r}) &= a_1(\mathbf{r})\sin g_1(\mathbf{r}). \end{aligned}\right\} \tag{49}$$

If a new set of axes is chosen, then each component of $\mathbf{V}$ with respect to the new set will again be of the form (47), since each new component is a linear combination of the old ones and can therefore involve only the sum of terms in $\cos\omega t$ and $\sin\omega t$.

We may regard $(p_x,\ p_y,\ p_z)$ and $(q_x,\ q_y,\ q_z)$ as components of two real vectors $\mathbf{p}$ and $\mathbf{q}$. Then

$$\mathbf{V}(\mathbf{r},\ t) = \mathbf{p}(\mathbf{r})\cos\omega t + \mathbf{q}(\mathbf{r})\sin\omega t. \tag{50}$$

By Fourier analysis, an arbitrary vector wave may be expressed as superposition of waves of this type.

* H. Poincaré, *Théorie Mathématique de la Lumière*, Vol. 2 (Paris, Georges Carré, 1892) Chap. 12.

† Examples illustrating the method and references to the relevant literature are given in H. G. Jerrard, *J. Opt. Soc. Amer.*, **44** (1954), 634. See also the paper by M. J. Walker, referred to in §10.9.3 and S. Pancharatnam, *Proc. Ind. Acad. Sci.*, A, **44** (1956), 247.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

As in the case of scalar waves, it is often convenient to use a complex representation. We write (50) in the form

$$\mathbf{V}(\mathbf{r},\ t) = \mathcal{R}\{\mathbf{U}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}\}, \tag{51}$$

where $\mathbf{U}$ is the complex vector

$$\mathbf{U}(\mathbf{r}) = \mathbf{p}(\mathbf{r}) + \mathrm{i}\mathbf{q}(\mathbf{r}), \tag{52}$$

$\mathcal{R}$ denoting the real part. When the operations on $\mathbf{V}$ are linear one can, as in the corresponding scalar case, operate directly on the complex quantity, omitting the symbol $\mathcal{R}$ altogether. The real part of the final expression is then understood to represent the physical quantity in question.

Operations with complex vectors follow the usual rules of vector algebra and of algebra of complex numbers. For example, the conjugate of $\mathbf{U}$ is the vector

$$\mathbf{U}^{\star} = \mathbf{p} - \mathrm{i}\mathbf{q}.$$

Similarly

$$\mathbf{U}^2 = \mathbf{U}\cdot\mathbf{U} = \mathbf{p}^2 - \mathbf{q}^2 + 2\mathrm{i}\mathbf{p}\cdot\mathbf{q},$$

$$\mathbf{U}\cdot\mathbf{U}^{\star} = (\mathbf{p}+\mathrm{i}\mathbf{q})\cdot(\mathbf{p}-\mathrm{i}\mathbf{q}) = \mathbf{p}^2 + \mathbf{q}^2,$$

etc.

To illustrate calculations with complex vectors we shall derive formulae needed later for the energy densities and the Poynting vector in a time-harmonic electromagnetic field. The electric and magnetic vectors then are of the form

$$\left.\begin{aligned} \mathbf{E}(\mathbf{r},\ t) &= \mathcal{R}\{\mathbf{E}_0(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}\} = \tfrac{1}{2}[\mathbf{E}_0(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t} + \mathbf{E}_0^{\star}(\mathbf{r})\mathrm{e}^{\mathrm{i}\omega t}],\\ \mathbf{H}(\mathbf{r},\ t) &= \mathcal{R}\{\mathbf{H}_0(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}\} = \tfrac{1}{2}[\mathbf{H}_0(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t} + \mathbf{H}_0^{\star}(\mathbf{r})\mathrm{e}^{\mathrm{i}\omega t}], \end{aligned}\right\} \tag{53}$$

$\mathbf{E}_0$ and $\mathbf{H}_0$ being complex vector functions of position. Since the optical frequencies are very large (ω is of order 10^{15} s^{-1}), one cannot observe the instantaneous values of any of the rapidly oscillating quantities, but only their time average taken over a time interval (say $-T' \leqslant t \leqslant T'$) which is large compared to the fundamental period $T = 2\pi/\omega$. In particular, the time-averaged electric energy density is

$$\langle w_e\rangle = \frac{1}{2T'}\int_{-T'}^{T'}\frac{\varepsilon}{8\pi}\mathbf{E}^2\,\mathrm{d}t = \frac{\varepsilon}{8\pi}\frac{1}{2T'}\int_{-T'}^{T'}\tfrac{1}{4}[\mathbf{E}_0^2\mathrm{e}^{-2\mathrm{i}\omega t} + 2\mathbf{E}_0\cdot\mathbf{E}_0^{\star} + \mathbf{E}_0^{\star 2}\mathrm{e}^{2\mathrm{i}\omega t}]\mathrm{d}t.$$

Now

$$\frac{1}{2T'}\int_{-T'}^{T'}\mathrm{e}^{-2\mathrm{i}\omega t}\,\mathrm{d}t = -\frac{1}{4\mathrm{i}\omega T'}[\mathrm{e}^{-2\mathrm{i}\omega t}]_{-T'}^{T'} = \frac{1}{4\pi}\frac{T}{T'}\sin 2\omega T'.$$

Since T' is assumed to be large compared to T, T/T' will be small compared to unity, so that the integral involving $\mathrm{e}^{-2\mathrm{i}\omega t}$ may be neglected. Similarly the integral involving $\mathrm{e}^{2\mathrm{i}\omega t}$ may also be neglected, and we finally obtain

$$\langle w_e\rangle = \frac{\varepsilon}{16\pi}\mathbf{E}_0\cdot\mathbf{E}_0^{\star}. \tag{54}$$

In a similar way the time average of the magnetic energy density is seen to be

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\langle w_m \rangle = \frac{\mu}{16\pi}\mathbf{H}_0 \cdot \mathbf{H}_0^\star. \tag{55}$$

The average of the Poynting vector is given by

$$\begin{aligned}
\langle \mathbf{S} \rangle &= \frac{1}{2T'}\int_{-T'}^{T} \frac{c}{4\pi}(\mathbf{E}\times\mathbf{H})\mathrm{d}t \\
&= \frac{c}{4\pi}\frac{1}{2T'}\int_{-T'}^{T} \tfrac{1}{4}[\mathbf{E}_0\times\mathbf{H}_0\mathrm{e}^{-2\mathrm{i}\omega t} + \mathbf{E}_0\times\mathbf{H}_0^\star + \mathbf{E}_0^\star\times\mathbf{H}_0 + \mathbf{E}_0^\star\times\mathbf{H}_0^\star\mathrm{e}^{2\mathrm{i}\omega t}]\mathrm{d}t \\
&\simeq \frac{c}{16\pi}(\mathbf{E}_0\times\mathbf{H}_0^\star + \mathbf{E}_0^\star\times\mathbf{H}_0) \\
&= \frac{c}{8\pi}\mathcal{R}(\mathbf{E}_0\times\mathbf{H}_0^\star).
\end{aligned} \tag{56}$$

The law of conservation of energy also takes a simple form. For a nonconducting medium ($\sigma = 0$) where no mechanical work is done, we find, on taking the time average of §1.1 (43) that

$$\operatorname{div}\langle \mathbf{S} \rangle = 0. \tag{57}$$

If we integrate (57) throughout an arbitrary volume which contains no radiator or absorber of energy, and apply Gauss' theorem, it follows that

$$\int \langle \mathbf{S} \rangle \cdot \mathbf{n}\,\mathrm{d}S = 0, \tag{58}$$

$\mathbf{n}$ being the outward normal to the surface; the integration is taken over the boundary. Thus the averaged total flux of energy through any closed surface is zero.

We now return to the general time-harmonic vector wave (50), and consider the behaviour of $\mathbf{V}$ at a point $\mathbf{r} = \mathbf{r}_0$ in space. In general, as time varies, the end point of $\mathbf{V}$ describes an ellipse, so that, like the plane wave, the wave (50) is in general also elliptically polarized. To see this we note first that, with varying time, the end point describes a curve in the plane specified by $\mathbf{p}(\mathbf{r}_0)$ and $\mathbf{q}(\mathbf{r}_0)$. Since $\mathbf{V}$ is periodic, the curve must be closed. Now we may set

$$(\mathbf{p} + \mathrm{i}\mathbf{q}) = (\mathbf{a} + \mathrm{i}\mathbf{b})\mathrm{e}^{\mathrm{i}\varepsilon}, \tag{59}$$

where ε is any scalar. In terms of $\mathbf{p}$, $\mathbf{q}$ and ε,

$$\left.\begin{aligned}
\mathbf{a} &= \mathbf{p}\cos\varepsilon + \mathbf{q}\sin\varepsilon, \\
\mathbf{b} &= -\mathbf{p}\sin\varepsilon + \mathbf{q}\cos\varepsilon.
\end{aligned}\right\} \tag{60}$$

Let us choose ε so that the vectors $\mathbf{a}$ and $\mathbf{b}$ are perpendicular to each other and let $|\mathbf{a}| \geqslant |\mathbf{b}|$. For $\mathbf{a}$ and $\mathbf{b}$ to be orthogonal, ε must satisfy the equation

$$(\mathbf{p}\cos\varepsilon + \mathbf{q}\sin\varepsilon)\cdot(-\mathbf{p}\sin\varepsilon + \mathbf{q}\cos\varepsilon) = 0, \tag{61}$$

i.e.

$$\tan 2\varepsilon = \frac{2\mathbf{p}\cdot\mathbf{q}}{\mathbf{p}^2 - \mathbf{q}^2}. \tag{62}$$

We now consider as parameters specifying the wave the five independent compo-

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

nents of the orthogonal vectors **a** and **b** and the corresponding phase factor ε, instead of the six rectangular components of **p** and **q**. Then from (51), (52) and (59),

$$\begin{aligned}\mathbf{V} &= \mathcal{R}\{(\mathbf{a}+\mathrm{i}\mathbf{b})\mathrm{e}^{-\mathrm{i}(\omega t-\varepsilon)}\}\\ &= \mathbf{a}\cos(\omega t-\varepsilon)+\mathbf{b}\sin(\omega t-\varepsilon).\end{aligned} \tag{63}$$

If we take Cartesian axes with origin at $\mathbf{r}_0$ and with the x and y directions along **a** and **b**, then

$$V_x = a\cos(\omega t-\varepsilon), \qquad V_y = b\sin(\omega t-\varepsilon), \qquad V_z = 0. \tag{64}$$

This represents an *ellipse* (the *polarization* ellipse)

$$\frac{V_x^2}{a^2}+\frac{V_y^2}{b^2}=1, \tag{65}$$

with semiaxes of lengths a and b along the x and y coordinate axes. By elementary geometry it may be shown that **p** and **q** are a pair of conjugate semidiameters of the ellipse.

As in the case of plane waves, the ellipse may be described in two possible senses, corresponding to left- and right-handed polarization; these are distinguished by the sign of the scalar triple product $[\mathbf{a},\ \mathbf{b},\ \nabla\varepsilon] = [\mathbf{p},\ \mathbf{q},\ \nabla\varepsilon]$.

The lengths of the semiaxes of the polarization ellipse are easily obtained from (60) and (62). From (60)

$$\begin{aligned}a^2 &= p^2\cos^2\varepsilon+q^2\sin^2\varepsilon+2\mathbf{p}\cdot\mathbf{q}\sin\varepsilon\cos\varepsilon\\ &= \tfrac{1}{2}(p^2+q^2)+\tfrac{1}{2}(p^2-q^2)\cos 2\varepsilon+\mathbf{p}\cdot\mathbf{q}\sin 2\varepsilon.\end{aligned}$$

From (62)

$$\sin 2\varepsilon = \frac{2\mathbf{p}\cdot\mathbf{q}}{\sqrt{(p^2-q^2)^2+4(\mathbf{p}\cdot\mathbf{q})^2}}, \qquad \cos 2\varepsilon = \frac{p^2-q^2}{\sqrt{(p^2-q^2)^2+4(\mathbf{p}\cdot\mathbf{q})^2}}.$$

Hence

$$a^2 = \tfrac{1}{2}\left[p^2+q^2+\sqrt{(p^2-q^2)^2+4(\mathbf{p}\cdot\mathbf{q})^2}\right].$$

Similarly one finds

$$b^2 = \tfrac{1}{2}\left[p^2+q^2-\sqrt{(p^2-q^2)^2+4(\mathbf{p}\cdot\mathbf{q})^2}\right]. \tag{66}$$

To find an expression for the angle between **a** and **p** we express the equation of the ellipse in parametric form:

$$V_x = a\cos\phi, \qquad V_y = b\sin\phi, \tag{67}$$

where ϕ is the eccentric angle (see Fig. 1.9). From elementary geometry we learn that this angle is related to the polar angle θ of the point $(V_x,\ V_y)$ by

$$\tan\theta = \frac{b}{a}\tan\phi. \tag{68}$$

Comparison of (64) and (67) shows that in the present case $\phi = \omega t-\varepsilon$. Now according to (50), $\mathbf{V} = \mathbf{p}$ when $t = 0$, so that the eccentric angle of **p** is $-\varepsilon$. Hence the angle ψ between **p** and **a** is given by

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

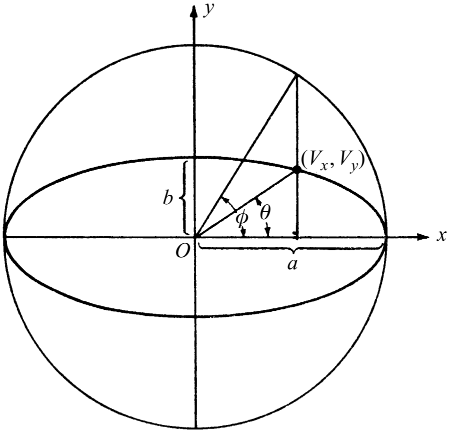

Fig. 1.9 Illustrating the notation relating to (67) and (68).

$$\tan\psi = \frac{b}{a}\tan\varepsilon. \tag{69}$$

If γ denotes the angle between **p** and **q**, and we introduce an auxiliary angle β such that

$$\frac{q}{p} = \tan\beta, \tag{70}$$

then (62) becomes

$$\begin{aligned}\tan 2\varepsilon &= \frac{2pq}{p^2 - q^2}\cos\gamma \\ &= \tan 2\beta\cos\gamma.\end{aligned} \tag{71}$$

Let us again summarize the results: If **p** and **q** are given, γ denotes the angle between these vectors, and β denotes the auxiliary angle defined by (70), then the principal semiaxes of the ellipse and the angle ψ which the major axis makes with **p** are given by

$$\left.\begin{aligned} a^2 &= \tfrac{1}{2}\left[p^2 + q^2 + \sqrt{(p^2 - q^2)^2 + 4p^2q^2\cos^2\gamma}\right], \\ b^2 &= \tfrac{1}{2}\left[p^2 + q^2 - \sqrt{(p^2 - q^2)^2 + 4p^2q^2\cos^2\gamma}\right], \\ \tan\psi &= \frac{b}{a}\tan\varepsilon, \end{aligned}\right\} \tag{72}$$

where

$$\tan 2\varepsilon = \tan 2\beta\cos\gamma. \tag{73}$$

As for plane waves, there are two cases of special interest, namely when the ellipse degenerates into a circle or a straight line. For a *circularly polarized* wave, **a** and **b** and consequently ε are not determined. According to (62), for this to be the case,

$$\mathbf{p}\cdot\mathbf{q} = \mathbf{p}^2 - \mathbf{q}^2 = 0. \tag{74}$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

For a *linearly polarized* wave, the minor axis vanishes ($b^2 = 0$) and (66) then gives

$$p^2 q^2 = (\mathbf{p}\cdot\mathbf{q})^2. \tag{75}$$

Finally we stress that the term *polarization* refers to the behaviour at a particular *point* in the field, and that the state of polarization will therefore in general be different at different points of the field. Thus a wave may be linearly or circularly polarized at some points and elliptically at others.* Only in special cases, as, for example, for the homogeneous plane wave, will the state of polarization be the same at every point in the field.

