2 Wave Properties of Light

2.1 INTRODUCTION

In Chapter 1 it was noted that electromagnetic radiation, including light, exhibits both wave and particle properties, and that the type of behaviour exhibited at any one time depended upon the special circumstances.

In this chapter we shall concentrate just on the wave properties. Most of these were discovered and examined in the nineteenth century, before the advent of quantum mechanics (in 1901). The success of the wave theory was remarkable, and it led to a number of important devices, some of which are described in this chapter.

We shall begin, in earnest, in Section 2.4, by looking at some aspects of the wave theory's crowning glory: Maxwell's equations for the electromagnetic field.

Before coming to that, however, there is, first, further emphasis on the range of radiation which lies within the electromagnetic spectrum and, second, a close look at the thinking behind the complex exponential representation of sinusoidal waves, because this is a most convenient and very widely used stratagem in all wave manipulations.

2.2 THE ELECTROMAGNETIC SPECTRUM

We already noted (Section 1.5) that Maxwell's equations allow the electromagnetic wave frequency ω to take any value from (almost) zero to infinity. At value zero we have electrostatics and magnetostatics. At any value other than zero we are dealing with electrodynamics and magnetodynamics.

The dynamic spectrum starts at the very low frequency radio waves and rises to the very high frequency gamma waves (Figure 1.4). Methods of generation, interactions with material media, and methods for detection all vary widely with frequency, leading to a variety of disciplines corresponding to the various frequency ranges (e.g., radio, microwave, infrared, optical, ultraviolet, X-ray, gamma rays, etc.). Photonics deals essentially with the optical range (wavelengths 400 nm to 700 nm) with short extensions to either side (i.e., into the near infrared and near ultraviolet). It deals essentially with that part of the spectrum where the sun's light is most intense at the earth's surface, and also where efficient laser action is presently possible.

2.3 WAVE REPRESENTATION

In the study of optical waves within this chapter, we shall be examining many of the quantitative relationships between them and with other physical entities. Waves are sinusoids of the form

$$E = e_0 \cos(\omega t + \varphi)$$

(We know from Fourier theory that any physical field disturbance can be expressed as the sum of such sinusoids.)

A particular problem in the manipulation of such quantities is that the trigonometric expansions are rather cumbersome—that is, in this case,

$$E_0 = e_0 \cos\omega t \cos\varphi - e_0 \sin\omega t \sin\varphi$$

and if, for example, we wish to add two such waves, of equal frequency, to produce a resultant, this will be another wave of the same frequency but with a different amplitude and phase. Our problem, in this case, is to find this resultant wave. That is to say, for

$$e_1 \cos(\omega t + \varphi_1) + e_2 \cos(\omega t + \varphi_2) = e_T \cos(\omega t + \varphi_T)$$

what are e_T and φ_T (Figure 2.1)?

The mathematics of this, although straightforward, is tedious and hence vulnerable to error; and the tedium increases rapidly with the number of waves.

A convenient solution to this is to express the sinusoid in its complex exponential form, for this allows factorizations that simplify the mathematics considerably. Because this stratagem is used extensively in this book and elsewhere, it is, perhaps, worth taking some time to appreciate it more fully.

A sinusoidal wave of the form

$$E = e_0 \cos(\omega t + \varphi)$$

is the real part of the expression

$$E' = e_0 \cos(\omega t + \varphi) + ie_0 \sin(\omega t + \varphi)$$

(E' might be expected to be a convenient mathematical entity because i [the square root of minus one] effectively is an operator that acts to rotate through $\pi/2$, and it thus brings together the cosine and sine that are, of course, $\pi/2$ out of phase.)

Now the well-known exponential expressions for cosine and sine are

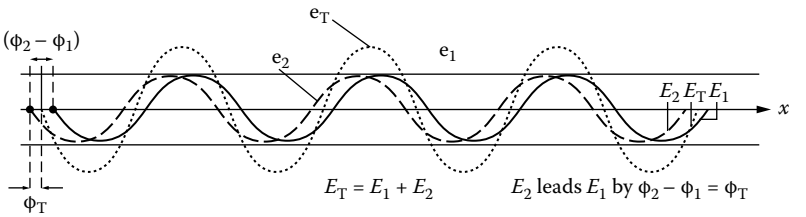


FIGURE 2.1 Addition of two waves of the same frequency.

$$\cos\varphi = \frac{1}{2}[\exp(i\phi) + \exp(-i\varphi)]$$

$$\sin\varphi = \frac{1}{2i}[\exp(i\phi) - \exp(-i\varphi)]$$

hence,

$$\cos\ \varphi + i\ \sin\ \varphi = \exp(i\varphi)$$

and thus,

$$\begin{aligned} E' &= e_0 \cos(\omega t + \varphi) + ie_0 \sin(\omega t + \varphi) \\ &= e_0 \exp[i(\omega t + \varphi)] \\ &= e_0 \exp(i\varphi)\exp(i\omega t) \end{aligned}$$

Our original wave, $e_0 \cos(\omega t + \varphi)$, is just the real part of this, and this is sometimes expressed by writing

$$e_0 \cos(\omega t + \varphi) = \mathrm{Re}[e_0 \exp(i\varphi)\exp(i\omega t)]$$

Now any sinusoidal wave can be written in this form, even a sine wave (as opposed to a cosine wave), because all that is necessary in that case is to subtract $\pi/2$ from the phase—that is,

$$e_0 \sin(\omega t + \varphi) = \mathrm{Re}\{e_0 \exp\left[i\left(\varphi - \frac{1}{2}\pi\right)\right]\exp(i\omega t)\}$$

The really important point about all of this, now, is that if we are dealing with a number of waves of the same frequency, then the frequency term may be factored out to leave a term that is a complex number whose modulus represents the wave amplitude, and whose argument represents the wave's phase. In our simple case,

$$E' = e_0 \exp(i\varphi)\exp(i\omega t) = E\exp(\omega t)$$

where

$$E = e_0 \exp(i\varphi)$$

and E is a complex number with the properties

$$|E| = e_0;\ \arg(E) = \varphi$$

Similarly, for any two waves of the same frequency,

$$E_1 = e_1 \cos(\omega t + \varphi_1)$$

$$E_2 = e_2 \cos(\omega t + \varphi_2)$$

The complex exponential forms are

$$E_1' = e_1 \exp(i\varphi_1)\exp(i\omega t)$$

$$E_2' = e_2 \exp(i\varphi_2)\exp(i\omega t)$$

Suppose, as earlier, we wish to find the amplitude and phase of the wave which result from the sum of these two waves. We may write

$$\begin{aligned} E_T' &= E_1' + E_2' \\ &= \exp(i\omega t)[e_1 \exp(i\varphi_1) + e_2 \exp(i\varphi_2)] \\ &= \exp(i\omega t)(E_1 + E_2) \end{aligned}$$

The complex amplitude term can always be written in the form

$$E_1 + E_2 = a + ib$$

where a and b are real numbers. Hence, the resultant wave amplitude is given by

$$e_T = |a + ib| = (a^2 + b^2)^{1/2}$$

and its phase by

$$\varphi_T = \arg(a + ib) = \tan^{-1}\left(\frac{b}{a}\right)$$

and hence, finally,

$$e_1 \cos(\omega t + \varphi_1) + e_2 \cos(\omega t + \varphi_2) = e_T \cos(\omega t + \varphi_T)$$

So that

$$e_1 \exp(i\varphi_1) + e_2 \exp(i\varphi_2) = e_T \exp(i\varphi_T)$$

and hence,

$$e_1 \cos\varphi_1 + e_2 \cos\varphi_2 + ie_1 \sin\varphi_1 + ie_2 \sin\varphi_2 = e_T(\cos\varphi_T + i \sin\varphi_T)$$

Here, then, we have

$$a = e_1 \cos\varphi_1 + e_2 \cos\varphi_2$$

$$b = e_1 \sin\varphi_1 + e_2 \sin\varphi_2$$

Hence,

$$e_T = (a^2 + b^2)^{1/2} = e_1^2 + e_2^2 + 2e_1e_2 \cos(\varphi_1 - \varphi_2)$$

$$\varphi_T = \tan^{-1}\left(\frac{b}{a}\right) = \tan^{-1}\left(\frac{e_1 \sin\varphi_1 + e_2 \sin\varphi_2}{e_1 \cos\varphi_1 + e_2 \cos\varphi_2}\right)$$

and the problem is solved.

This is a much more convenient mathematical technique than that which involves the trigonometric identities for the expansion of the cosines, and the convenience becomes markedly more noticeable as the number of waves increases.

For these reasons, the complex exponential representation is to be used extensively in this chapter. Occasionally, however, it is convenient to revert to the real sinusoid. This usually is in those cases where the optical frequency is of direct importance, so that its removal as a separate factored complex quantity cannot conveniently represent the true physical condition.

2.4 ELECTROMAGNETIC WAVES

2.4.1 Velocity and Refractive Index

In 1864 Clerk Maxwell showed conclusively that light waves were electromagnetic in nature. He did this by expressing the then known laws of electromagnetism in such a way as to allow him to derive from them a wave equation (see Appendix I). This wave equation permitted free-space solutions that corresponded to electromagnetic waves with a velocity equal to the known experimental value of the velocity of light. The consequent recognition of light as an electromagnetic phenomenon was probably the single most important advance in the progression of its understanding.

All the important features of light waves follow from a detailed examination of Maxwell's equations (see Appendix I). Taking Cartesian axes Ox, Oy, Oz (Figure 2.2), a typical sinusoidal solution is given by

$$\begin{aligned} E_x &= E_0 \exp[i(\omega t - kz)] \\ H_y &= H_0 \exp[i(\omega t - kz)] \end{aligned} \tag{2.1}$$

which states that the electric field oscillates sinusoidally in the xz plane, the magnetic field oscillates in the yz plane (i.e., orthogonally to the E field) and in phase with the

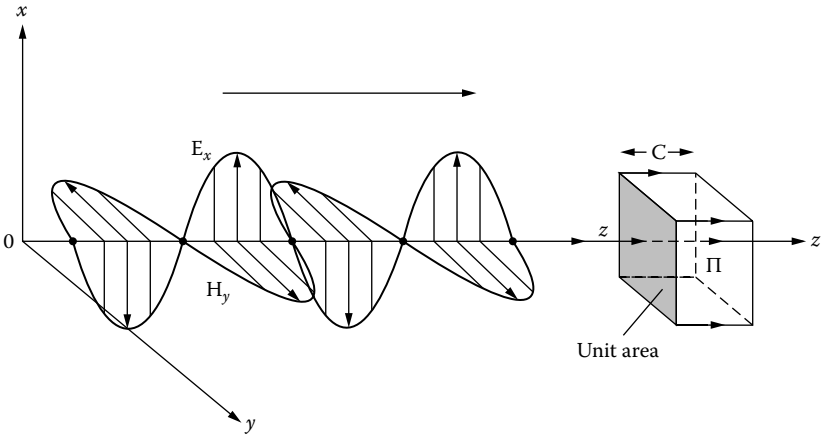


FIGURE 2.2 Electromagnetic wave and energy flow (Poynting vector: Π).

E field, and the wave propagates in the Oz direction (Figure 2.2). The frequency and wavelength of the wave are given by

$$f = \frac{\omega}{2\pi}$$

$$\lambda = \frac{2\pi}{k}$$

and $f\lambda = \omega/k = c$ where c is the wave velocity. This latter is related to the electromagnetic properties of the medium in which the wave propagates via the relation

$$c = (\varepsilon\mu)^{-1/2} \tag{2.2}$$

where ε is the electric permittivity of the medium, and μ is its magnetic permeability. The relation (2.2) can also be written in the form

$$c = (\varepsilon_R \varepsilon_0 \mu_R \mu_0)^{-1/2}$$

that is,

$$\varepsilon = \varepsilon_R \varepsilon_0, \mu = \mu_R \mu_0$$

where ε_R, μ_R are the permittivity and permeability factors for the medium relative to those for free space, ε_0, μ_0; ε_R is often called the dielectric constant. The electric displacement $\boldsymbol{D}$ and the magnetic flux density $\boldsymbol{B}$ are defined by the relations

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}$$

$$\boldsymbol{B} = \mu \boldsymbol{H}$$

(For reasons of symmetry, $\boldsymbol{D}$ is sometimes called the electric flux density.) We can, therefore, also write

$$c = \frac{c_0}{(\varepsilon_R \mu_R)^{1/2}} \tag{2.3}$$

where c_0 is the velocity of the electromagnetic wave in free space and has the experimentally determined value 2.997925×10^8 ms^{-1}.