1.5 Reflection and refraction of a plane wave

In §1.1.3 relations were derived which the field vectors must satisfy across surfaces at which the physical properties of the medium are discontinuous. These formulae will now be applied to the study of the propagation of a plane wave incident on a plane boundary between two homogeneous isotropic media.

1.5.1 The laws of reflection and refraction

When a plane wave falls on to a boundary between two homogeneous media of different optical properties, it is split into two waves: a transmitted wave proceeding into the second medium and a reflected wave propagated back into the first medium. The existence of these two waves can be demonstrated from the boundary conditions, since it is easily seen that these conditions cannot be satisfied without postulating both the transmitted and the reflected wave. We shall tentatively assume that these waves are also plane, and derive expressions for their directions of propagation and their amplitudes.

A plane wave propagated in the direction specified by the unit vector† $\mathbf{s}^{(i)}$ is completely determined when the time behaviour at one particular point in space is known. For if $\mathbf{F}(t)$ represents the time behaviour at any one point, the time behaviour at another point, whose position vector relative to the first point is $\mathbf{r}$, is given by $\mathbf{F}[t - (\mathbf{r}\cdot\mathbf{s})/v]$. At the boundary between the two media, the time variation of the secondary fields will be the same as that of the incident primary field. Hence, if $\mathbf{s}^{(r)}$ and $\mathbf{s}^{(t)}$ denote unit vectors in the direction of propagation of the reflected and transmitted wave, one has, on equating the arguments of the three wave functions at a point $\mathbf{r}$ on the boundary plane $z = 0$:

$$t - \frac{\mathbf{r}\cdot\mathbf{s}^{(i)}}{v_1} = t - \frac{\mathbf{r}\cdot\mathbf{s}^{(r)}}{v_1} = t - \frac{\mathbf{r}\cdot\mathbf{s}^{(t)}}{v_2}, \tag{1}$$

v_1 and v_2 being the velocities of propagation in the two media. Written out more explicitly, one has, with $\mathbf{r} \equiv x, y, 0$:

* General properties of time-harmonic electromagnetic waves of arbitrary form but with at least one of the field vectors linearly polarized have been investigated by A. Nisbet and E. Wolf, *Proc. Camb. Phil. Soc.*, **50** (1954), 614.

† The suffixes i, r and t refer throughout to the incident, reflected and transmitted (refracted) waves respectively.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\frac{xs_x^{(i)} + ys_y^{(i)}}{v_1} = \frac{xs_x^{(r)} + ys_y^{(r)}}{v_1} = \frac{xs_x^{(t)} + ys_y^{(t)}}{v_2}. \tag{2}$$

Since (2) must hold for all values x and y on the boundary,

$$\frac{s_x^{(i)}}{v_1} = \frac{s_x^{(r)}}{v_1} = \frac{s_x^{(t)}}{v_2}, \qquad \frac{s_y^{(i)}}{v_1} = \frac{s_y^{(r)}}{v_1} = \frac{s_y^{(t)}}{v_2}. \tag{3}$$

The plane specified by $\mathbf{s}^{(i)}$ and the normal to the boundary is called the *plane of incidence*. Eqs. (3) show that both $\mathbf{s}^{(t)}$ and $\mathbf{s}^{(r)}$ lie in this plane.

Taking the plane of incidence as the x, z-plane and denoting by θ_i, θ_r and θ_t the angle which $\mathbf{s}^{(i)}$, $\mathbf{s}^{(r)}$ and $\mathbf{s}^{(t)}$ make with Oz, one has (see Fig. 1.10)

$$\left.\begin{aligned} s_x^{(i)} &= \sin\theta_i, & s_y^{(i)} &= 0, & s_z^{(i)} &= \cos\theta_i, \\ s_x^{(r)} &= \sin\theta_r, & s_y^{(r)} &= 0, & s_z^{(r)} &= \cos\theta_r, \\ s_x^{(t)} &= \sin\theta_t, & s_y^{(t)} &= 0, & s_z^{(t)} &= \cos\theta_t. \end{aligned}\right\} \tag{4}$$

For waves propagated from the first into the second medium, the z components of the $\mathbf{s}$ vectors are positive; for those propagated in the opposite sense, they are negative:

$$s_z^{(i)} = \cos\theta_i \geqslant 0, \qquad s_z^{(r)} = \cos\theta_r \leqslant 0, \qquad s_z^{(t)} = \cos\theta_t \geqslant 0. \tag{5}$$

The first set in (3) gives, on substituting from (4)

$$\frac{\sin\theta_i}{v_1} = \frac{\sin\theta_r}{v_1} = \frac{\sin\theta_t}{v_2}. \tag{6}$$

Hence, $\sin\theta_r = \sin\theta_i$, and we find, also using (5), that $\cos\theta_r = -\cos\theta_i$, so that

$$\theta_r = \pi - \theta_i. \tag{7}$$

This relation, together with the statement that the reflected wave normal $\mathbf{s}^{(r)}$ is in the plane of incidence, constitute the *law of reflection*.

Also from (6), using Maxwell's relation §1.2 (14) connecting the refractive index and the dielectric constant,

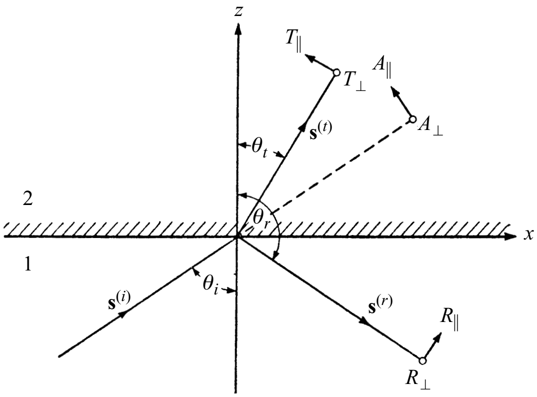

Fig. 1.10 Refraction and reflection of a plane wave. Plane of incidence.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\frac{\sin\theta_i}{\sin\theta_t} = \frac{v_1}{v_2} = \sqrt{\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}} = \frac{n_2}{n_1} = n_{12}. \tag{8}$$

The relation $\sin\theta_i/\sin\theta_t = n_2/n_1$, together with the statement that the refracted wave normal $\mathbf{s}^{(t)}$ is in the plane of incidence, constitute the *law of refraction* (or Snell's law).

When $n_2 > n_1$, then $n_{12} > 1$, and one says that the second medium is optically denser than the first medium. In this case, by (8),

$$\sin\theta_t = \frac{1}{n_{12}}\sin\theta_i < \sin\theta_i, \tag{9}$$

so that there is a real angle θ_t of refraction for every angle of incidence. If, however, the second medium is optically less dense than the first medium (i.e. if $n_{12} < 1$), one obtains a real value for θ_t only for those incident angles θ_i for which $\sin\theta_i \leqslant n_{12}$. For larger values of the angle of incidence, so-called *total reflection* takes place. It will be considered separately in §1.5.4.

1.5.2 *Fresnel formulae*

Next we consider the amplitudes of the reflected and the transmitted waves. We shall assume that the two (homogeneous and isotropic) media are both of zero conductivity and consequently perfectly transparent; their magnetic permeabilities will then in fact differ from unity by negligible amounts, and accordingly we take $\mu_1 = \mu_2 = 1$.

Let A be the amplitude of the electric vector of the incident field. We take A to be complex, with its phase equal to the constant part of the argument of the wave function; the variable part is

$$\tau_i = \omega\left(t - \frac{\mathbf{r}\cdot\mathbf{s}^{(i)}}{v_1}\right) = \omega\left(t - \frac{x\sin\theta_i + z\cos\theta_i}{v_1}\right). \tag{10}$$

We resolve each vector into components parallel (denoted by subscript $\|$) and perpendicular (subscript $\perp$) to the plane of incidence. The choice of the positive directions for the parallel components is indicated in Fig. 1.10. The perpendicular components must be visualized at right angles to the plane of the figure.

The components of the electric vector of the incident field then are

$$E_x^{(i)} = -A_\| \cos\theta_i \mathrm{e}^{-\mathrm{i}\tau_i}, \quad E_y^{(i)} = A_\perp \mathrm{e}^{-\mathrm{i}\tau_i}, \quad E_z^{(i)} = A_\| \sin\theta_i \mathrm{e}^{-\mathrm{i}\tau_i}. \tag{11}$$

The components of the magnetic vector are immediately obtained by using §1.4 (4) (with $\mu = 1$):

$$\mathbf{H} = \sqrt{\varepsilon}\mathbf{s}\times\mathbf{E}. \tag{12}$$

This gives

$$H_x^{(i)} = -A_\perp \cos\theta_i\sqrt{\varepsilon_1}\mathrm{e}^{-\mathrm{i}\tau_i}, \; H_y^{(i)} = -A_\|\sqrt{\varepsilon_1}\mathrm{e}^{-\mathrm{i}\tau_i}, \; H_z^{(i)} = A_\perp \sin\theta_i\sqrt{\varepsilon_1}\mathrm{e}^{-\mathrm{i}\tau_i}. \tag{13}$$

Similarly if T and R are the complex amplitudes of the transmitted and reflected waves, the corresponding components of the electric and magnetic vectors are:

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Transmitted field

$$\left.\begin{aligned} &E_x^{(t)} = -T_\parallel \cos\theta_t \mathrm{e}^{-\mathrm{i}\tau_t}, && E_y^{(t)} = T_\perp \mathrm{e}^{-\mathrm{i}\tau_t}, && E_z^{(t)} = T_\parallel \sin\theta_t \mathrm{e}^{-\mathrm{i}\tau_t}, \\ &H_x^{(t)} = -T_\perp \cos\theta_t \sqrt{\varepsilon_2}\mathrm{e}^{-\mathrm{i}\tau_t}, && H_y^{(t)} = -T_\parallel \sqrt{\varepsilon_2}\mathrm{e}^{-\mathrm{i}\tau_t}, && H_z^{(t)} = T_\perp \sin\theta_t \sqrt{\varepsilon_2}\mathrm{e}^{-\mathrm{i}\tau_t}, \end{aligned}\right\} \tag{14}$$

with

$$\tau_t = \omega\left(t - \frac{\mathbf{r}\cdot\mathbf{s}^{(t)}}{v_2}\right) = \omega\left(t - \frac{x\sin\theta_t + z\cos\theta_t}{v_2}\right). \tag{15}$$

Reflected field

$$\left.\begin{aligned} &E_x^{(r)} = -R_\parallel \cos\theta_r \mathrm{e}^{-\mathrm{i}\tau_r}, && E_y^{(r)} = R_\perp \mathrm{e}^{-\mathrm{i}\tau_r}, && E_z^{(r)} = R_\parallel \sin\theta_r \mathrm{e}^{-\mathrm{i}\tau_r}, \\ &H_x^{(r)} = -R_\perp \cos\theta_r \sqrt{\varepsilon_1}\mathrm{e}^{-\mathrm{i}\tau_r}, && H_y^{(r)} = -R_\parallel \sqrt{\varepsilon_1}\mathrm{e}^{-\mathrm{i}\tau_r}, && H_z^{(r)} = R_\perp \sin\theta_r \sqrt{\varepsilon_1}\mathrm{e}^{-\mathrm{i}\tau_r}, \end{aligned}\right\} \tag{16}$$

with

$$\tau_r = \omega\left(t - \frac{\mathbf{r}\cdot\mathbf{s}^{(r)}}{v_1}\right) = \omega\left(t - \frac{x\sin\theta_r + z\cos\theta_r}{v_1}\right). \tag{17}$$

The boundary conditions §1.1 (23), §1.1 (25) demand that across the boundary the tangential components of **E** and **H** should be continuous. Hence we must have

$$\left.\begin{aligned} E_x^{(i)} + E_x^{(r)} = E_x^{(t)}, \qquad E_y^{(i)} + E_y^{(r)} = E_y^{(t)}, \\ H_x^{(i)} + H_x^{(r)} = H_x^{(t)}, \qquad H_y^{(i)} + H_y^{(r)} = H_y^{(t)}, \end{aligned}\right\} \tag{18}$$

the conditions §1.1 (15) and §1.1 (19), being then automatically fulfilled for the normal components of **B** and **D**. On substituting into (18) for all the components, and using the fact that $\cos\theta_r = \cos(\pi - \theta_i) = -\cos\theta_i$, we obtain the four relations

$$\left.\begin{aligned} \cos\theta_i(A_\parallel - R_\parallel) &= \cos\theta_t T_\parallel, \\ A_\perp + R_\perp &= T_\perp, \\ \sqrt{\varepsilon_1}\cos\theta_i(A_\perp - R_\perp) &= \sqrt{\varepsilon_2}\cos\theta_t T_\perp, \\ \sqrt{\varepsilon_1}(A_\parallel + R_\parallel) &= \sqrt{\varepsilon_2} T_\parallel. \end{aligned}\right\} \tag{19}$$

We note that the equations (19) fall into two groups, one of which contains only the components parallel to the plane of incidence, whilst the other contains only those which are perpendicular to the plane of incidence. *These two kinds of waves are, therefore, independent of one another.*

We can solve (19) for the components of the reflected and transmitted waves in terms of those of the incident wave, giving (using again the Maxwell relation $n = \sqrt{\varepsilon}$)

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\left.\begin{aligned}T_{\|} &= \frac{2n_1\cos\theta_i}{n_2\cos\theta_i + n_1\cos\theta_t}A_{\|},\\ T_{\perp} &= \frac{2n_1\cos\theta_i}{n_1\cos\theta_i + n_2\cos\theta_t}A_{\perp},\end{aligned}\right\} \tag{20}$$

$$\left.\begin{aligned}R_{\|} &= \frac{n_2\cos\theta_i - n_1\cos\theta_t}{n_2\cos\theta_i + n_1\cos\theta_t}A_{\|},\\ R_{\perp} &= \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t}A_{\perp}.\end{aligned}\right\} \tag{21}$$

Eqs. (20) and (21) are called *Fresnel formulae*, having first been derived in a slightly less general form by Fresnel in 1823, on the basis of his elastic theory of light. They are usually written in the following alternative form, which may be obtained from (20) and (21) by using the law of refraction (8):

$$\left.\begin{aligned}T_{\|} &= \frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i+\theta_t)\cos(\theta_i-\theta_t)}A_{\|},\\ T_{\perp} &= \frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i+\theta_t)}A_{\perp},\end{aligned}\right\} \tag{20a}$$

$$\left.\begin{aligned}R_{\|} &= \frac{\tan(\theta_i-\theta_t)}{\tan(\theta_i+\theta_t)}A_{\|},\\ R_{\perp} &= -\frac{\sin(\theta_i-\theta_t)}{\sin(\theta_i+\theta_t)}A_{\perp}.\end{aligned}\right\} \tag{21a}$$

Since θ_i and θ_t are real (the case of total reflection being excluded for the present), the trigonometrical factors on the right-hand sides of (20a) and (21a) will also be real. Consequently the phase of each component of the reflected or transmitted wave is either equal to the phase of the corresponding component of the incident wave or differs from it by π. Since $T_{\|}$ and $T_{\perp}$ have the same signs as $A_{\|}$ and $A_{\perp}$, the phase of the transmitted wave is actually equal to that of the incident wave. In the case of the reflected wave, the phase will, however, depend on the relative magnitudes of θ_i and θ_t. For, if the second medium is optically denser than the first ($\varepsilon_2 > \varepsilon_1$), then $\theta_t < \theta_i$; according to (21), the signs of $R_{\perp}$ and $A_{\perp}$ are different and the phases therefore differ* by π. Under the same circumstances $\tan(\theta_i - \theta_t)$ is positive, but the denominator $\tan(\theta_i + \theta_t)$ becomes negative for $\theta_i + \theta_t > \pi/2$, and the phase of $R_{\|}$ and $A_{\|}$ then differ by π. Similar considerations apply when the second medium is optically less dense than the first.

For *normal incidence*, $\theta_i = 0$ and consequently $\theta_t = 0$, and (20) and (21) reduce to

* From (11) and (16) it follows that in the plane $z = 0$

$$\frac{E_y^{(r)}}{E_y^{(i)}} = \frac{R_{\perp}}{A_{\perp}}, \qquad \frac{E_x^{(r)}}{E_x^{(i)}} = -\frac{R_{\|}}{A_{\|}}.$$

This result implies that in the plane $z = 0$, the phases of $E_y^{(r)}$ and $E_y^{(i)}$ differ by π, whereas the phases of $E_x^{(r)}$ and $E_x^{(i)}$ are equal to each other. This difference in the behaviour of the phases of the y- and x-components is rather formal, arising from the way in which the angle of reflection θ_r was defined (see Fig. 1.10).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\left.\begin{aligned} T_{\|} &= \frac{2}{n+1} A_{\|}, \\ T_{\perp} &= \frac{2}{n+1} A_{\perp}, \end{aligned}\right\} \tag{22}$$

$$\left.\begin{aligned} R_{\|} &= \frac{n-1}{n+1} A_{\|}, \\ R_{\perp} &= -\frac{n-1}{n+1} A_{\perp}, \end{aligned}\right\} \tag{23}$$

where $n = n_2/n_1$.

1.5.3 The reflectivity and transmissivity; polarization on reflection and refraction

We shall now examine how the energy of the incident field is divided between the two secondary fields.

According to §1.4 (8), the light intensity is given (again assuming $\mu = 1$) by

$$S = \frac{c}{4\pi}\sqrt{\varepsilon}E^2 = \frac{cn}{4\pi}E^2. \tag{24}$$

The amount of energy in the primary wave which is incident on a unit area of the boundary per second is therefore

$$J^{(i)} = S^{(i)} \cos\theta_i = \frac{cn_1}{4\pi}|A|^2 \cos\theta_i, \tag{25}$$

and the energies of the reflected and transmitted wave leaving a unit area of the boundary per second are given by similar expressions:

$$\left.\begin{aligned} J^{(r)} &= S^{(r)} \cos\theta_i = \frac{cn_1}{4\pi}|R|^2 \cos\theta_i, \\ J^{(t)} &= S^{(t)} \cos\theta_t = \frac{cn_2}{4\pi}|T|^2 \cos\theta_t. \end{aligned}\right\} \tag{26}$$

The ratios

$$\mathcal{R} = \frac{J^{(r)}}{J^{(i)}} = \frac{|R|^2}{|A|^2} \quad \text{and} \quad \mathcal{T} = \frac{J^{(t)}}{J^{(i)}} = \frac{n_2 \cos\theta_t}{n_1 \cos\theta_i}\frac{|T|^2}{|A|^2} \tag{27}$$

are called the *reflectivity* and *transmissivity* respectively.* It can easily be verified that, in agreement with the law of conservation of energy,

$$\mathcal{R} + \mathcal{T} = 1. \tag{28}$$

The reflectivity and transmissivity depend on the polarization of the incident wave. They may be expressed in terms of the reflectivity and transmissivity associated with polarizations in the parallel and perpendicular directions, respectively.

Let α_i be the angle which the **E** vector of the incident wave makes with the plane of incidence. Then

* If $\mu \neq 1$, the factor n_2/n_1 in the expression for $\mathcal{T}$ must be replaced by $\sqrt{\varepsilon_2/\mu_2}\Big/\sqrt{\varepsilon_1/\mu_1}$, as is immediately evident from §1.4 (8).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$A_{\|} = A\cos\alpha_i, \qquad A_{\perp} = A\sin\alpha_i. \tag{29}$$

Let

$$\left.\begin{aligned} J_{\|}^{(i)} &= \frac{cn_1}{4\pi}|A_{\|}|^2\cos\theta_i = J^{(i)}\cos^2\alpha_i, \\ J_{\perp}^{(i)} &= \frac{cn_1}{4\pi}|A_{\perp}|^2\cos\theta_i = J^{(i)}\sin^2\alpha_i, \end{aligned}\right\} \tag{30}$$

and

$$\left.\begin{aligned} J_{\|}^{(r)} &= \frac{cn_1}{4\pi}|R_{\|}|^2\cos\theta_i, \\ J_{\perp}^{(r)} &= \frac{cn_1}{4\pi}|R_{\perp}|^2\cos\theta_i. \end{aligned}\right\} \tag{31}$$

Then

$$\begin{aligned} \mathcal{R} = \frac{J^{(r)}}{J^{(i)}} = \frac{J_{\|}^{(r)} + J_{\perp}^{(r)}}{J^{(i)}} &= \frac{J_{\|}^{(r)}}{J_{\|}^{(i)}}\cos^2\alpha_i + \frac{J_{\perp}^{(r)}}{J_{\perp}^{(i)}}\sin^2\alpha_i \\ &= \mathcal{R}_{\|}\cos^2\alpha_i + \mathcal{R}_{\perp}\sin^2\alpha_i, \end{aligned} \tag{32}$$

where

$$\left.\begin{aligned} \mathcal{R}_{\|} &= \frac{J_{\|}^{(r)}}{J_{\|}^{(i)}} = \frac{|R_{\|}|^2}{|A_{\|}|^2} = \frac{\tan^2(\theta_i - \theta_t)}{\tan^2(\theta_i + \theta_t)}, \\ \mathcal{R}_{\perp} &= \frac{J_{\perp}^{(r)}}{J_{\perp}^{(i)}} = \frac{|R_{\perp}|^2}{|A_{\perp}|^2} = \frac{\sin^2(\theta_i - \theta_t)}{\sin^2(\theta_i + \theta_t)}. \end{aligned}\right\} \tag{33}$$

In a similar way, we obtain

$$\mathcal{T} = \frac{J^{(t)}}{J^{(i)}} = \mathcal{T}_{\|}\cos^2\alpha_i + \mathcal{T}_{\perp}\sin^2\alpha_i, \tag{34}$$

with

$$\left.\begin{aligned} \mathcal{T}_{\|} &= \frac{J_{\|}^{(t)}}{J_{\|}^{(i)}} = \frac{n_2\cos\theta_t}{n_1\cos\theta_i}\frac{|T_{\|}|^2}{|A_{\|}|^2} = \frac{\sin 2\theta_i \sin 2\theta_t}{\sin^2(\theta_i + \theta_t)\cos^2(\theta_i - \theta_t)}, \\ \mathcal{T}_{\perp} &= \frac{J_{\perp}^{(t)}}{J_{\perp}^{(i)}} = \frac{n_2\cos\theta_t}{n_1\cos\theta_i}\frac{|T_{\perp}|^2}{|A_{\perp}|^2} = \frac{\sin 2\theta_i \sin 2\theta_t}{\sin^2(\theta_i + \theta_t)}. \end{aligned}\right\} \tag{35}$$

Again we may verify that

$$\mathcal{R}_{\|} + \mathcal{T}_{\|} = 1, \qquad \mathcal{R}_{\perp} + \mathcal{T}_{\perp} = 1. \tag{36}$$

For *normal incidence* the distinction between the parallel and perpendicular components disappears, and one has from (22), (23) and (27)

$$\left.\begin{aligned} \mathcal{R} &= \left(\frac{n-1}{n+1}\right)^2, \\ \mathcal{T} &= \frac{4n}{(n+1)^2}. \end{aligned}\right\} \tag{37}$$

It is seen from (37) that

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\lim_{n\to 1} \mathcal{R} = 0, \qquad \lim_{n\to 1} \mathcal{T} = 1. \tag{38}$$

Similar results also hold for the limiting values of $\mathcal{R}_\parallel$, $\mathcal{T}_\parallel$, and $\mathcal{R}_\perp$, $\mathcal{T}_\perp$, as can easily be seen from (33) and (35), making use of the fact that, according to the law of refraction, $\theta_t \to \theta_i$ as $n \to 1$. Hence the smaller the difference in the optical densities of the two media, the less energy is carried away by the reflected wave.

The denominators in (33) and (35) are finite, except when $\theta_i + \theta_t = \pi/2$. Then $\tan(\theta_i + \theta_t) = \infty$ and consequently $\mathcal{R}_\parallel = 0$. In this case (see Fig. 1.11) the reflected and transmitted rays are perpendicular to each other, and it follows from the law of refraction (since now $\sin\theta_t = \sin[(\pi/2) - \theta_i] = \cos\theta_i$) that

$$\tan\theta_i = n. \tag{39}$$

The angle θ_i given by (39) is called the *polarizing* or *Brewster angle*; its significance was noted first in 1815 by David Brewster (1781–1868): *If light is incident under this angle, the electric vector of the reflected light has no component in the plane of incidence*. One usually says that the light is then polarized 'in the plane of incidence'. According to this traditional terminology the plane of polarization is therefore the plane which contains the magnetic vector and the direction of propagation. For reasons already explained in §1.4.2 it is, however, better not to use this term. The above result is often called *Brewster's law.*

In Fig. 1.12 the reflectivity for glass of refractive index 1.52 is plotted against the angle of incidence θ_i. The numbers along the upper horizontal refer to the angle of refraction θ_t. The zero value of the curve (a) (for $\mathcal{R}_\parallel$) corresponds to the polarizing angle $\tan^{-1} 1.52 = 56°40'$.

The refractive indices with respect to air are usually of the order of 1.5 at optical wavelengths, but at radio wavelengths they are much larger, there being a corresponding increase in the polarizing angle. For example, at optical wavelengths the refractive index of water is about 1.3 and the polarizing angle 53°. For radio wavelengths its value is about 9, the polarizing angle then being in the neighbourhood of 84°.

According to (32), the curve (b) in Fig. 1.12 is seen to correspond to $\alpha = 45°$. The same curve, as will now be shown, represents also the reflectivity $\overline{\mathcal{R}}$ for natural light,

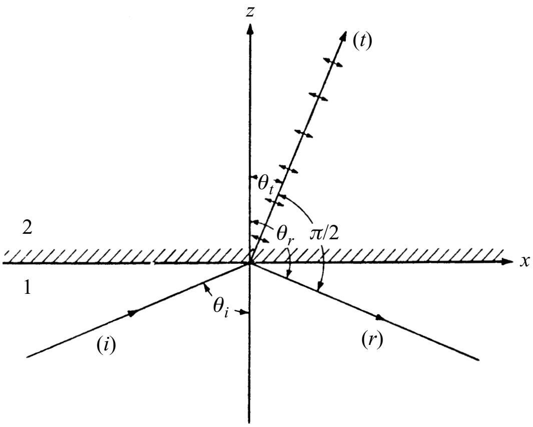

Fig. 1.11 Illustrating the polarizing (Brewster's) angle.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

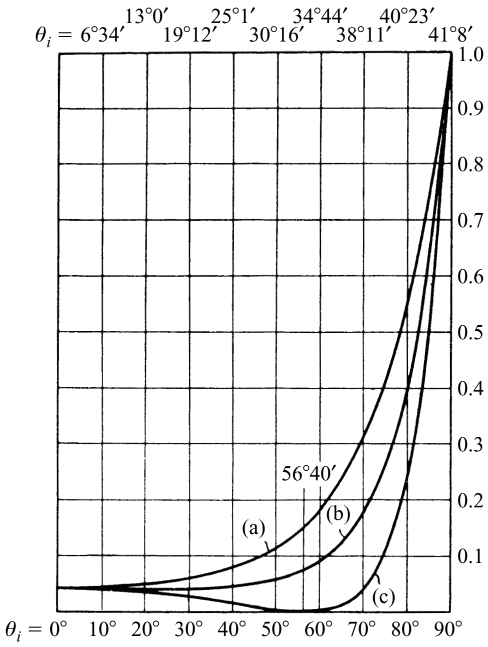

Fig. 1.12 Intensity of reflected light as a function of the angle of incidence: (a) $\mathcal{R}_\perp$; (b) $\frac{1}{2}(\mathcal{R}_\parallel + \mathcal{R}_\perp)$; (c) $\mathcal{R}_\parallel$. [After O. D. Chwolson, *Lehrb. d. Physik* (Braunschweig, Vieweg), Bd. 2. 2 Aufl. (1922), p. 716.]

i.e. for light obtained from a body which is made to glow, by raising its temperature. The directions of vibration of natural light vary rapidly in a random, irregular manner. The corresponding reflectivity $\overline{\mathcal{R}}$ may be obtained by averaging over all directions. Since the averages of $\sin^2\alpha$ and $\cos^2\alpha$ are $\frac{1}{2}$, we obtain for the average values of $J_\parallel^{(i)}$ and $J_\perp^{(i)}$ the relations

$$\overline{J}_\parallel^{(i)} = \overline{J}_\perp^{(i)} = \tfrac{1}{2}J^{(i)}. \tag{40}$$

For the reflected light, however, the two components will in general differ from each other. For, using (40), one has

$$\left.\begin{aligned} \overline{J}_\parallel^{(r)} &= \frac{1}{2}\frac{\overline{J}_\parallel^{(r)}}{\overline{J}_\parallel^{(i)}}J^{(i)} = \tfrac{1}{2}\mathcal{R}_\parallel J^{(i)}, \\ \overline{J}_\perp^{(r)} &= \frac{1}{2}\frac{\overline{J}_\perp^{(r)}}{\overline{J}_\perp^{(i)}}J^{(i)} = \tfrac{1}{2}\mathcal{R}_\perp J^{(i)}. \end{aligned}\right\} \tag{41}$$

The reflected light is then said to be partially polarized and its *degree of polarization P* may be defined to be*

$$P = \left|\frac{\mathcal{R}_\parallel - \mathcal{R}_\perp}{\mathcal{R}_\parallel + \mathcal{R}_\perp}\right|. \tag{42}$$

* A more general definition of the *degree of polarization* is given in §10.9.2, where its physical significance is also discussed.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

The reflectivity $\overline{\mathcal{R}}$ is now given by

$$\overline{\mathcal{R}} = \frac{\overline{\mathcal{J}}^{(r)}}{\overline{\mathcal{J}}^{(i)}} = \frac{\overline{\mathcal{J}}_{\parallel}^{(r)} + \overline{\mathcal{J}}_{\perp}^{(r)}}{\overline{\mathcal{J}}^{(i)}} = \tfrac{1}{2}(\mathcal{R}_{\parallel} + \mathcal{R}_{\perp}) \tag{43}$$

and is therefore again represented by the curve (b) in Fig. 1.12. The degree of polarization may now be expressed in the form

$$P = \frac{1}{\overline{\mathcal{R}}}\tfrac{1}{2}\{|\mathcal{R}_{\parallel} - \mathcal{R}_{\perp}|\};$$

the quantity in the brace brackets is sometimes called the *polarized proportion* of the light reflected.

Similar results hold for the transmitted light. We also have for natural light

$$\overline{\mathcal{R}} + \overline{\mathcal{T}} = 1. \tag{44}$$

Returning to the case where the incident light is linearly polarized we see that this is true also for the reflected and for the transmitted light, since the phases change only by 0 or π. The directions of vibrations in the reflected and the transmitted light are, however, turned in opposite directions with respect to the incident light. This can be seen from the following:

The angle which we denoted by α, i.e. the angle between the plane of vibration and the plane of incidence, may be called *the azimuth* of the vibration, and we shall regard it as positive when the plane of vibration turns clockwise around the direction of propagation (Fig. 1.13). It may be assumed that the azimuthal angle is in the range $-\pi/2$ to $\pi/2$. We have for the incident, reflected and transmitted electric wave

$$\tan\alpha_i = \frac{A_{\perp}}{A_{\parallel}}, \qquad \tan\alpha_r = \frac{R_{\perp}}{R_{\parallel}}, \qquad \tan\alpha_t = \frac{T_{\perp}}{T_{\parallel}}. \tag{45}$$

Using the Fresnel formulae (20) and (21),

$$\tan\alpha_r = -\frac{\cos(\theta_i - \theta_t)}{\cos(\theta_i + \theta_t)}\tan\alpha_i, \tag{46}$$

$$\tan\alpha_t = \cos(\theta_i - \theta_t)\tan\alpha_i. \tag{47}$$

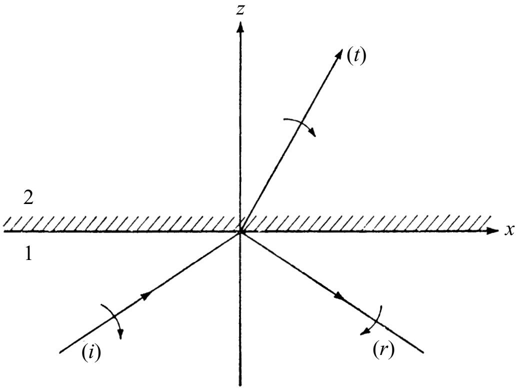

Fig. 1.13 Illustrating the signs of the azimuthal angles.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Since $0 \leqslant \theta_i \leqslant \pi/2$, $0 < \theta_t < \pi/2$,

$$|\tan\alpha_r| \geqslant |\tan\alpha_i|, \tag{48}$$

$$|\tan\alpha_t| \leqslant |\tan\alpha_i|. \tag{49}$$

In (48) the equality sign holds only for normal or tangential incidence ($\theta_i = \theta_t = 0$ or $\theta_i = \pi/2$); in (49) it holds only for normal incidence. The two inequalities imply that on reflection the plane of vibration is turned away from the plane of incidence, whereas on refraction it is turned towards it. In Fig. 1.14 the behaviour of α_r and α_t is illustrated for $n = 1.52$ and $\alpha_i = 45°$. We see that, when θ_i is equal to the polarizing angle $56°40'$, $\alpha_r = 90°$. In fact, according to (46), $\tan\alpha_r = \infty$ (i.e. $\alpha_r = \pi/2$), for $\theta_i + \theta_t = \pi/2$, whatever the value of α_i may be.