For most optical media of any importance, we have $\mu_R \sim 1$ and $\varepsilon_R > 1$. These materials belong to the class known as dielectrics and they are electrical insulators. Thus, we may write Equation (2.3) in the form

$$c \approx \frac{c_0}{(\varepsilon_R)^{1/2}}$$

and note that $c < c_0$. The ratio c_0/c is, by definition, the refractive index n of the medium, so that

$$n \approx \varepsilon_R^{1/2} \tag{2.4}$$

where n is thus the factor by which light travels more slowly in an optical medium than it does in free space. Now ε_R is a measure of the ease with which the medium can be polarized electrically by the action of an external electric field. (See Section 4.2 for more details of this.) This polarization depends on the mobility of the electrons, within the molecule, in the face of resistance by molecular forces. Clearly then, ε_R will depend on the frequency of the applied electric field, because it will depend on how quickly these forces can respond to the field. Thus, Equation (2.4) will be true only if n and ε_R refer to the same frequency of wave; hence, we also note that n is frequency dependent.

2.4.2 Energy, Power, and Intensity

Let us now consider the energy content of the wave. For an electric field, the energy per unit volume, u_E, is given by (see, for example, Reference [1])

$$u_E = \frac{1}{2}\varepsilon E^2$$

and for a magnetic field,

$$u_H = \frac{1}{2}\mu H^2$$

Maxwell's equations relate E and H for an electromagnetic wave according to (see Appendix I)

$$H = \left(\frac{\varepsilon}{\mu}\right)^{1/2} E$$

Hence, the total energy density in the wave is given by

$$u = u_E + u_H = \varepsilon E^2 = \mu H^2 \tag{2.5}$$

Consider now the plane wave propagating in the direction Oz (Figure 2.2). The total energy flowing across unit area in unit time in the direction Oz will be that contained within a volume c m^3, where c is the wave velocity. Hence, the power flux across unit area is given by

$$\frac{\text{power}}{\text{area}} = c\varepsilon\, E^2 = \left(\frac{\varepsilon}{\mu}\right)^{1/2} E^2$$

Clearly, if the electric field E varies sinusoidally, this quantity also will vary sinusoidally; for example, if

The average value of this quantity over one period of oscillation is called the 'intensity' of the wave (sometimes the irradiance) and clearly represents the measurable power per unit area for any device that cannot respond to optical frequencies (i.e., the vast majority).

Hence, we have

$$I = \left\langle \frac{\text{power}}{\text{area}} \right\rangle = \left(\frac{\varepsilon}{\mu} \right)^{1/2} \left\langle E^2 \right\rangle = \left(\frac{\varepsilon}{\mu} \right)^{1/2} \frac{1}{2} E_0^2 \tag{2.6a}$$

(where $\langle\rangle$ denotes the average value) because cos $2\omega t$ averages to zero.

Clearly, I is proportional to the square of the electric field amplitude, and also, from Equation (2.5), it will be proportional to the square of the magnetic field amplitude. The quantity I has MKS units of watts.metres^{-2}.

More generally, the intensity is expressed in terms of the Poynting vector $\boldsymbol{\Pi}$ (see Appendix I):

$$\boldsymbol{\Pi} = \boldsymbol{E} \times \boldsymbol{H}$$

where $\boldsymbol{E}$ and $\boldsymbol{H}$ are now vector quantities and $\boldsymbol{E} \times \boldsymbol{H}$ is their vector product (see Appendix I and Chapter 1). The intensity of the wave will be the value of $\boldsymbol{\Pi}$ averaged over one period of the wave. If $\boldsymbol{E}$ and $\boldsymbol{H}$ are spatially orthogonal and in phase, as in the case of a wave propagating in an isotropic dielectric medium, then

$$I = \langle \boldsymbol{\Pi} \rangle = = c\varepsilon \, E^2 = c\mu H^2$$

as before. As is to be expected, in some more exotic cases (e.g., anisotropic media), the $\boldsymbol{E}$ and $\boldsymbol{H}$ components are neither orthogonal nor in phase, but $\langle \boldsymbol{\Pi} \rangle$ will still provide the average power flow across unit area. If, for example, $\boldsymbol{E}$ and $\boldsymbol{H}$ happened to be in phase quadrature, then we should have

$$I = \langle \boldsymbol{\Pi} \rangle = \langle E_0 \cos \omega t \; H_0 \sin \omega t \rangle = 0$$

and thus there is no mean power flow. (This result should be noted for reference to the case of 'evanescent' waves, which will be considered later.)

In an optical medium with $\mu_R \approx 1$, Equation (2.6a) can be written

$$I = \left(\frac{\varepsilon_R \varepsilon_0}{\mu_R \mu_0} \right)^{1/2} \frac{1}{2} E_0^2 = n \left(\frac{\varepsilon_0}{\mu_0} \right)^{1/2} \frac{1}{2} E_0^2 \tag{2.6b}$$

where n is, again, the refractive index of the medium.

The quantity $(\mu_0/\varepsilon_0)^{1/2}$ is sometimes called the 'impedance of free space' and given the symbol Z_0. This is because, in free space,

$$\frac{E}{H} = \left(\frac{\mu_0}{\varepsilon_0} \right)^{1/2} = Z_0$$

Because $\boldsymbol{E}$ has dimensions of volts.metres^{-1} and $\boldsymbol{H}$ of amps.metres^{-1}, Z_0 clearly has the dimensions of impedance (ohms); Z_0 is real and has the MKS value:

$$\left(\frac{\mu_0}{\varepsilon_0} \right)^{1/2} = \left(\frac{4\pi \times 10^{-7}}{8.854 \times 10^{-12}} \right)^{1/2} = 376.7 \text{ ohms}$$

It follows that (2.6b) can be written

$$I = \frac{n}{2Z_0} E_0^2 = \frac{n}{753.46} E_0^2 = 1.33 \times 10^{-3} n E_0^2 \tag{2.6c}$$

This is a useful relationship in two ways. First, it relates a quantity that is directly measurable (I) with one that is not (E_0). Second, it provides the actual numerical relationship between I and E_0, and this is valuable when designing devices and systems, as we shall discover later.

2.4.3 Optical Polarisation

We should now give brief consideration to what is known as the 'polarization' of the optical wave. (This topic will be dealt with more comprehensively in Chapter 3.)

The 'typical' sinusoidal solution of Maxwell's wave equation given by Equations (2.1) is, of course, only one of an infinite number of such sinusoidal solutions. The general solution for a sinusoid of angular frequency ω is given by

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}(\boldsymbol{r}) \exp(i\omega t)$$

where $\boldsymbol{E}(\boldsymbol{r}, t)$, $\boldsymbol{E}(\boldsymbol{r})$ are, in general, complex vectors, and $\boldsymbol{r}$ is a real radius vector in the xy plane.

If, for simplicity, we consider just plane, monochromatic (single-frequency) waves propagating in free space in the direction Oz, we may, for the $\boldsymbol{E}$ field, write the general solution to the wave equation in the form

$$E_x = e_x \cos(\omega t - kz + \delta_x)$$

$$E_y = e_y \cos(\omega t - kz + \delta_y)$$

where δ_x, δ_y are arbitrary (but constant) phase angles. Thus, we are able to describe this solution completely by means of two waves: one in which the electric field lies entirely in the xz plane, and the other in which it lies entirely in the yz plane (Figure 2.3). If these waves are observed at a particular value of z, say z', then they take the oscillatory form:

$$E_x = e_x \cos(\omega t + \delta_x'); \ \delta_x' = \delta_x - kz'$$

$$E_y = e_y \cos(\omega t + \delta_y'); \ \delta_y' = \delta_y - kz'$$

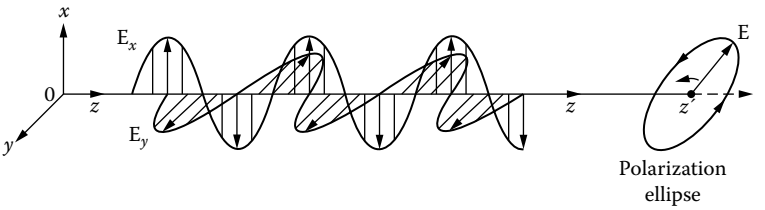


FIGURE 2.3 Electric field components for an elliptically polarized wave.

and the tip of each vector appears to oscillate sinusoidally with time along a line. E_x is said to be linearly polarized in the direction Ox, and E_y linearly polarized in the direction Oy.

The tip of the vector which is the sum of E_x and E_y will, in general, describe an ellipse whose Cartesian equation in the xy plane at the chosen z' will be given by eliminating ωt from the expression for E_x and E_y—that is,

$$\frac{E_x^2}{e_x^2} + \frac{E_y^2}{e_y^2} - \frac{2E_x E_y}{e_x e_y}\cos\delta = \sin^2\delta$$

$$\delta = \delta_y' - \delta_x'$$

This ellipse will degenerate into a straight line (and the overall polarization state of the light will thus be linear) if (a) $e_x \neq 0$; $e_y = 0$ or (b) $e_x = 0$; $e_y \neq 0$ or (c) $\delta = m\pi$, where m is a positive or negative integer. This corresponds to the condition that E_x and E_y are either in phase or in anti-phase.

The ellipse becomes a circle (and the light is thus circularly polarized) if (a) $e_x = e_y$ and (b) $\delta = (2m + 1)\,\pi/2$—that is, the waves are equal in amplitude and are in phase quadrature.

The polarization properties of light waves are especially important for propagation within anisotropic media, in which the physical properties vary with direction. In this case, the propagation characteristics for the component E_x will, in general, differ from those for E_y, so that the values to e_1, e_2, and δ will vary along the propagation path. The polarization state of the light will now become dependent upon the propagation distance, and on the state of the medium. This, also, will be covered in detail in Chapter 3.

2.5 REFLECTION AND REFRACTION

We have seen in Section 2.4 that Maxwell's equations allow a set of solutions of the form

$$E_x = E_0 \exp[i(\omega t - kz)]$$

$$H_y = H_0 \exp[i(\omega t - kz)]$$

with $\omega/k = (\varepsilon\mu)^{-1/2} = c$.

These represent plane waves travelling in the Oz direction. We shall now investigate the behaviour of such waves, with particular regard to the effects which occur at the boundaries between different optical media.

Of course, other types of solution are also possible. An important solution is that of a wave that spreads spherically from a point to a distance r:

$$E_r = \frac{E_0}{r}\exp[i(\omega t - kr)]$$

Here the factor $1/r$ in the amplitude is necessary to ensure conservation of energy (via the Poynting vector) for, clearly, the total area over which the energy flux occurs

is $4\pi r^2$, so that the intensity falls as $1/r^2$. (Remember that intensity is proportional to the square of the amplitude.)

It is interesting and valuable to note that the propagation of a plane wave (such as in Figure 2.4) is equivalent to the propagation of spherical waves radiating from each point on the propagating wavefront of the plane wave. On a given wavefront, the waves at each point begin in phase (this is the definition of a wavefront), so that they remain strictly in phase only in a direction at right angles to the front (Figure 2.4). Hence, the plane wave appears to propagate in that direction. This principle of equivalence, first enunciated by Huygens and later shown by Kirchhoff to be mathematically sound [2], is very useful in the study of wave propagation phenomena generally.