It follows from Brewster's law that polarized light may be produced simply by allowing reflection to take place at the polarizing angle. One of the oldest instruments based on this principle is the so-called *Nörrenberg's Reflecting Polariscope* (after Nörrenberg, 1787–1862). It consists essentially of two glass plates (Fig. 1.15) which

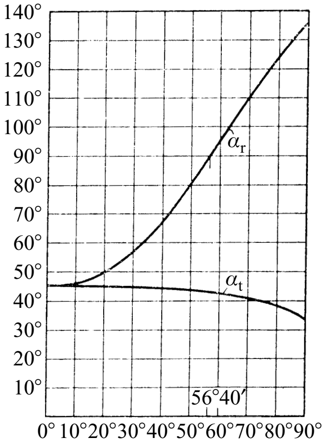

Fig. 1.14 Azimuthal angles as functions of the angle of incidence. [After O. D. Chwolson, *Lehrb. d. Physik* (Braunschweig, Vieweg), Bd. 2. 2 Aufl. (1922), p. 716.]

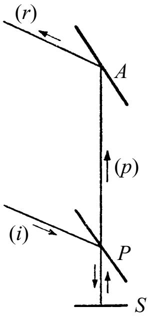

Fig. 1.15 Principle of the *Nörrenberg Reflecting Polariscope:* P = polarizing glass plate; S = reflecting mirror; A = analyser; i = incident beam; p = polarized beam; r = beam reflected at A.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

the rays meet at the polarizing angle. The first plate plays the role of a *polarizer*, i.e. of an instrument producing linearly polarized light from unpolarized light; the second plays the role of an *analyzer*, i.e. an arrangement which detects linearly polarized light. This instrument has, however, several disadvantages, chiefly that the fraction of the light reflected at the polarizing angle is comparatively small, and the path of the ray through the instrument is rather complicated. It is preferable to employ instruments which polarize the incident light without changing its direction of propagation. This can be done, for example, by using a *pile of thin plane-parallel plates*. If a beam of unpolarized light is incident on the pile, it is partially polarized on each refraction, and one can achieve a reasonably high degree of polarization even with a small number of plates. After passing through both boundaries of a plate the intensities of the two components will be in the ratio

$$\left(\frac{\mathcal{T}_\perp}{\mathcal{T}_\parallel}\right)^2 = \cos^4(\theta_i - \theta_t) < 1, \tag{50}$$

a result obtained by applying (35) twice in succession. This shows that, on emerging from the plate, the parallel component is stronger than the perpendicular component, the degree of polarization being greater, the greater θ_i is. If θ_i is equal to the polarizing angle, $\theta_i + \theta_t = \pi/2$, $\tan\theta_i = n$, and we then have

$$\left(\frac{\mathcal{T}_\perp}{\mathcal{T}_\parallel}\right)^2 = \sin^4 2\theta_i = \left(\frac{2n}{1+n^2}\right)^4. \tag{51}$$

For $n = 1.5$ this expression has the value 0.73. Hence, if the light passes through five plates, for example, we obtain the ratio $0.73^5 \simeq 0.2$.

In the past, polarized light was as a rule produced by double refraction in crystals of calcite or quartz, as described in §15.4.1. Today the most convenient method is by the use of so-called *polaroid* films. Their action is based on a property known as *dichroism*. By this we mean the property shown by certain materials of having different absorption coefficients for light polarized in different directions. Polyvinyl alcohol films impregnated by iodine, for example, can be made which transmit nearly 80 per cent of light polarized in one plane, and less than 1 per cent of light polarized at right angles to this plane. The theory of this effect will be briefly discussed in §15.6.3.

1.5.4 Total reflection

So far we have excluded the case when the law of refraction

$$\sin\theta_t = \frac{\sin\theta_i}{n_{12}} \tag{52}$$

does not give a real value for the angle of refraction θ_t. We shall now examine this case. It occurs when light is propagated from an optically denser medium into one which is optically less dense, i.e. when

$$n_{12} = \frac{n_2}{n_1} = \sqrt{\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}} < 1,$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

provided that the angle of incidence θ_i exceeds the critical value $\overline{\theta}_i$ given by

$$\sin\overline{\theta}_i = n_{12}. \tag{53}$$

When $\theta_i = \overline{\theta}_i$, $\sin\theta_t = 1$, i.e. $\theta_t = 90°$, so that the light emerges in a direction tangent to the boundary. If θ_i exceeds the limiting value $\overline{\theta}_i$, no light enters the second medium. All the incident light is then reflected back into the first medium and we speak of *total reflection*.

Nevertheless the electromagnetic field in the second medium does not disappear, only there is no longer a flow of energy across the boundary. For, if (omitting the subscript 12 on n_{12}) we set

$$\sin\theta_t = \frac{\sin\theta_i}{n}, \qquad \cos\theta_t = \pm\mathrm{i}\sqrt{\frac{\sin^2\theta_i}{n^2} - 1}, \tag{54}$$

in the phase factor (15) of the transmitted wave, we have

$$\mathrm{e}^{-\mathrm{i}\tau_t} = \mathrm{e}^{-\mathrm{i}\omega\left(t - \frac{x\,\sin\theta_i}{nv_2}\right)}\,\mathrm{e}^{\mp\frac{\omega z}{v_2}\sqrt{\frac{\sin^2\theta_i}{n^2}-1}}. \tag{55}$$

Eq. (55) represents an inhomogeneous wave which is propagated along the boundary in the plane of incidence (i.e. in the x direction), and which varies exponentially with the distance z from the boundary. Naturally only the negative sign in front of the square root in (55) corresponds to the physical situation, since otherwise the amplitude would tend to infinity with increasing distance. The amplitude is seen to decrease very rapidly with the depth z of penetration, the effective depth of penetration being of the order of $v_2/\omega = \lambda_2/2\pi$, i.e. of the order of a wavelength. The wave is not transversal since, as will be shown below, the component of the electric vector in the direction of propagation does not vanish.

Experimental verification of the disturbance in the second (less dense) medium is somewhat troublesome, since any arrangement used for its detection will perturb the boundary conditions. A rough confirmation may be obtained by bringing up a second refracting medium within about a quarter of a wavelength of the interface at which total reflection is taking place, and observing the entry of the radiation into the second medium.*

To apply the Fresnel formulae (21a) to the case of total reflection, we rewrite them in the form

$$\left.\begin{aligned}
R_\parallel &= \frac{\sin\theta_i\cos\theta_i - \sin\theta_t\cos\theta_t}{\sin\theta_i\cos\theta_i + \sin\theta_t\cos\theta_t}A_\parallel,\\
R_\perp &= -\frac{\sin\theta_i\cos\theta_t - \sin\theta_t\cos\theta_i}{\sin\theta_i\cos\theta_t + \sin\theta_t\cos\theta_i}A_\perp,
\end{aligned}\right\} \tag{56}$$

and substitute into these expressions from (54), remembering that the upper sign is to be taken in front of the square root. We then obtain

* An elegant way of doing this was described by W. Culshaw and D. S. Jones, *Proc. Phys. Soc.*, B, **66** (1953), 859, using electrically generated waves of wavelength 1.25 cm.

$$\left.\begin{aligned}R_\parallel &= \frac{n^2\cos\theta_i - \mathrm{i}\sqrt{\sin^2\theta_i - n^2}}{n^2\cos\theta_i + \mathrm{i}\sqrt{\sin^2\theta_i - n^2}}A_\parallel,\\ R_\perp &= \frac{\cos\theta_i - \mathrm{i}\sqrt{\sin^2\theta_i - n^2}}{\cos\theta_i + \mathrm{i}\sqrt{\sin^2\theta_i - n^2}}A_\perp.\end{aligned}\right\} \tag{57}$$

Hence

$$|R_\parallel| = |A_\parallel|, \qquad |R_\perp| = |A_\perp|, \tag{58}$$

i.e. for each component, the intensity of the light which is totally reflected is equal to the intensity of the incident light.

Although there is a field in the second medium, it is easy to see that no energy flows across the boundary. More precisely it will be shown that, although the component of the Poynting vector in the direction normal to the boundary is in general finite, its time average vanishes; this implies that the energy flows to and fro, but that there is no lasting flow into the second medium.

We write down the x- and y-components of the transmitted field, for $z = 0$, and make use of (54). (Real expressions must now be used for **E** and **H** since the energy flow is a quadratic function of the components.) Denoting the conjugate of a complex quantity by an asterisk, we have from (14),

$$\begin{aligned}E_x^{(t)} &= -\tfrac{1}{2}(T_\parallel\cos\theta_t \mathrm{e}^{-\mathrm{i}\tau_t^o} + T_\parallel^\star\cos^\star\theta_t\mathrm{e}^{+\mathrm{i}\tau_t^o})\\ &= -\frac{\mathrm{i}}{2}\sqrt{\frac{\sin^2\theta_i}{n^2} - 1}\left(T_\parallel\mathrm{e}^{-\mathrm{i}\tau_t^o} - T_\parallel^\star\mathrm{e}^{+\mathrm{i}\tau_t^o}\right),\\ E_y^{(t)} &= \tfrac{1}{2}(T_\perp\mathrm{e}^{-\mathrm{i}\tau_t^o} + T_\perp^\star\mathrm{e}^{+\mathrm{i}\tau_t^o}),\\ H_x^{(t)} &= -\tfrac{1}{2}(T_\perp\cos\theta_t\sqrt{\varepsilon_2}\mathrm{e}^{-\mathrm{i}\tau_t^o} + T_\perp^\star\cos^\star\theta_t\sqrt{\varepsilon_2}\mathrm{e}^{+\mathrm{i}\tau_t^o})\\ &= -\frac{i}{2}\sqrt{\varepsilon_2}\sqrt{\frac{\sin^2\theta_i}{n^2} - 1}\left(T_\perp\mathrm{e}^{-\mathrm{i}\tau_t^o} - T_\perp^\star\mathrm{e}^{+\mathrm{i}\tau_t^o}\right),\\ H_y^{(t)} &= -\tfrac{1}{2}\sqrt{\varepsilon_2}(T_\parallel\mathrm{e}^{-\mathrm{i}\tau_t^o} + T_\parallel^\star\mathrm{e}^{+\mathrm{i}\tau_t^o}),\end{aligned}$$

where

$$\tau_t^o = \omega\left(t - \frac{x\sin\theta_i}{nv_2}\right).$$

If we now form the time average of

$$S_z^{(t)} = \frac{c}{4\pi}(E_x^{(t)}H_y^{(t)} - E_y^{(t)}H_x^{(t)})$$

over an interval $-t' \leqslant t \leqslant t'$, where t' is large compared with the period $\mathrm{T} = 2\pi/\omega$, then both terms disappear for $z = 0$; for one contains the factor

$$\frac{1}{2t'}\int_{-t'}^{t'}(T_\parallel^2\mathrm{e}^{-2\mathrm{i}\tau_t^o} - T_\parallel^{\star 2}\mathrm{e}^{+2\mathrm{i}\tau_t^o})\mathrm{d}t = \left(T_\parallel^2\mathrm{e}^{+\frac{2\mathrm{i}\omega x\sin\theta_i}{nv_2}} - T_\parallel^{\star 2}\mathrm{e}^{-\frac{2\mathrm{i}\omega x\sin\theta_i}{nv_2}}\right)O\left(\frac{\mathrm{T}}{t'}\right),$$

(O denoting the order symbol) which is negligibly small when $t' \gg \mathrm{T}$; the other contains a similar factor with $T_\perp$ in place of $T_\parallel$.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

On the other hand, if the other two components of $\mathbf{S}^{(t)}$ for $z = 0$ are calculated, namely $S_x^{(t)}$ and $S_y^{(t)}$, their time averaged values are in general found to be finite. Energy therefore does not penetrate into the second medium, but flows along the boundary in the plane of incidence.

The preceding analysis applies to a stationary state, and is based on the assumption that the boundary surface and the wave-fronts are of infinite extent. It does not explain how the energy initially entered the second medium. In an actual experiment, the incident wave will be bounded both in space and time*; at the beginning of the process a small amount of energy will penetrate into the second medium and will give rise to a field there.

Finally, we determine the changes in the phases of the components of the reflected and the incident wave. On account of (58) we may set

$$\frac{R_\parallel}{A_\parallel} = \mathrm{e}^{\mathrm{i}\delta_\parallel}, \qquad \frac{R_\perp}{A_\perp} = \mathrm{e}^{\mathrm{i}\delta_\perp}. \tag{59}$$

Now according to (57), $R_\parallel/A_\parallel$ and $R_\perp/A_\perp$ are each of the form $z(z^\star)^{-1}$. Hence if α is the argument of z (i.e. $z = a\mathrm{e}^{\mathrm{i}\alpha}$, with a, α both real), then

$$\mathrm{e}^{\mathrm{i}\delta} = z(z^\star)^{-1} = \mathrm{e}^{2\mathrm{i}\alpha}, \qquad \text{i.e.} \quad \tan\frac{\delta}{2} = \tan\alpha,$$

and therefore

$$\left.\begin{aligned} \tan\frac{\delta_\parallel}{2} &= -\frac{\sqrt{\sin^2\theta_i - n^2}}{n^2\cos\theta_i}, \\ \tan\frac{\delta_\perp}{2} &= -\frac{\sqrt{\sin^2\theta_i - n^2}}{\cos\theta_i}. \end{aligned}\right\} \tag{60}$$

The two components are seen to undergo phase jumps of different amounts. Linearly polarized light will in consequence become elliptically polarized on total reflection.

One can also immediately write down an expression for the relative phase difference $\delta = \delta_\perp - \delta_\parallel$:

$$\tan\frac{\delta}{2} = \frac{\tan\dfrac{\delta_\perp}{2} - \tan\dfrac{\delta_\parallel}{2}}{1 + \tan\dfrac{\delta_\perp}{2}\tan\dfrac{\delta_\parallel}{2}} = \frac{\left(\dfrac{1}{n^2} - 1\right)\dfrac{\sqrt{\sin^2\theta_i - n^2}}{\cos\theta_i}}{1 + \dfrac{\sin^2\theta_i - n^2}{n^2\cos^2\theta_i}},$$

i.e.

$$\tan\frac{\delta}{2} = \frac{\cos\theta_i\sqrt{\sin^2\theta_i - n^2}}{\sin^2\theta_i}. \tag{61}$$

This expression vanishes for grazing incidence ($\theta_i = \pi/2$), and for incidence at the critical angle $\overline{\theta}_i$ ($\sin\overline{\theta}_i = n$). Between these two values there lies the maximum value of the relative phase difference; it is determined from the equation

* The total reflection of a beam of light of finite cross-section has become of considerable interest in recent years in connection with the so-called Goos–Hänchen effect. For a review of some of the pertinent literature, see H. K. V. Lotsch, *Optik*, **32** (1970), 116, 189, 299, 553.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i}(\tan\delta/2) = \frac{2n^2 - (1+n^2)\sin^2\theta_i}{\sin^3\theta_i\sqrt{\sin^2\theta_i - n^2}} = 0.$$

This is satisfied when

$$\sin^2\theta_i = \frac{2n^2}{1+n^2}. \tag{62}$$

On substituting from (62) into (61), we obtain for the maximum δ_m of the relative phase difference δ, the expression

$$\tan\frac{\delta_m}{2} = \frac{1-n^2}{2n}. \tag{63}$$

From (61) it is seen that, with n given, there are two values of the angle θ_i of incidence for every value of δ.