The laws of reflection and refraction were first formulated in terms of 'rays' of light. It had been noticed (c. 1600) that, when dealing with 'point' sources, the light passed through apertures consistently with the view that it was composed of rays travelling in straight lines from the point. (It was primarily this observation that led to Newton's 'corpuscular' theory.) The practical concept was legitimized by allowing such light to pass through a small hole so as to isolate a 'ray'. Such rays were produced, and their behaviour in respect of reflection and refraction at material boundaries was formulated, thus,

(i) On reflection at a boundary between two media, the reflected ray lies in the same plane as that of the incident ray and the normal to the boundary at the point of incidence (the plane of incidence); the angle of reflection equals the angle of incidence.
(ii) On refraction at a boundary, the refracted ray also lies in the plane of incidence, and the sine of the angle of refraction bears a constant ratio to the sine of the angle of incidence (Snell's law).

These two laws form the basis of what is known as geometrical optics, or, 'ray' optics. The majority of bulk optics (e.g., lens design, reflectometers, prismatics) can be formulated with its aid. However, it has severe limitations. For example, it cannot predict the intensities of the refracted and reflected rays.

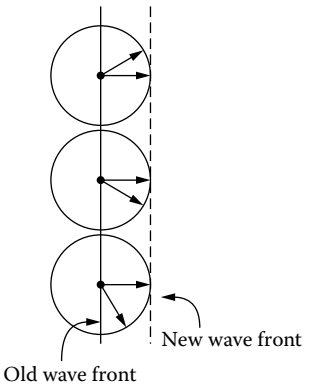


FIGURE 2.4 Huygens' construction.

If, in the attempt to isolate a ray of light of increasing fineness, the aperture is made too small, the ray divergence appears to increase, rather than diminish. This occurs when the aperture size becomes comparable with the wavelength of the light, and it is under this condition that the geometrical theory breaks down. 'Diffraction' has occurred, and this is, quintessentially, a wave phenomenon. The wave theory provides a more complete, but necessarily more complex, view of light propagation. We shall now deal with the phenomena of reflection and refraction using the wave theory, but we should

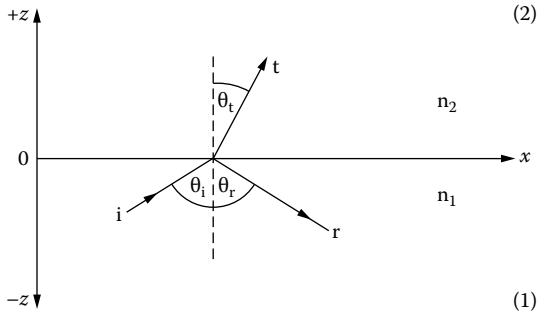


FIGURE 2.5 Reflection and refraction at a boundary between two media.

remember that under certain conditions (apertures much larger than the wavelength), the ray theory is useful for its simplicity: a wave can be replaced by a set of rays in the direction of propagation, normal to surfaces of constant phase, and obeying simple geometrical rules.

Let us consider two nonconducting dielectric media with refractive indices n_1 and n_2, separated by a plane boundary that we take to be the xy plane at $z = 0$ (Figure 2.5). Let us now consider a plane wave lying in the xz plane which is propagating in medium 1 and is incident on the boundary at angle υ_i, as shown in the figure. All the field components, such as (E_i, H_i), will vary as

$$(E_i, H_i)\exp\{i\omega[t - n_1(x\sin\vartheta_i + z\cos\vartheta_i)/c]\}$$

(see Figure 2.6) using the exponential forms of the wave, detailed in Section 2.3 and taking c to be the velocity of light in free space.

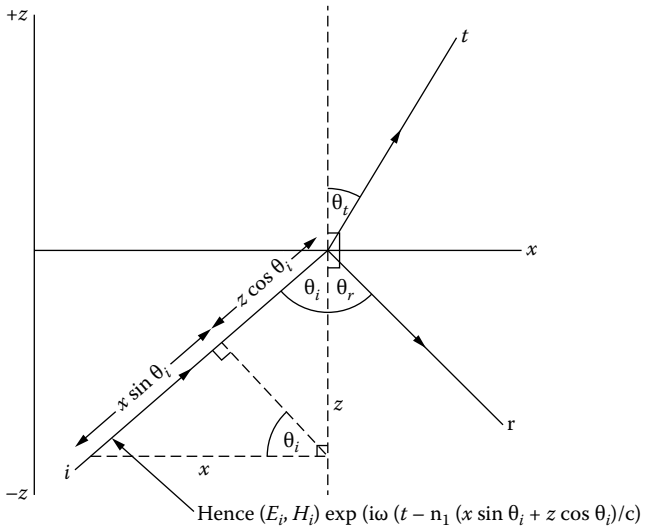


FIGURE 2.6 Trigonometry of the incident ray.

After striking the boundary, there will, in general, be a reflected and a refracted (transmitted, *t*) wave. This fact is a direct consequence of the boundary conditions that must be satisfied at the interface between the two media. These conditions follow from Maxwell's equations, and essentially may be stated as follows:

(i) Tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$ are continuous across the boundary.
(ii) Normal components of $\boldsymbol{B}$ and $\boldsymbol{D}$ are continuous across the boundary.

The above conditions must be true at all times and at all places on the boundary plane. They can only be true at all times at a given point if the frequencies of all the waves (i.e., incident, reflected, refracted) are the same; otherwise, clearly, amplitude discontinuities would occur across the boundary. Further, because the phase and amplitude of the incident wave must be constant on the boundary plane along any line for which x is constant (see Figure 2.7), it follows that the phases and amplitudes of the reflected and refracted waves must also be constant along such a line, if continuity in accordance with the boundary conditions is to be maintained. This is equivalent to saying that the reflected and refracted rays travel in the same direction and thus in the same plane (the xz plane) as the incident ray, which proves one of the previously stated laws of reflection and refraction.

To go further, it is necessary to give proper mathematical expression to the waves. Any given wave is, of course, a sinusoid, whose amplitude, frequency, and phase define the wave completely, and in Section 2.3 it was shown that the most convenient representation of such waves was via their complex exponential form.

Suppose (Figure 2.6) that the reflected and refracted waves make angles ϑ_r and ϑ_t, respectively, to the boundary in the xz plane. Then these waves will vary as:

$$\text{reflected: } \exp\{i\omega[t - n_1(x \sin\vartheta_r - z\cos\vartheta_r)/c]\}$$

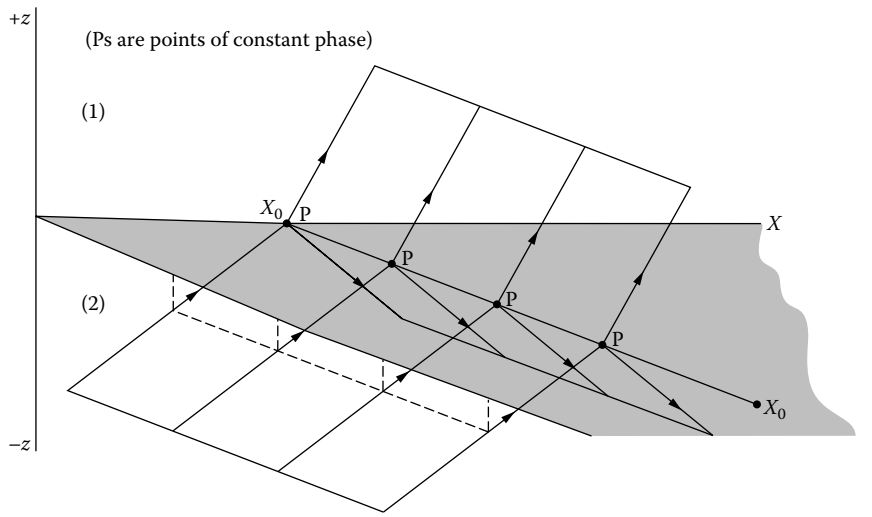


FIGURE 2.7 Line of constant phase in boundary plane.

(Note that the reflected ray travels in the *negative* z direction.)

$$\text{refracted: } \exp\{i\omega[t - n_2(x\sin\vartheta_t + z\cos\vartheta_t)/c]\}$$

whereas the incident wave, for reference, was

$$\text{incident: } \exp\{i\omega[t - n_1(x\sin\vartheta_i + z\cos\vartheta_i)/c]\}$$

At the boundary ($z = 0$), these variations must be identical for any x, t, if continuity is to be maintained; hence,

$$n_1 x\sin\vartheta_i = n_1 x\sin\vartheta_r = n_2 x\sin\vartheta_t$$

Thus, we have

$$\vartheta_i = \vartheta_r \text{ (law of reflection)}$$

$n_1 \sin\vartheta_i = n_2 \sin\vartheta_t$ (Snell's law of refraction)

We must now consider the relative amplitudes of the waves. To do this we match the components of $\boldsymbol{E}$, $\boldsymbol{H}$, $\boldsymbol{D}$, $\boldsymbol{B}$, separately. A further complication is that the values of these quantities at the boundary will depend on the direction of vibration of the $\boldsymbol{E}$, $\boldsymbol{H}$, fields of the incident wave, relative to the plane of the wave. Therefore, we need to consider two linear, orthogonal polarization components separately, one in the xz plane, the other normal to it. (Any other polarization state can be resolved into these two linear components, so that our solution will be complete.)

Let us consider the two stated linear components in turn:

(a) *$\boldsymbol{E}$ in the plane of incidence; $\boldsymbol{H}$ normal to the plane of incidence*

The incident wave can now be written in the following form (see Figure 2.6):

$$\begin{aligned} E_x^i &= -E_i\cos\vartheta_i \exp\{i\omega[t - n_1(x\sin\vartheta_i + z\cos\vartheta_i)/c]\} \\ E_z^i &= E_i\sin\vartheta_i \exp\{i\omega[t - n_1(x\sin\vartheta_i + z\cos\vartheta_i)/c]\} \\ H_y^i &= H_i \exp\{i\omega[t - n_1(x\sin\vartheta_i + z\cos\vartheta_i)/c]\} \end{aligned} \tag{2.7}$$

Now we can again enlist the help of Maxwell's equations to relate $\boldsymbol{H}$ and $\boldsymbol{E}$ for a plane wave (see Appendix I).

We have

$$\frac{E}{H} = Z = \left(\frac{\mu}{\varepsilon}\right)^{1/2}$$

Z is now known as the characteristic impedance of the medium. Because we are dealing, in this case, with nonconducting dielectrics, we have $\mu = 1$ and $n = \varepsilon^{1/2}$; hence,

$$Z = \frac{1}{n}$$

Thus,

$$H_i = nE_i \tag{2.8}$$

and the expression for H_y^i becomes

$$H_y^i = n_1 E_i \exp\{i\omega[t - n_1(x \sin \vartheta_i + z \cos \vartheta_i)/c]\}$$

Clearly we can construct similar sets of equations for the reflected and refracted waves. Having done this, we can impose the boundary conditions to obtain the required relationships between wave amplitudes. We shall now derive these relationships—that is, that between the reflected and incident electric field amplitudes, and that between the refracted and incident electric field amplitudes for this case.

We know that the exponential factors are all identical at the boundary if we are going to be able to satisfy the boundary conditions at all; let us, therefore, write the universal exponential factor as F.

For the incident (i) wave, from Equations (2.7), we have

$$\begin{aligned} E_x^i &= -E_i \cos \vartheta_i F \\ E_z^i &= E_i \sin \vartheta_i F \\ H_y^i &= H_i F \end{aligned} \tag{i}$$

For the reflected (r) wave,

$$\begin{aligned} E_x^r &= E_r \cos \vartheta_r F \\ E_z^r &= E_r \sin \vartheta_r F \\ H_y^r &= H_r F \end{aligned} \tag{r}$$

For the refracted (t) wave,

$$\begin{aligned} E_x^t &= -E_r \cos \vartheta_t F \\ E_z^t &= E_t \sin \vartheta_t F \\ H_y^t &= H_t F \end{aligned} \tag{t}$$

Imposing the condition that the tangential components (i.e., x components) of E must be continuous across the boundary, we have

$$E_x^i + E_x^r = E_x^t$$

or

$$-E_i \cos\vartheta_i + E_r \cos\vartheta_r = -E_t \cos\vartheta_t \tag{2.9}$$

using the appropriate equations from (*i*), (*r*), and (*t*) and cancelling the factor *F*.