The phase change which takes place on total reflection may be used (as shown already by Fresnel) to produce circularly polarized light from light which is linearly polarized. The amplitude components of the incident light are made equal ($|A_\parallel| = |A_\perp|$) by taking the incident wave to be polarized in a direction which makes an angle of 45° with the normal to the plane of incidence (i.e. $\alpha_i = 45°$). Then, by (58), $|R_\parallel| = |R_\perp|$. Further, n and θ_i are chosen in such a way that the relative phase difference δ is equal to 90°. To attain this value of δ by a single reflection, it would be necessary, according to (63), that

$$\tan\frac{\pi}{4} = 1 < \frac{1-n^2}{2n},$$

i.e.

$$n = n_{12} < \sqrt{2} - 1 = 0.414.$$

This means that the refractive index $n_{21} = 1/n$ of the denser with respect to the less dense medium would have to be at least 2.41. This value is rather large, although there are nonabsorbing substances whose refractive index comes close to, and even exceeds, this value. Fresnel made use of two total reflections on glass. When $n_{21} = 1.51$, one obtains, according to (62) and (63), a maximum relative phase difference $\delta_m = 45°\,56'$ when the angle of incidence θ_i equals $51°\,20'$. It is therefore just possible to attain the value $\delta = 45°$, namely, with either of the following angles of incidence:

$$\theta_i = 48°\,37', \qquad \theta_i = 54°\,37'.$$

A phase difference of 90° may therefore be obtained by means of two successive total reflections at either of these angles. For this purpose a glass block is used, of the form shown in Fig. 1.16, known as *Fresnel's rhomb.*

Fresnel's rhomb may, of course, be also used to produce elliptically polarized light; the azimuth of the incident (linearly polarized) light must then be taken different from 45°. One may also invert the procedure and produce, by means of Fresnel's rhomb, linearly polarized light from elliptically polarized light.

Measurement of the critical angle $\overline{\theta}_i$ gives a convenient and accurate way of determining the index of refraction $n = \sin\overline{\theta}_i$. Instruments used for this purpose are called *refractometers*.

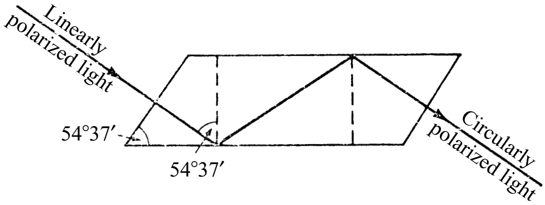

Fig. 1.16 Fresnel's rhomb.

1.6 Wave propagation in a stratified medium. Theory of dielectric films

A medium whose properties are constant throughout each plane perpendicular to a fixed direction is called a *stratified medium*. If the z-axis of a Cartesian reference system is taken along this special direction, then

$$\varepsilon = \varepsilon(z), \qquad \mu = \mu(z). \tag{1}$$

We shall consider the propagation of a plane, time-harmonic electromagnetic wave through such a medium; this is a natural generalization of the simple case treated in the previous section.

The theory of stratified media is of considerable importance in optics, in connection with *multilayers*, i.e. a succession of thin plane-parallel films. Such films may be produced with the help of high-vacuum evaporation techniques, and their thickness may be controlled with very high accuracy. They have many useful applications. For example, as will be demonstrated later, they may be employed as *antireflection films*, i.e. as coatings which reduce the reflectivity of a given surface. On the other hand thin films will, under appropriate conditions, *enhance* the reflectivity so that when deposited on a glass surface they may be used as beam-splitters, i.e. arrangements employed in interferometry for the division of an incident beam into two parts. Under appropriate conditions a multilayer may also be employed as a filter which transmits (or reflects) only selected regions of the spectrum. Multilayers may also be used as polarizers.

The subject of dielectric and metallic films has been very extensively treated in the scientific literature and many schemes for the computation of the optical effects of multilayers have been proposed. We shall give an outline of the general theory as developed in elegant and important investigations by F. Abelès,[*] and consider in detail some special cases of particular interest. For the treatment of problems involving only a small number of films it is naturally not necessary to use the general theory, and accordingly we shall later (§7.6) describe an alternative and older method based on the concept of multiple reflections.

Only dielectric stratified media will be treated in this section. The extension of the analysis to conducting media will be described in §14.4.

[*] F. Abelès, *Ann. de Physique*, **5** (1950), 596–640 and 706–782. For a detailed treatment of the subject of thin films see a more specialized treatise e.g. H. Mayer, *Physik dünner Schichten* (Stuttgart, Wissenschaftliche Verlagsgessellschaft, 1950); S. Methfessel, *Dünne Schichten* (Halle (Saale), VEB Wilhelm Knapp Verlag, 1953); or O. S. Heavens, *Optical Properties of Thin Solid Films* (London, Butterworths Scientific Publications, 1955).

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

1.6.1 The basic differential equations

Consider a plane, time-harmonic electromagnetic wave propagated through a stratified medium. In the special case when the wave is linearly polarized with its electric vector perpendicular to the plane of incidence we shall speak of a *transverse electric wave* (denoted by *TE*); when it is linearly polarized with its magnetic vector perpendicular to the plane of incidence we shall speak of a *transverse magnetic wave* (denoted by *TM*).* Any arbitrarily polarized plane wave may be resolved into two waves, one of which is a *TE* wave and the other a *TM* wave. Since according to §1.5 the boundary conditions at a discontinuity surface for the perpendicular and parallel components are independent of each other, these two waves will also be mutually independent. Moreover, Maxwell's equations remain unchanged when **E** and **H** and simultaneously ε and $-\mu$ are interchanged. Thus any theorem relating to *TM* waves may immediately be deduced from the corresponding result for *TE* waves by making this change. It will, therefore, be sufficient to study in detail the *TE* waves only.

We take the plane of incidence to be the y, z-plane,† z being the direction of stratification. For a *TE* wave, $E_y = E_z = 0$ and Maxwell's equations reduce to the following six scalar equations (time dependence $\exp(-\mathrm{i}\omega t)$ being assumed):

$$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} + \frac{\mathrm{i}\varepsilon\omega}{c} E_x = 0, \qquad (1a) \qquad\qquad \frac{\mathrm{i}\omega\mu}{c} H_x = 0, \qquad (2a)$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = 0, \qquad (1b) \qquad\qquad \frac{\partial E_x}{\partial z} - \frac{\mathrm{i}\omega\mu}{c} H_y = 0, \qquad (2b)$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = 0, \qquad (1c) \qquad\qquad \frac{\partial E_x}{\partial y} + \frac{\mathrm{i}\omega\mu}{c} H_z = 0. \qquad (2c)$$

These equations show that H_y, H_z and E_x are functions of y and z only. Eliminating H_y and H_z between (1a), (2b) and (2c) (or by taking the x-component of the wave equation §1.2 (5) for E) it follows that

$$\frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} + n^2 k_0{}^2 E_x = \frac{\mathrm{d}(\ln\mu)}{\mathrm{d}z}\frac{\partial E_x}{\partial z}, \qquad (3)$$

where

$$n^2 = \varepsilon\mu, \qquad k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda_0}. \qquad (4)$$

To solve (3) we take, as a trial solution, a product of two functions, one involving y only and the other involving z only:

$$E_x(y, z) = Y(y)U(z). \qquad (5)$$

Eq. (3) then becomes

[*] The terms 'E-polarized' and 'H-polarized' are also used (see §11.4.1). It should be mentioned that the terms 'transverse electric wave' and 'transverse magnetic wave' have different meanings in the theory of wave guides.

[†] Not the x, z-plane as in the previous section.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\frac{1}{Y}\frac{\mathrm{d}^2 Y}{\mathrm{d}y^2} = -\frac{1}{U}\frac{\mathrm{d}^2 U}{\mathrm{d}z^2} - n^2 k_0^2 + \frac{\mathrm{d}(\ln\mu)}{\mathrm{d}z}\frac{1}{U}\frac{\mathrm{d}U}{\mathrm{d}z}. \tag{6}$$

Now the term on the left is a function of y only whilst the terms on the right depend only on z. Hence (6) can only hold if each side is equal to a constant ($-K^2$ say):

$$\frac{1}{Y}\frac{\mathrm{d}^2 Y}{\mathrm{d}y^2} = -K^2, \tag{7}$$

$$\frac{\mathrm{d}^2 U}{\mathrm{d}z^2} - \frac{\mathrm{d}(\ln\mu)}{\mathrm{d}z}\frac{\mathrm{d}U}{\mathrm{d}z} + n^2 k_0^2 U = K^2 U. \tag{8}$$

It will be convenient to set

$$K^2 = k_0^2\alpha^2. \tag{9}$$

Then (7) gives

$$Y = [\text{constant}]\,\mathrm{e}^{\mathrm{i}k_0\alpha y},$$

and consequently E_x is of the form

$$E_x = U(z)\mathrm{e}^{\mathrm{i}(k_0\alpha y-\omega t)}, \tag{10}$$

where $U(z)$ is a (possibly complex) function of z. From (2b) and (2c) we see that H_y and H_z are given by expressions of the same form:

$$H_y = V(z)\mathrm{e}^{\mathrm{i}(k_0\alpha y-\omega t)}, \tag{11}$$

$$H_z = W(z)\mathrm{e}^{\mathrm{i}(k_0\alpha y-\omega t)}. \tag{12}$$

On account of (1a), (2b) and (2c), the amplitude functions U, V and W are related by the following equations:

$$V' = \mathrm{i}k_0(\alpha W + \varepsilon U), \tag{13a}$$

$$U' = \mathrm{i}k_0\mu V, \tag{13b}$$

$$\alpha U + \mu W = 0, \tag{13c}$$

the prime denoting differentiation with respect to z. Substituting for W from (13c) into (13a) we have, together with (13b), a pair of simultaneous first-order differential equations* for U and V:

$$\left.\begin{aligned} U' &= \mathrm{i}k_0\mu V,\\ V' &= \mathrm{i}k_0\left(\varepsilon - \frac{\alpha^2}{\mu}\right)U. \end{aligned}\right\} \tag{14}$$

[*] Eqs. (14) are of the same form as the equations of an electric transmission line, i.e.

$$\frac{\mathrm{d}V}{\mathrm{d}z} = -ZI, \qquad \frac{\mathrm{d}I}{\mathrm{d}z} = -YV,$$

where V is the voltage across the line, I is the current in the line, Z is the series impedance, and Y the shunt admittance. The theory of stratified media may therefore be developed in a strict analogy with the theory of electric transmission lines, as has been done by several authors, for example, R. B. Muchmore, *J. Opt. Soc. Amer.*, **38** (1948), 20; K. Schuster, *Ann. d. Physik* (6), **4** (1949), 352; R. Kronig, R. S. Blaisse and J. J. v. d. Sande, *J. Appl. Sci. Res.*, **B**, **1** (1947), 63.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

Elimination between these equations finally gives the following second-order linear differential equations for U and V:

$$\frac{d^2 U}{dz^2} - \frac{d(\ln \mu)}{dz}\frac{dU}{dz} + k_0^2(n^2 - \alpha^2)U = 0, \tag{15}$$

$$\frac{d^2 V}{dz^2} - \frac{d\left[\ln\left(\varepsilon - \dfrac{\alpha^2}{\mu}\right)\right]}{dz}\frac{dV}{dz} + k_0^2(n^2 - \alpha^2)V = 0. \tag{16}$$

According to the substitution rule which is a consequence of the symmetry of Maxwell's equations, it immediately follows that *for the TM wave* ($H_y = H_z = 0$), the nonvanishing components of the field vectors are of the form:

$$H_x = U(z)e^{i(k_0 \alpha y - \omega t)}, \tag{17}$$

$$E_y = -V(z)e^{i(k_0 \alpha y - \omega t)}, \tag{18}$$

$$E_z = -W(z)e^{i(k_0 \alpha y - \omega t)}, \tag{19}$$

where

$$\left.\begin{aligned} U' &= ik_0 \varepsilon V, \\ V' &= ik_0\left(\mu - \frac{\alpha^2}{\varepsilon}\right)U, \end{aligned}\right\} \tag{20}$$

and W is related to U by means of the equation

$$\alpha U + \varepsilon W = 0. \tag{21}$$

U and V now satisfy the following second-order linear differential equations:

$$\frac{d^2 U}{dz^2} - \frac{d(\ln \varepsilon)}{dz}\frac{dU}{dz} + k_0^2(n^2 - \alpha^2)U = 0, \tag{22}$$

$$\frac{d^2 V}{dz^2} - \frac{d\left[\ln\left(\mu - \dfrac{\alpha^2}{\varepsilon}\right)\right]}{dz}\frac{dV}{dz} + k_0^2(n^2 - \alpha^2)V = 0. \tag{23}$$

U, V and W are in general complex functions of z. The surfaces of constant amplitude of E_x are given by

$$|U(z)| = \text{constant},$$

whilst the surfaces of constant phase have the equation

$$\phi(z) + k_0 \alpha y = \text{constant},$$

where $\phi(z)$ is the phase of U. The two sets of surfaces do not in general coincide so that E_x (and similarly H_y and H_z) is an inhomogeneous wave. For a small displacement (dy, dz) along a cophasal surface, $\phi'(z)dz + k_0 \alpha dy = 0$; hence if θ denotes the angle which the normal to the co-phasal surface makes with OZ, then

$$\tan\theta = -\frac{dz}{dy} = \frac{k_0 \alpha}{\phi'(z)}.$$

In the special case when the wave is a homogeneous plane wave,

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\phi(z) = k_0 nz \cos\theta, \qquad \alpha = n \sin\theta. \tag{24}$$

Hence the relation

$$\alpha = \text{constant}$$

imposed by (9) may be regarded as a generalization of *Snell's law of refraction* to stratified media.

1.6.2 The characteristic matrix of a stratified medium

The solutions, subject to appropriate boundary conditions, of the differential equations which we have just derived, and various theorems relating to stratified media, can most conveniently be expressed in terms of matrices. We shall therefore give a brief account of the main definitions relating to matrices before discussing the consequences of our equations.

I. By a matrix one understands a system of real or complex numbers, arranged in a rectangular or a square array:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

a_{ij} denoting the element in the ith row and the jth column. The matrix is denoted symbolically by $\boldsymbol{A}$ or $[a_{ij}]$, and is said to be an m by n matrix (or $m \times n$ matrix), since it contains m rows and n columns. In the special case when $m = n$, $\boldsymbol{A}$ is said to be a *square matrix* of order m. If $\boldsymbol{A}$ is a square matrix, the determinant whose elements are the same, and are in the same positions as the elements of $\boldsymbol{A}$, is said to be the *determinant of the matrix* $\boldsymbol{A}$; it is denoted by $|\boldsymbol{A}|$ or $|a_{ij}|$. If $|\boldsymbol{A}| = 1$, $\boldsymbol{A}$ is said to be *unimodular.*

By definition two matrices are *equal* only if they have the same number of rows (m) and the same number of columns (n), and if their corresponding elements are equal. If $\boldsymbol{A} = [a_{ij}]$ and $\boldsymbol{B} = [b_{ij}]$ are two matrices with the same number of rows and the same number of columns, then their *sum* $\boldsymbol{A} + \boldsymbol{B}$ is defined as the matrix $\boldsymbol{C}$ whose elements are $c_{ij} = a_{ij} + b_{ij}$. Similarly their *difference* $\boldsymbol{A} - \boldsymbol{B}$ is defined as the matrix $\boldsymbol{D}$ with elements $d_{ij} = a_{ij} - b_{ij}$.