Now doing the same for the tangential *H* field (*y* components),

$$H_i + H_r = H_t \tag{2.10}$$

We also know, from (2.8), that

$$H_i = n_1 E_i;\ H_r = n_1 E_r;\ H_t = n_2 E_t$$

hence, the *H* field condition (2.10) becomes

$$n_1 E_i + n_1 E_r = n_2 E_t \tag{2.11}$$

We may now eliminate E_t from (2.9) and (2.11) to obtain (remembering, also, that $\vartheta_r = \vartheta_i$)

$$\frac{E_r}{E_i} = \frac{n_2 \cos\vartheta_i - n_1 \cos\vartheta_t}{n_2 \cos\vartheta_i + n_1 \cos\vartheta_t} \tag{2.12a}$$

which is the required relationship.

Note also that, since, from Snell's law, $n_1 \sin\theta_i = n_2 \sin\theta_t$, this can be written

$$\frac{E_r}{E_i} = \frac{\tan(\vartheta_i - \vartheta_t)}{\tan(\vartheta_i + \vartheta_t)}$$

Similarly, we may eliminate E_r from (2.9) and (2.11) to obtain

$$\frac{E_t}{E_i} = \frac{2n_1 \cos\vartheta_i}{n_2 \cos\vartheta_i + n_1 \cos\vartheta_t} \tag{2.12b}$$

We must now consider the wave with the other, orthogonal, polarization:

(b)***E*** *normal to the plane of incidence*; ***H*** *in the plane of incidence*

Using the same methods as before, we obtain the relations

$$\frac{E'_r}{E'_i} = \frac{n_1 \cos\vartheta_i - n_2 \cos\vartheta_t}{n_1 \cos\vartheta_i + n_2 \cos\vartheta_t} \tag{2.12c}$$

$$\frac{E'_t}{E'_i} = \frac{2n_1 \cos\vartheta_i}{n_1 \cos\vartheta_i + n_2 \cos\vartheta_t} \tag{2.12d}$$

The above four expressions (2.12a through 2.12d) are known as *Fresnel's equations*; Fresnel derived them from the elastic-solid theory of light, which prevailed at his time. The equations contain several points worthy of emphasis.

First, we note that there is a possibility of eliminating the reflected wave. For E in the plane of incidence, we find from Equation (2.12a) that this occurs when

$$n_1 \cos \vartheta_t = n_2 \cos \vartheta_i$$

But from Snell's law, we also have

$$n_1 \sin \vartheta_i = n_2 \sin \vartheta_t$$

so that, combining the two relations,

$$\sin 2\vartheta_i = \sin 2\vartheta_t$$

Now, of course, this equation has an infinite number of solutions, but the only one of interest is that for which $\vartheta_i \neq \vartheta_t$ ($\vartheta_i = \vartheta_t$ only if $n_1 = n_2$) and for which both ϑ_i and ϑ_t lie in the range $0 \rightarrow \pi/2$. The required solution is

$$\vartheta_i + \vartheta_t = \frac{1}{2}\pi$$

and simple geometry then requires that the reflected and refracted rays are normal to each other (Figure 2.8). Clearly, from Snell's law, this occurs when

$$n_1 \sin \vartheta_i = n_2 \cos \vartheta_i$$

that is,

$$\tan \vartheta_i = \frac{n_2}{n_1}$$

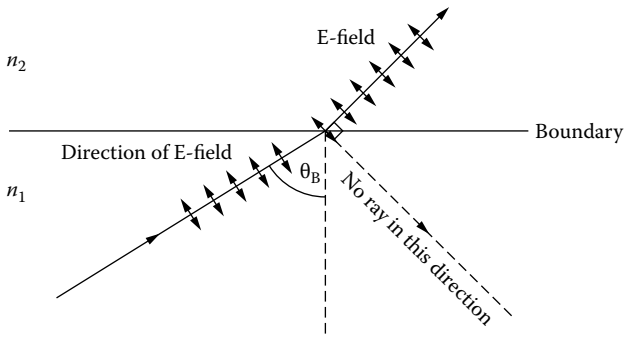


FIGURE 2.8 Elimination of the reflected ray at the Brewster angle (ϑ_B).

This particular value of ϑ_i is known as Brewster's angle (ϑ_B).

For example, for the glass/air boundary we find $\vartheta_B = 56.3°$.

It is instructive to understand the physical reason for the disappearance of the reflected ray at this angle when the electric field lies in the plane of incidence. Referring to Figure 2.8 we note that the incident wave sets up oscillations of the elementary dipoles in the second medium (see Chapter 4 for details), and at the Brewster angle, these oscillations take place in the direction of the reflected ray, because the refracted and reflected rays are orthogonal. Hence, these oscillations cannot generate any transverse waves in the required direction of reflection. Because light waves are, by their very nature, transverse, the reflected ray must be absent. If we ask the same question of the polarization that has $\boldsymbol{E}$ normal to the plane of incidence, we find from (2.12c):

$$n_1 \cos\vartheta_i = n_2 \cos\vartheta_t$$

which, with Snell's law, gives

$$\tan\vartheta_i = \tan\vartheta_t$$

There is no solution of this equation which satisfies the required conditions, so the reflected wave cannot be eliminated in this case. If, then, a wave of arbitrary polarization is incident on the boundary at the Brewster angle, only the polarization with $\boldsymbol{E}$ normal to the plane of incidence is reflected. This is a useful way of linearly polarizing a wave.