A matrix having every element zero is called a *null matrix*. The square matrix with elements $a_{ij} = 0$ when $i \neq j$ and $a_{ii} = 1$ for every value of i is called *unit matrix* and will be denoted by $\boldsymbol{I}$.

The product of a matrix $\boldsymbol{A}$ and a number λ (real or complex) is defined as the matrix $\boldsymbol{B}$ with elements $b_{ij} = \lambda a_{ij}$.

The product $\boldsymbol{AB}$ of two matrices is defined only when the number of columns in $\boldsymbol{A}$ is equal to the number of rows in $\boldsymbol{B}$. If $\boldsymbol{A}$ is an $m \times p$ matrix and $\boldsymbol{B}$ is a $p \times n$ matrix the product is then by definition the $m \times n$ matrix with elements

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}.$$

The process of multiplication of two matrices is thus analogous to the row-by-column rule for multiplication of determinants of equal orders. In general $\boldsymbol{AB} \neq \boldsymbol{BA}$. For example

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

whilst

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

In the special case when $\boldsymbol{AB} = \boldsymbol{BA}$, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to *commute*.

The above definitions and properties of matrices are the only ones necessary for our purposes, and we can, therefore, now return to our discussion of propagation of electromagnetic waves through a stratified medium.

II. Since the functions $U(z)$ and $V(z)$ of §1.6.1 each satisfy a second-order linear differential equation [(15) and (16)], it follows that U and V may each be expressed as a linear combination of two particular solutions, say U_1, U_2 and V_1, V_2. These particular solutions cannot be arbitrary; they must be coupled by the first-order differential equations (14):

$$\left.\begin{aligned} U_1' &= \mathrm{i}k_0 \mu V_1, \\ V_1' &= \mathrm{i}k_0 \left(\varepsilon - \frac{\alpha^2}{\mu} \right) U_1, \end{aligned}\right\} \qquad \left.\begin{aligned} U_2' &= \mathrm{i}k_0 \mu V_2, \\ V_2' &= \mathrm{i}k_0 \left(\varepsilon - \frac{\alpha^2}{\mu} \right) U_2. \end{aligned}\right\} \tag{25}$$

From these relations it follows that

$$V_1 U_2' - U_1' V_2 = 0, \qquad U_1 V_2' - V_1' U_2 = 0,$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}z}(U_1 V_2 - U_2 V_1) = 0.$$

This relation implies that *the determinant*

$$D = \begin{vmatrix} U_1 & V_1 \\ U_2 & V_2 \end{vmatrix} \tag{26}$$

associated with any two arbitrary solutions of (14) is a constant, i.e. that D is an invariant of our system of equations.*

* This also follows from a well-known property of a Wronskian of second-order linear differential equations. Moreover, it may also be shown that, if U_1 is known, U_2 may be obtained by integration from the relation

$$U_2 = \mathrm{i}kDU_1 \int \frac{\mu}{U_1^2} \, \mathrm{d}z.$$

See F. Abelès, *Ann. de Physique*, **5** (1950), 603.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

For our purposes the most convenient choice of the particular solutions is

$$\left.\begin{aligned} U_1 &= f(z), \quad & U_2 &= F(z), \\ V_1 &= g(z), \quad & V_2 &= G(z), \end{aligned}\right\} \tag{27}$$

such that

$$f(0) = G(0) = 0 \quad \text{and} \quad F(0) = g(0) = 1. \tag{28}$$

Then the solutions with

$$U(0) = U_0, \qquad V(0) = V_0, \tag{29}$$

may be expressed in the form

$$\left.\begin{aligned} U &= FU_0 + fV_0, \\ V &= GU_0 + gV_0, \end{aligned}\right\}$$

or, in matrix notation,

$$\boldsymbol{Q} = \boldsymbol{N}\boldsymbol{Q}_0, \tag{30}$$

where

$$\boldsymbol{Q} = \begin{bmatrix} U(z) \\ V(z) \end{bmatrix}, \qquad \boldsymbol{Q}_0 = \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}, \qquad \boldsymbol{N} = \begin{bmatrix} F(z) & f(z) \\ G(z) & g(z) \end{bmatrix}. \tag{31}$$

On account of the relation $D =$ constant, the determinant of the square matrix $\boldsymbol{N}$ is a constant. The value of this constant may immediately be found by taking $z = 0$, giving

$$|\boldsymbol{N}| = Fg - fG = 1.$$

It is usually more convenient to express U_0 and V_0 as functions of $U(z)$ and $V(z)$. Solving for U_0 and V_0, we obtain

$$\boldsymbol{Q}_0 = \boldsymbol{M}\boldsymbol{Q}, \tag{32}$$

where

$$\boldsymbol{M} = \begin{bmatrix} g(z) & -f(z) \\ -G(z) & F(z) \end{bmatrix}. \tag{33}$$

This matrix is also unimodular,

$$|\boldsymbol{M}| = 1. \tag{34}$$

The significance of $\boldsymbol{M}$ is clear: it relates the x- and y-components of the electric (or magnetic) vectors in the plane $z = 0$ to the components in an arbitrary plane $z =$ constant. Now we saw that knowledge of U and V is sufficient for the complete specification of the field. Hence *for the purposes of determining the propagation of a plane monochromatic wave through a stratified medium, the medium only need be specified by an appropriate two by two unimodular matrix* $\boldsymbol{M}$. For this reason we shall call $\boldsymbol{M}$ the *characteristic matrix* of the stratified medium. The constancy of the determinant $|\boldsymbol{M}|$ may be shown to imply the conservation of energy.*

* To show this, one evaluates the reflectivity and transmissivity (51) in terms of the matrix elements. If further one uses the fact that for a nonabsorbing medium the characteristic matrix is of the form indicated by (45), it follows that the conservation law $\mathcal{R} + \mathcal{T} = 1$ will be satisfied provided that $|\boldsymbol{M}| = 1$.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

We shall now consider the form of the characteristic matrix for cases of particular interest.

(a) A homogeneous dielectric film

In this case ε, μ and $n = \sqrt{\varepsilon\mu}$ are constants. If θ denotes the angle which the normal to the wave makes with the z-axis, we have by (24),

$$\alpha = n \sin\theta.$$

For a *TE* wave, we have according to (15) and (16),

$$\left.\begin{aligned} \frac{\mathrm{d}^2 U}{\mathrm{d}z^2} + (k_0^2 n^2 \cos^2\theta) U = 0, \\ \frac{\mathrm{d}^2 V}{\mathrm{d}z^2} + (k_0^2 n^2 \cos^2\theta) V = 0. \end{aligned}\right\} \tag{35}$$

The solutions of these equations, subject to the relations (14), are easily seen to be

$$\left.\begin{aligned} U(z) &= A\cos(k_0 nz\cos\theta) + B\sin(k_0 nz\cos\theta), \\ V(z) &= \frac{1}{\mathrm{i}}\sqrt{\frac{\varepsilon}{\mu}}\cos\theta[B\cos(k_0 nz\cos\theta) - A\sin(k_0 nz\cos\theta)]. \end{aligned}\right\} \tag{36}$$

Hence the particular solutions (27) which satisfy the boundary conditions (28) are

$$\left.\begin{aligned} U_1 &= f(z) = \frac{\mathrm{i}}{\cos\theta}\sqrt{\frac{\mu}{\varepsilon}}\sin(k_0 nz\cos\theta), \\ V_1 &= g(z) = \cos(k_0 nz\cos\theta), \\ U_2 &= F(z) = \cos(k_0 nz\cos\theta), \\ V_2 &= G(z) = \mathrm{i}\sqrt{\frac{\varepsilon}{\mu}}\cos\theta\sin(k_0 nz\cos\theta). \end{aligned}\right\} \tag{37}$$

If we set

$$p = \sqrt{\frac{\varepsilon}{\mu}}\cos\theta, \tag{38}$$

the characteristic matrix is seen to be

$$\boldsymbol{M}(z) = \begin{bmatrix} \cos(k_0 nz\cos\theta) & -\dfrac{\mathrm{i}}{p}\sin(k_0 nz\cos\theta) \\ -\mathrm{i}p\sin(k_0 nz\cos\theta) & \cos(k_0 nz\cos\theta) \end{bmatrix}. \tag{39}$$

For a *TM* wave, the same equations hold, with p replaced by

$$q = \sqrt{\frac{\mu}{\varepsilon}}\cos\theta. \tag{40}$$

(b) A stratified medium as a pile of thin homogeneous films

Consider two adjacent stratified media, the first one extending from $z = 0$ to $z = z_1$, and the second from $z = z_1$ to $z = z_2$. If $\boldsymbol{M}_1(z)$ and $\boldsymbol{M}_2(z)$ are the characteristic matrices of the two media, then

$$\boldsymbol{Q}_0 = \boldsymbol{M}_1(z_1)\boldsymbol{Q}(z_1), \qquad \boldsymbol{Q}(z_1) = \boldsymbol{M}_2(z_2 - z_1)\boldsymbol{Q}(z_2),$$

so that

$$\boldsymbol{Q}_0 = \boldsymbol{M}(z_2)\boldsymbol{Q}(z_2),$$

where

$$\boldsymbol{M}(z_2) = \boldsymbol{M}_1(z_1)\boldsymbol{M}_2(z_2 - z_1).$$

This result may immediately be generalized to the case of a succession of stratified media extending from $0 \leqslant z \leqslant z_1$, $z_1 \leqslant z \leqslant z_2$, $\ldots$, $z_{N-1} \leqslant z \leqslant z_N$. If the characteristic matrices are $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\ldots$, $\boldsymbol{M}_N$, then

$$\left.\begin{aligned} &\boldsymbol{Q}_0 = \boldsymbol{M}(z_N)\boldsymbol{Q}(z_N),\\ &\text{where}\\ &\boldsymbol{M}(z_N) = \boldsymbol{M}_1(z_1)\boldsymbol{M}_2(z_2 - z_1)\cdots \boldsymbol{M}_N(z_N - z_{N-1}).\end{aligned}\right\} \tag{41}$$

With the help of (41) an approximate expression for the characteristic matrix of any stratified medium may easily be derived*: we regard the medium as consisting of a very large number of thin films of thickness δz_1, δz_2, δz_3, $\ldots$, δz_n. If the maximum thickness is sufficiently small, it is permissible to regard ε, μ and n to be constant throughout each film. From (39) it is seen that the characteristic matrix of the jth film is then approximately given by

$$\boldsymbol{M}_j = \begin{bmatrix} 1 & -\dfrac{\mathrm{i}}{p_j}k_0 n_j \delta z_j \cos\theta_j \\ -\mathrm{i}p_j k_0 n_j \delta z_j \cos\theta, & 1 \end{bmatrix}.$$

Hence the characteristic matrix of the whole medium, considered as a pile of thin films, is approximately equal to (again retaining terms up to the first power in δz only):

$$\boldsymbol{M} = \prod_{j=1}^{N} \boldsymbol{M}_j = \begin{bmatrix} 1 & -\mathrm{i}k_0 B \\ -\mathrm{i}k_0 A & 1 \end{bmatrix}, \tag{42}$$

where

$$A = \sum_{j=1}^{N} p_j n_j \delta z_j \cos\theta_j = \sum_{j=1}^{N}\left(\varepsilon_j - \frac{\alpha^2}{\mu_j}\right)\delta z_j,$$

$$B = \sum_{j=1}^{N} \frac{n_j}{p_j}\delta z_j \cos\theta_j = \sum_{j=1}^{N} \mu_j \delta z_j.$$

* For a fuller treatment of stratified media of continuously varying refractive index see R. Jacobsson, *Progress in Optics*, Vol. 5, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 247.

Proceeding to the limit as $N \to \infty$ in such a way that $\max|\delta z_j| \to 0$, we obtain

$$\boldsymbol{M} = \begin{bmatrix} 1 & -\mathrm{i}k_0\mathcal{B} \\ -\mathrm{i}k_0\mathcal{A} & 1 \end{bmatrix}, \tag{43}$$

where

$$\mathcal{A} = \int\left(\varepsilon - \frac{\alpha^2}{\mu}\right)\mathrm{d}z, \qquad \mathcal{B} = \int \mu \mathrm{d}z, \tag{44}$$

the integration being taken throughout the whole z range. Eq. (43) gives a first approximation to the characteristic matrix of an arbitrary stratified medium. Improved approximations can be obtained by retaining higher-order terms* in the expansions of $\cos(k_0 n\delta z \cos\theta)$ and $\sin(k_0 n\delta z \cos\theta)$ and in the product (42).

Since, for a nonabsorbing medium, ε and μ are real, it is also seen that *the characteristic matrix of a nonabsorbing stratified medium has the form*

$$\boldsymbol{M} = \begin{bmatrix} a & \mathrm{i}b \\ \mathrm{i}c & d \end{bmatrix}, \tag{45}$$

where a, b, c and d are real.

1.6.3 The reflection and transmission coefficients

Consider a plane wave incident upon a stratified medium that extends from $z = 0$ to $z = z_1$ and that is bounded on each side by a homogeneous, semiinfinite medium. We shall derive expressions for the amplitudes and intensities of the reflected and transmitted waves.†

Let A, R and T denote as before the amplitudes (possibly complex) of the electric vectors of the incident, reflected and transmitted waves. Further, let ε_1, μ_1 and ε_l, μ_l be the dielectric constant and the magnetic permeability of the first and the last medium, and let θ_1 and θ_l be the angles which the normals to the incident and the transmitted waves make with the z direction (direction of stratification).

The boundary conditions of §1.1 demand that the tangential components of **E** and **H** shall be continuous across each of the two boundaries of the stratified medium. This gives, if the relation §1.4 (4)

$$\mathbf{H} = \sqrt{\frac{\varepsilon}{\mu}}\,\mathbf{s} \times \mathbf{E}$$

is also used, the following relations for a *TE* wave:

$$\left.\begin{aligned} U_0 &= A + R, & U(z_1) &= T, \\ V_0 &= p_1(A - R), & V(z_1) &= p_l T, \end{aligned}\right\} \tag{46}$$

where

* This is discussed fully in the paper by F. Abelès, *Ann. d. Physique*, **5** (1950), 611.

† We consider the amplitudes of the electric vectors when studying a *TE* wave and those of the magnetic vectors when studying a *TM* wave.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$p_1 = \sqrt{\frac{\varepsilon_1}{\mu_1}}\cos\theta_1, \qquad p_l = \sqrt{\frac{\varepsilon_l}{\mu_l}}\cos\theta_l. \tag{47}$$

The four quantities U_0, V_0, U and V given by (46) are connected by the basic relation (32); hence

$$\left.\begin{aligned} A + R &= (m'_{11} + m'_{12}p_l)T, \\ p_1(A - R) &= (m'_{21} + m'_{22}p_l)T, \end{aligned}\right\} \tag{48}$$

m'_{ij} being the elements of the characteristic matrix of the medium, evaluated for $z = z_1$.

From (48) we obtain the reflection and transmission coefficients of the film:

$$r = \frac{R}{A} = \frac{(m'_{11} + m'_{12}p_l)p_1 - (m'_{21} + m'_{22}p_l)}{(m'_{11} + m'_{12}p_l)p_1 + (m'_{21} + m'_{22}p_l)}, \tag{49}$$

$$t = \frac{T}{A} = \frac{2p_1}{(m'_{11} + m'_{12}p_l)p_1 + (m'_{21} + m'_{22}p_l)}. \tag{50}$$

In terms of r and t, the *reflectivity* and *transmissivity* are

$$\mathcal{R} = |r|^2, \qquad \mathcal{T} = \frac{p_l}{p_1}|t|^2. \tag{51}$$

The phase δ_r of r may be called *the phase change on reflection* and the phase δ_t of t *the phase change on transmission*. The phase change δ_r is referred to the first surface of discontinuity, whilst the phase change δ_l is referred to the plane boundary between the stratified medium and the last semiinfinite medium.

The corresponding formulae for a *TM* wave are immediately obtained from (49)–(51) on replacing the quantities p_1 and p_l by

$$q_1 = \sqrt{\frac{\mu_1}{\varepsilon_1}}\cos\theta_1, \qquad q_l = \sqrt{\frac{\mu_l}{\varepsilon_l}}\cos\theta_l. \tag{52}$$

r and t are then the ratios of the amplitudes of the magnetic and not the electric vectors.

1.6.4 A homogeneous dielectric film[*]

The properties of an homogeneous dielectric film situated between two homogeneous media are of particular interest in optics, and we shall, therefore, study this case more fully. We assume all the media to be nonmagnetic ($\mu = 1$).

The characteristic matrix of a homogeneous dielectric film is given by (39). Denoting by subscripts 1, 2 and 3 quantities which refer to the three media (see Fig. 1.17), and by h the thickness of the film, we have

$$m'_{11} = m'_{22} = \cos\beta, \qquad m'_{12} = -\frac{\mathrm{i}}{p_2}\sin\beta, \qquad m'_{21} = -\mathrm{i}p_2\sin\beta, \tag{53}$$

[*] An alternative derivation of the main formulae relating to the properties of a single dielectric film will be found in §7.6.1. The formulae may, of course, also be derived directly by applying the boundary conditions of §1.1.3 at each boundary of the film [see M. Born, *Optik* (Berlin, Springer, 1933, reprinted 1965), p. 125; or H. Mayer, *Physik dünner Schichten* (Stuttgart, Wissenschaftliche Verlagsgesellschaft 1950), p. 145.]

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

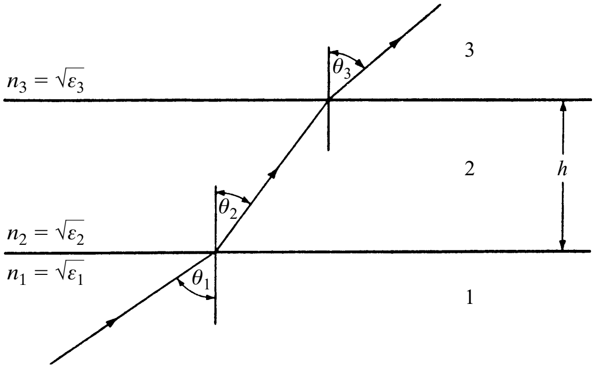
$n_3 = \sqrt{\varepsilon_3}$
θ_3
3
2
h
θ_2
$n_2 = \sqrt{\varepsilon_2}$
$n_1 = \sqrt{\varepsilon_1}$
θ_1
1

Fig. 1.17 Propagation of an electromagnetic wave through a homogeneous film.

where

$$\beta = \frac{2\pi}{\lambda_0} n_2 h \cos\theta_2$$

and

$$p_j = n_j \cos\theta_j \qquad (j = 1, 2, 3). \tag{54}$$

The reflection and transmission coefficients r and t may be obtained by substituting these expressions into (49) and (50), with $l = 3$. The resulting formulae may be conveniently expressed in terms of the corresponding coefficients r_{12}, t_{12} and r_{23}, t_{23} associated with the reflection and transmission at the first and the second surface respectively. According to the Fresnel formulae §1.5 (20) and §1.5 (21) we have for a *TE* wave,

$$r_{12} = \frac{n_1\cos\theta_1 - n_2\cos\theta_2}{n_1\cos\theta_1 + n_2\cos\theta_2} = \frac{p_1 - p_2}{p_1 + p_2}, \tag{55}$$

$$t_{12} = \frac{2n_1\cos\theta_1}{n_1\cos\theta_1 + n_2\cos\theta_2} = \frac{2p_1}{p_1 + p_2}, \tag{56}$$

with analogous expressions for r_{23} and t_{23}. In terms of these expressions, the formulae for r and t become*

$$r = \frac{r_{12} + r_{23}\mathrm{e}^{2\mathrm{i}\beta}}{1 + r_{12}r_{23}\mathrm{e}^{2\mathrm{i}\beta}}, \tag{57}$$

$$t = \frac{t_{12}t_{23}\mathrm{e}^{\mathrm{i}\beta}}{1 + r_{12}r_{23}\mathrm{e}^{2\mathrm{i}\beta}}; \tag{58}$$

the reflectivity and transmissivity are therefore given by

$$\mathcal{R} = |r|^2 = \frac{r_{12}^2 + r_{23}^2 + 2r_{12}r_{23}\cos 2\beta}{1 + r_{12}^2 r_{23}^2 + 2r_{12}r_{23}\cos 2\beta}; \tag{59}$$

* These formulae were first derived in a different manner by G. B. Airy, *Phil. Mag.*, **2** (1833), 20; also *Ann. Phys. und Chem.* (Ed. J. C. Poggendorff), **41** (1837), 512.

and

$$\mathcal{T} = \frac{p_3}{p_1}|t|^2 = \frac{n_3 \cos\theta_3}{n_1 \cos\theta_1} \frac{t_{12}^2 t_{23}^2}{1 + r_{12}^2 r_{23}^2 + 2r_{12}r_{23}\cos 2\beta}. \tag{60}$$

A straightforward calculation gives, as expected,

$$\mathcal{R} + \mathcal{T} = 1.$$

The phase changes can also easily be calculated from (57) and (58), and are found to be given by

$$\tan\delta_r = \tan(\arg r) = \frac{r_{23}(1 - r_{12}^2)\sin 2\beta}{r_{12}(1 + r_{23}^2) + r_{23}(1 + r_{12}^2)\cos 2\beta}, \tag{61}$$

$$\tan\delta_t = \tan(\arg t) = \frac{1 - r_{12}r_{23}}{1 + r_{12}r_{23}}\tan\beta. \tag{62}$$

Let us now briefly consider the implications of these formulae. We first note that (59) and (60) remain unchanged when β is replaced by $\beta + \pi$, i.e. when h is replaced by $h + \Delta h$, where

$$\Delta h = \frac{\lambda_0}{2n_2\cos\theta_2}. \tag{63}$$

Hence *the reflectivity and transmissivity of dielectric films which differ in thickness by an integral multiple of* $\lambda_0/2n_2\cos\theta_2$ *are the same*.

Next we determine the optical thickness for which the reflection coefficient has a maximum or a minimum. If we set

$$H = n_2 h, \tag{64}$$

we find from (59) that

$$\frac{d\mathcal{R}}{dH} = 0 \qquad \text{when} \qquad \sin 2\beta = 0,$$

i.e. when

$$H = \frac{m\lambda_0}{4\cos\theta_2}, \qquad (m = 0, 1, 2, \ldots).$$

We must distinguish two cases:

(1) When *m is odd*, i.e. when H has any of the values

$$H = \frac{\lambda_0}{4\cos\theta_2}, \qquad \frac{3\lambda_0}{4\cos\theta_2}, \qquad \frac{5\lambda_0}{4\cos\theta_2}, \ldots$$

then $\cos 2\beta = -1$ and (59) reduces to

$$\mathcal{R} = \left(\frac{r_{12} - r_{23}}{1 - r_{12}r_{23}}\right)^2. \tag{65}$$

In particular for *normal incidence*, one has from (55)

$$r_{12} = \frac{n_1 - n_2}{n_1 + n_2}, \qquad r_{23} = \frac{n_2 - n_3}{n_2 + n_3}, \tag{66}$$

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

and (65) becomes

$$\mathcal{R} = \left(\frac{n_1 n_3 - n_2^2}{n_1 n_3 + n_2^2}\right)^2 . \tag{67}$$

(2) When *m is even*, i.e. when the optical thickness has any of the values

$$H = \frac{\lambda_0}{2\cos\theta_2}, \qquad \frac{2\lambda_0}{2\cos\theta_2}, \qquad \frac{3\lambda_0}{2\cos\theta_2}, \ldots$$

then $\cos 2\beta = 1$ and (59) reduces to

$$\mathcal{R} = \left(\frac{r_{12} + r_{23}}{1 + r_{12}r_{23}}\right)^2 . \tag{68}$$

In particular, for *normal incidence*, this becomes

$$\mathcal{R} = \left(\frac{n_1 - n_3}{n_1 + n_3}\right)^2 , \tag{69}$$

and is seen to be independent of n_2. Now the only difference in the case of oblique incidence is the replacement of n_j by $n_j \cos\theta_j$ ($j = 1, 2, 3, \ldots$) in all the formulae; hence *a plate whose optical thickness is* $m\lambda_0/2\cos\theta_2$ ($m = 1, 2, 3, \ldots$) *has no influence on the intensity of the reflected (or transmitted) radiation.*

Next we must determine the nature of these extreme values. After a straightforward calculation we find that when $H = m\lambda_0/4\cos\theta_2$ ($m = 1, 2, \ldots$)

$$\left.\begin{aligned} &\left(\frac{\mathrm{d}^2\mathcal{R}}{\mathrm{d}H^2}\right) \gtrless 0 \\ \text{according as}\quad &(-1)^m r_{12}r_{23}(1 + r_{12}^2 r_{23}^2 - r_{12}^2 - r_{23}^2) \lessgtr 0, \end{aligned}\right\} \tag{70}$$

so that with the upper sign there is a minimum and with the lower sign a maximum. In particular, for *normal incidence*, r_{12} and r_{23} are given by (66) and we have

$$\left.\begin{aligned} &\textit{maximum}, \text{ if } (-1)^m(n_1 - n_2)(n_2 - n_3) > 0, \\ &\textit{minimum}, \text{ if } (-1)^m(n_1 - n_2)(n_2 - n_3) < 0. \end{aligned}\right\} \tag{71}$$

Usually the first medium is air ($n_1 \sim 1$) and we see that *with a film whose optical thickness has any of the values* $\lambda_0/4$, $3\lambda_0/4$, $5\lambda_0/4$, … *the reflectivity is then a maximum or a minimum according to whether the refractive index of the film is greater or smaller than the refractive index of the last medium; for a film whose optical thickness has any of the values* $\lambda_0/2$, $2\lambda_0/2$, $3\lambda_0/2$, … *the opposite is the case.*

These results, which are illustrated in Fig. 1.18, are found to be in good agreement with experiment.*

It is evident from the preceding analysis that a plate whose optical thickness is a quarter of the wavelength and whose refractive index is low enough may be used as an *antireflection film*, i.e. a film by means of which the reflectivity of a surface is reduced.

* See, for example, K. Hammer, *Z. tech. Phys.*, **24** (1943), 169.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

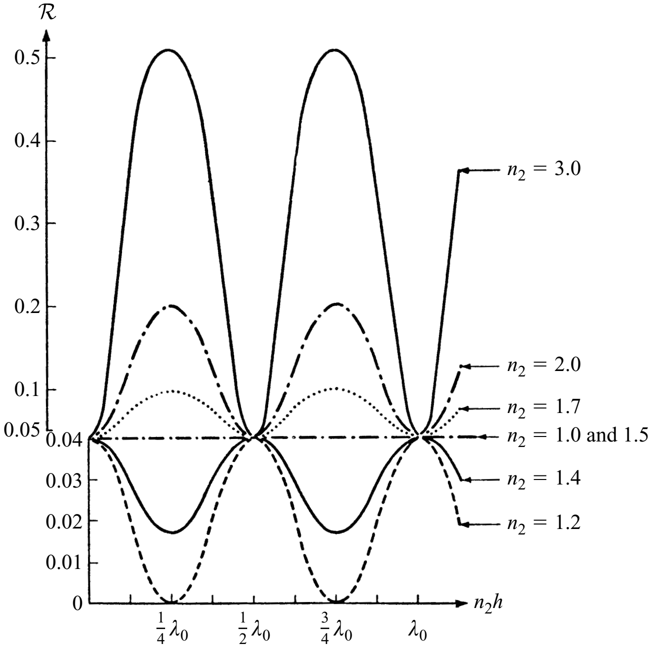

Fig. 1.18 The reflectivity of a dielectric film of refractive index n_2 as a function of its optical thickness. ($\theta_1 = 0$, $n_1 = 1$, $n_3 = 1.5$). [After R. Messner, *Zeiss Nachr.*, **4** (H9) (1943), 253.]

(The surface is then said to be 'bloomed'.) The two substances most commonly used for this purpose are cryolite ($n \sim 1.35$) and magnesium fluoride (MgF$_2$, $n \sim 1.38$)*.

According to (67), the reflectivity at normal incidence would be strictly zero if

$$n_2 = \sqrt{n_1 n_3}. \tag{72}$$

With $n_1 = 1$, $n_3 = 1.5$ this demands $n_2 \sim 1.22$, a condition which cannot be satisfied in practice. A fuller analysis of (59) shows, however, that with oblique incidence it is possible to have zero reflectivity for a *TM* wave (electric vector parallel to the plane of incidence) but not for a *TE* wave (electric vector perpendicular to the plane of incidence), i.e. under favourable conditions one can have simultaneously $\mathcal{R}_\parallel = 0$, $\mathcal{R}_\perp \neq 0$. Hence a thin film of a suitable dielectric material may also be used as a *polarizer*, working by reflection. Such a polarizer may be regarded as a generalization of the simple arrangement discussed earlier in connection with Brewster's angle. To obtain a large value for $\mathcal{R}_\perp$ (with $\mathcal{R}_\parallel = 0$), the refractive index n_2 of the film must be as large as possible.† For example, with $n_1 = 1$, $n_2 = 2.5$, $n_3 = 1.53$ one obtains $\mathcal{R}_\parallel = 0$, $\mathcal{R}_\perp = 0.79$ when $\theta_1 = 74° 30'$.

If a glass surface is coated with a material of sufficiently high refractive index, the reflectivity of the surface will, according to the preceding analysis, be greatly enhanced

* The design and performance of multilayer antireflection films are discussed by A. Musset and A. Thelen in *Progress in Optics*, Vol. 8, ed. E. Wolf (Amsterdam, North-Holland Publishing Company and New York, American Elsevier Publishing Company, 1970), p. 201.

† Cf. H. Schröder, *Optik* **3** (1948), 499.

(see Figs. 1.18 and 1.19). The surface will then act as a good beam-splitter. Coatings of titanium dioxide (TiO_2, $n \sim 2.45$) or zinc sulphide (ZnS, $n \sim 2.3$) are very suitable for this purpose, giving a maximum reflectivity of about 0.3. There are other substances which have high refractive indices, but they absorb some of the incident light. For example, with a coating of stibnite (Sb_2S_3, $n \sim 2.8$) one can attain the values $\mathcal{R} = \mathcal{T} = 0.46$, but 8 per cent of the incident light is then absorbed by the film.

It is also of interest to examine the case when total reflection takes place on the first boundary. In this case

$$n_1 \sin\theta_1 > n_2, \qquad n_1 \sin\theta_1 < n_3,$$

and [see §1.5 (54)],

$$n_2 \cos\theta_2 = \mathrm{i}\sqrt{n_1^2 \sin^2\theta_1 - n_2^2}. \tag{73}$$

The coefficients for reflection at the two boundaries now are (see §1.5 (21))

$$\left.\begin{aligned} r_{12} &= \frac{n_1\cos\theta_1 - \mathrm{i}\sqrt{n_1^2\sin^2\theta_1 - n_2^2}}{n_1\cos\theta_1 + \mathrm{i}\sqrt{n_1^2\sin^2\theta_1 - n_2^2}}, \\ r_{23} &= \frac{\mathrm{i}\sqrt{n_1^2\sin^2\theta_1 - n_2^2} - n_3\cos\theta_3}{\mathrm{i}\sqrt{n_1^2\sin^2\theta_1 - n_2^2} + n_3\cos\theta_3}. \end{aligned}\right\} \tag{74}$$

If we set

$$k_0 n_2 h \cos\theta_2 = \mathrm{i}b, \tag{75}$$

where, according to (73),

$$b = \frac{2\pi}{\lambda_0} h \sqrt{n_1^2 \sin^2\theta_1 - n_2^2}, \tag{76}$$

we obtain for the reflection coefficients the following expression, in place of (57):

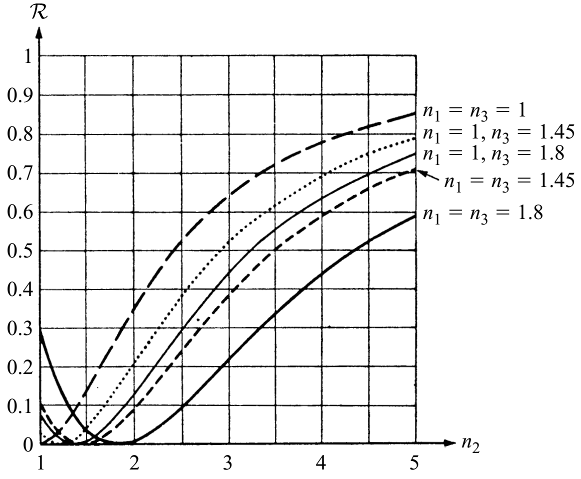

Fig. 1.19 The reflectivity at normal incidence of a quarter-wave film ($n_2 h = \lambda_0/4$) as a function of the refractive index n_2 of the film. (After K. Hammer, *Z. tech. Phys.*, **24**, (1943), 169.)

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$r = \frac{r_{12} + r_{23}e^{-2b}}{1 + r_{12}r_{23}e^{-2b}}. \tag{77}$$

Since $|r_{12}| = |r_{23}| = 1$, r_{12} and r_{23} are of the form

$$r_{12} = e^{i\phi_{12}}, \qquad r_{23} = e^{i\phi_{23}}, \tag{78}$$

where the ϕ's are real; hence the reflectivity now is

$$\mathcal{R} = |r|^2 = \frac{e^{2b} + e^{-2b} + 2\cos(\phi_{12} - \phi_{23})}{e^{2b} + e^{-2b} + 2\cos(\phi_{12} + \phi_{23})}. \tag{79}$$

In contrast with the previous case, $\mathcal{R}$ is now no longer a periodic function of the thickness of the film. Eq. (76) shows that if the dependence of the refractive index on the wavelength is neglected, b is inversely proportional to wavelength. Since for sufficiently large values of b, $\mathcal{R}$ will be practically unity, the shorter wavelengths will not be transmitted; the film then acts as a *low-pass filter*, i.e. one which transmits the long wavelengths only.

We have seen that by the use of dielectric films of suitable material, many useful effects can be obtained. It will be apparent that, with a number of such films arranged in succession, the desired features may be still further enhanced. The characteristic matrix of such a *multilayer* may be obtained with the help of the theorem expressed by (41).* We shall discuss in detail only the case when the multilayer is periodic.

1.6.5 Periodically stratified media

A stratified periodic medium with period h is characterized by a dielectric constant ε and a magnetic permeability μ which are functions of z only and are such that

$$\varepsilon(z + jh) = \varepsilon(z), \qquad \mu(z + jh) = \mu(z),$$

j being an integer in some fixed range $1 \leqslant j \leqslant N$.

Let $\boldsymbol{M}(h)$ be the characteristic matrix corresponding to one period and write†

$$\boldsymbol{M}(h) = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}. \tag{80}$$

According to (41) we then have, on account of the periodicity,

$$\boldsymbol{M}(Nh) = \underbrace{\boldsymbol{M}(h)\cdot\boldsymbol{M}(h)\cdots\boldsymbol{M}(h)}_{N\text{ times}} = [\boldsymbol{M}(h)]^N. \tag{81}$$

To evaluate the elements of the matrix $\boldsymbol{M}(Nh)$ we use a result from the theory of matrices, according to which the Nth power of a unimodular matrix $\boldsymbol{M}(h)$ is‡

* Formulae relating to multilayers have been given by many writers, e.g. R. L. Mooney, *J. Opt. Soc. Amer.*, **36** (1946), 256; W. Weinstein, *ibid*, **37** (1947), 576.

† We now omit the prime on the matrix elements.

‡ The correctness of this result may be verified by induction, using the recurrence relation

$$\mathcal{U}_j(x) = 2x\mathcal{U}_{j-1}(x) - \mathcal{U}_{j-2}(x),$$

which follows as an identity from the definition of the Chebyshev polynomials.

A direct proof based on the theory of matrices was given by F. Abelès, *Ann. de Physique*, **5** (1950), 777.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$[\boldsymbol{M}(h)]^N = \begin{bmatrix} m_{11}\mathcal{U}_{N-1}(a) - \mathcal{U}_{N-2}(a) & m_{12}\mathcal{U}_{N-1}(a) \\ m_{21}\mathcal{U}_{N-1}(a) & m_{22}\mathcal{U}_{N-1}(a) - \mathcal{U}_{N-2}(a) \end{bmatrix}, \tag{82}$$

where

$$a = \tfrac{1}{2}(m_{11} + m_{22}), \tag{83}$$

and $\mathcal{U}_N$ are the *Chebyshev polynomials* of the second kind*:

$$\mathcal{U}_N(x) = \frac{\sin[(N+1)\cos^{-1} x]}{\sqrt{1-x^2}}. \tag{84}$$

A multilayer usually consists of a succession of homogeneous layers of alternately low and high refractive indices n_2 and n_3 and of thickness h_2 and h_3, placed between two homogeneous media of refractive indices n_1 and n_l. (See Fig. 1.20.) We again assume the media to be nonmagnetic ($\mu = 1$) and set

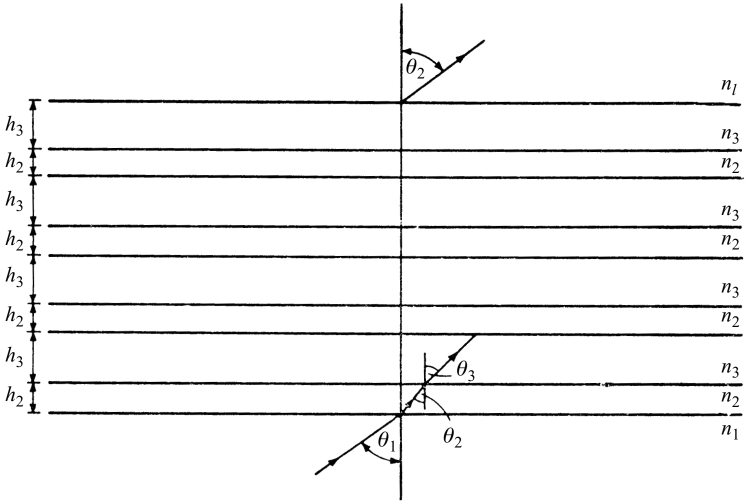

Fig. 1.20 A periodic multilayer.

* These polynomials satisfy the following orthogonality and normalizing conditions:

$$\int_{-1}^{+1} \mathcal{U}_n(x)\mathcal{U}_m(x)\sqrt{1-x^2}\,\mathrm{d}x = 0 \text{ when } n \neq m$$
$$= \frac{\pi}{2} \text{ when } n = m.$$

The explicit expressions of the first six polynomials are:

$$\mathcal{U}_0(x) = 1, \qquad \mathcal{U}_3(x) = 8x^3 - 4x,$$
$$\mathcal{U}_1(x) = 2x, \qquad \mathcal{U}_4(x) = 16x^4 - 12x^2 + 1,$$
$$\mathcal{U}_2(x) = 4x^2 - 1, \qquad \mathcal{U}_5(x) = 32x^5 - 32x^3 + 6x.$$

Tables of Chebyshev polynomials have been published by the National Bureau of Standards Washington (Applied Mathematics Series **9** (1952)), where the main properties of the polynomials are also summarized. See also *Higher Transcendental Functions* [Bateman Manuscript Project (New York, McGraw-Hill, Vol. 2 1953), p. 183].

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\left.\begin{aligned}\beta_2 &= \frac{2\pi}{\lambda_0} n_2 h_2 \cos\theta_2, &\qquad \beta_3 &= \frac{2\pi}{\lambda_0} n_3 h_3 \cos\theta_3,\\ p_2 &= n_2 \cos\theta_2, & p_3 &= n_3\cos\theta_3,\\ &&h = h_2 + h_3.\end{aligned}\right\} \tag{85}$$

The characteristic matrix $\boldsymbol{M}_2(h)$ of one period then is, according to (39) and (41),

$$\boldsymbol{M}_2(h) = \begin{bmatrix} \cos\beta_2 & -\dfrac{\mathrm{i}}{p_2}\sin\beta_2 \\ -\mathrm{i}p_2\sin\beta_2 & \cos\beta_2\end{bmatrix}\begin{bmatrix} \cos\beta_3 & -\dfrac{\mathrm{i}}{p_3}\sin\beta_3 \\ -\mathrm{i}p_3\sin\beta_3 & \cos\beta_3\end{bmatrix}$$
$$= \begin{bmatrix} \cos\beta_2\cos\beta_3 - \dfrac{p_3}{p_2}\sin\beta_2\sin\beta_3 & -\dfrac{\mathrm{i}}{p_3}\cos\beta_2\sin\beta_3 - \dfrac{\mathrm{i}}{p_2}\sin\beta_2\cos\beta_3 \\ -\mathrm{i}p_2\sin\beta_2\cos\beta_3 - \mathrm{i}p_3\cos\beta_2\sin\beta_3 & \cos\beta_2\cos\beta_3 - \dfrac{p_2}{p_3}\sin\beta_2\sin\beta_3\end{bmatrix}. \tag{86}$$

Hence according to (81), the characteristic matrix $\boldsymbol{M}_{2N}(Nh)$ of the multilayer (with $2N$ films in all) is given by the following formula due to Abelès:

$$\boldsymbol{M}_{2N}(Nh) = \begin{bmatrix}\mathcal{M}_{11} & \mathcal{M}_{12}\\ \mathcal{M}_{21} & \mathcal{M}_{22}\end{bmatrix} \tag{87}$$

where

$$\left.\begin{aligned}\mathcal{M}_{11} &= \left(\cos\beta_2\cos\beta_3 - \frac{p_3}{p_2}\sin\beta_2\sin\beta_3\right)\mathcal{U}_{N-1}(a) - \mathcal{U}_{N-2}(a),\\ \mathcal{M}_{12} &= -\mathrm{i}\left(\frac{1}{p_3}\cos\beta_2\sin\beta_3 + \frac{1}{p_2}\sin\beta_2\cos\beta_3\right)\mathcal{U}_{N-1}(a),\\ \mathcal{M}_{21} &= -\mathrm{i}(p_2\sin\beta_2\cos\beta_3 + p_3\cos\beta_2\sin\beta_3)\mathcal{U}_{N-1}(a),\\ \mathcal{M}_{22} &= \left(\cos\beta_2\cos\beta_3 - \frac{p_2}{p_3}\sin\beta_2\sin\beta_3\right)\mathcal{U}_{N-1}(a) - \mathcal{U}_{N-2}(a),\end{aligned}\right\} \tag{88}$$

and

$$a = \cos\beta_2\cos\beta_3 - \tfrac{1}{2}\left(\frac{p_2}{p_3} + \frac{p_3}{p_2}\right)\sin\beta_2\sin\beta_3. \tag{89}$$

The reflection and transmission coefficients of the multilayer are immediately obtained by substituting these expressions into (49) and (50).

Of particular interest is the case when the two basic layers are of the same optical thickness (usually $\lambda_0/4$), i.e. when

$$n_2 h_2 = n_3 h_3, \tag{90}$$

and the incidence is normal ($\theta_1 = 0$). Then

$$\beta_2 = \beta_3 = \frac{2\pi}{\lambda_0} n_2 h_2 = \frac{2\pi}{\lambda_0} n_3 h_3, \tag{91}$$

and if we denote this common value by β, the argument of the Chebyshev polynomials reduces to

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$a = \cos^2\beta - \tfrac{1}{2}\left(\frac{n_2}{n_3} + \frac{n_3}{n_2}\right)\sin^2\beta. \tag{92}$$

It is seen that a cannot exceed unity, but that for some values of β it may become smaller than -1. Then $\cos^{-1}a$ will be imaginary and consequently, since for any χ

$$\sin \mathrm{i}\chi = \mathrm{i}\sinh\chi = \mathrm{i}\frac{\mathrm{e}^{\chi} - \mathrm{e}^{-\chi}}{2},$$

$\mathcal{U}_N$ will have exponential behaviour. It follows that the reflectivity of such a multilayer will increase rapidly with the number of the periods.

With *quarter-wave films* ($n_2h_2 = n_3h_3 = \lambda_0/4$) at normal incidence (again assuming nonmagnetic media),

$$\beta = \pi/2, \qquad p_2 = n_2, \qquad p_3 = n_3, \tag{93}$$

and (86) reduces to

$$\boldsymbol{M}_2(h) = \begin{bmatrix} -\dfrac{n_3}{n_2} & 0 \\ 0 & -\dfrac{n_2}{n_3} \end{bmatrix}. \tag{94}$$

The characteristic matrix (87) of the multilayer whose basic period is such a double layer is, as can be directly verified by multiplying (94) N times by itself,

$$\boldsymbol{M}_{2N}(Nh) = \begin{bmatrix} \left(-\dfrac{n_3}{n_2}\right)^N & 0 \\ 0 & \left(-\dfrac{n_2}{n_3}\right)^N \end{bmatrix}. \tag{95}$$

According to (49) and (51) the reflectivity is

$$\mathcal{R}_{2N} = \left(\frac{1 - \dfrac{n_l}{n_1}\left(\dfrac{n_2}{n_3}\right)^{2N}}{1 + \dfrac{n_l}{n_1}\left(\dfrac{n_2}{n_3}\right)^{2N}}\right)^2. \tag{96}$$

This shows that for a fixed number N of the double layers,* $\mathcal{R}_{2N}$ increases when the ratio n_2/n_3 is increased, and that if this ratio is fixed $\mathcal{R}_{2N}$ increases with N.

Sometimes, for example, for plate coatings of the Fabry–Perot interferometer (see §7.6) the layers are arranged in succession characterized by the sequence n_2, n_3, n_2, $n_3, \ldots,$ n_2, n_3, n_2 of refractive indices. The characteristic matrix of this multilayer is

$$\boldsymbol{M}_{2N+1} = \boldsymbol{M}_{2N} \cdot \boldsymbol{M}, \tag{97}$$

where $\boldsymbol{M}_{2N}$ is given by (87) and $\boldsymbol{M}$ is the characteristic matrix of the last film in the sequence. In particular with quarter-wave films at normal incidence, $\boldsymbol{M}_{2N}$ reduces to (95), $\beta_2 = \pi/2$, and (97) then becomes

* A thorough discussion of the properties of a system of double layers will be found in a paper by C. Dufour and A. Herpin, *Rev. Opt.*, **32** (1953), 321.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press

$$\boldsymbol{M}_{2N+1} = \begin{bmatrix} 0 & -\dfrac{\mathrm{i}}{n_2}\left(-\dfrac{n_3}{n_2}\right)^N \\ -\mathrm{i}n_2\left(-\dfrac{n_2}{n_3}\right)^N & 0 \end{bmatrix}. \tag{98}$$

Substitution into (49) and (51) gives the required reflectivity:

$$\mathcal{R}_{2N+1} = \left(\frac{1 - \left(\dfrac{n_2}{n_1}\right)\left(\dfrac{n_2}{n_l}\right)\left(\dfrac{n_2}{n_3}\right)^{2N}}{1 + \left(\dfrac{n_2}{n_1}\right)\left(\dfrac{n_2}{n_l}\right)\left(\dfrac{n_2}{n_3}\right)^{2N}} \right)^2. \tag{99}$$

The reflectivity is seen to increase rapidly with the ratio n_2/n_3 and with N (see Table 1.3).

Table 1.3. *Reflectivity $\mathcal{R}_{2N+1}$ of multilayers formed by a periodic succession of quarter-wave films of zinc sulphide and cryolite at normal incidence* ($n_1 = 1$, $n_2 = 2.3$, $n_3 = 1.35$, $n_l = 1.52$, $n_2 h_2 = n_3 h_3 = \lambda_0/4$, $\lambda_0 = 5460$ Å, $\theta_1 = 0$).

N	$\mathcal{R}_{2N+1}$
0	0.306
1	0.672
2	0.872 (0.865)
3	0.954 (0.945)
4	0.984 (0.97)

The values in brackets are experimental results obtained by P. Giacomo, *Compt. Rend. Acad. Sci., Paris*, **235** (1952), 1627.

https://doi.org/10.1017/9781108769914.004 Published online by Cambridge University Press