The second point worthy of emphasis is the condition at normal incidence. Here we have $\vartheta_i = \vartheta_r = \vartheta_t = 0$; hence, the relations, identical for both polarizations, become

$$\frac{E_r}{E_i} = \frac{E_r'}{E_i'} = \frac{n_1 - n_2}{n_1 + n_2} \tag{2.13a}$$

$$\frac{E_t}{E_i} = \frac{E_t'}{E_i'} = \frac{2n_1}{n_1 + n_2} \tag{2.13b}$$

Now the wave intensities are proportional to the squares of the electric field amplitudes *but only for a given medium*, because, from Equation (2.6c), the intensity is proportional to the refractive index as well as to the square of the field.

Hence, because the incident and reflected waves propagate in the same medium, it is appropriate to write

$$\frac{I_r}{I_i} = \frac{E_r^2}{E_i^2} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 \tag{2.13c}$$

but for the transmitted (refracted) wave, we have

$$\frac{I_t}{I_i} = \frac{n_2 E_t^2}{n_1 E_i^2} = \frac{4n_1 n_2}{(n_1 + n_2)^2} \tag{2.13d}$$

Note that now

$$I_r + I_t = I_i$$

so that energy is conserved, as required.

Equation (2.13c) and Equation (2.13d) are useful expressions, for they tell us how much light is lost by normal reflection when transmitting from one medium (say air) to another (say glass). For example, when passing through a glass lens (air → glass → air), taking the refractive index of the glass as 1.5 we find from (2.13c) that the fractional loss at the front face of the lens (assumed approximately normal) is

$$\frac{I_r}{I_i} = \frac{(0.5)^2}{(2.5)^2} = 0.04$$

Another 4% will be lost at the back face, giving a total 'Fresnel' loss of the order of 8%. This figure can be reduced by 'anti-reflection' coatings, which will be further discussed later (Section 10.2).

Finally, we should notice that all the expressions for the ratios of field amplitudes are mathematically real, and thus any change of phase that occurs at a boundary must be either 0 or π. We shall now look at a rather different type of reflection where this is not the case.

2.6 TOTAL INTERNAL REFLECTION

We return to Snell's law:

$$n_1 \sin\vartheta_i = n_2 \sin\vartheta_t$$

or

$$\sin\vartheta_t = \frac{n_1}{n_2}\sin\vartheta_i \tag{2.14}$$

The factor $\sin\vartheta_i$ is, of course, always less than unity. However, if $n_2 < n_1$ (i.e., the second medium is less optically dense than the first, which contains the incident ray), then it may be that

$$\sin\vartheta_i > \frac{n_2}{n_1}$$

that is,

$$\frac{n_1}{n_2}\sin\vartheta_i > 1$$

If this is so, then we have from Equation (2.14),

$$\sin\vartheta_t > 1 \tag{2.15}$$

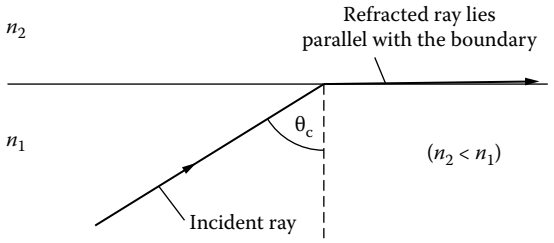


FIGURE 2.9 Critical angle (ϑ_c) for total internal reflection (TIR).

Equation (2.15) clearly cannot be satisfied for any real value of ϑ_t and there can be no real refracted ray. The explanation of this is that the refracted ray angle (ϑ_t), under these conditions of passage from a less dense to a more dense medium, is always greater than the incident angle (ϑ_i). Consequently, ϑ_t will reach a value of 90° (i.e., parallel to the boundary) before ϑ_i, and any greater value of ϑ_i cannot yield a refracted ray (Figure 2.9). The value of ϑ_i for which (2.15) just becomes true we define as the critical angle, ϑ_c:

$$\sin \vartheta_c = \frac{n_2}{n_1}$$

For all values of $\vartheta_i > \vartheta_c$, the light is totally reflected at the boundary: the phenomenon is called total internal reflection (TIR). However, Fresnel's equations must still apply, for we made no limitations on the values of the quantities when imposing the boundary conditions. Furthermore, if the fields are to be continuous across the boundary, as required by Maxwell's equations, then there must be a field disturbance of some kind in the second medium. We can use Fresnel's equations to investigate this disturbance.

We write

$$\cos \vartheta_t = (1 - \sin^2 \vartheta_t)^{1/2} \tag{2.16}$$

Since $\sin \vartheta_t > 1$ for $\vartheta_t > \vartheta_c$, and since also the function $\cosh \gamma \geq 1$ for all real γ, we may, for convenience, use the substitution

$$\sin \vartheta_t = \cosh \gamma (\vartheta_i > \vartheta_c)$$

and henceforth, therefore, the TIR condition (2.15) is, implicitly, imposed. We now have, from (2.16),

$$\cos \vartheta_t = i\,(\cosh^2 \gamma - 1)^{1/2} = \pm i \sinh \gamma$$

Hence, we may write the field components in the second medium to vary as

$$\exp\{i\omega[t - n_2 (x \cosh \gamma - iz \sinh \gamma)/c]\}$$

or

$$\exp[(-\omega n_2 z \sinh\gamma)/c]\exp[i\omega\ (t - n_2 x\cosh\gamma)/c]$$

where we have used the fact that $\cosh\gamma = \frac{1}{2}(e^{\gamma} + e^{-\gamma})$, which tends to infinity as $\gamma \to \pm\infty$, and has a minimum of 1 at $\gamma = 0$.

This represents a wave travelling in the Ox direction in the second medium (i.e., parallel to the boundary) with amplitude decreasing exponentially in the Oz direction (at right angles to the boundary). The rate at which the amplitude falls with z can be written

$$\exp[(-2\pi z \sinh\ \gamma)/\gamma_2]$$

or, in terms of the original parameters,

$$\exp[-k_2 z(n_1^2 \sin^2\vartheta_i - n_2^2)^{1/2}/n_2]$$

with γ_2 being the wavelength of the light and k_2 the wavenumber, in the second medium. This shows that the wave is attenuated significantly over distances ~ γ_2. For example, at the glass/air interface, the critical angle will be ~$\sin^{-1}$ (1/1.5)—that is, ~42°. For a wave in the glass incident on the glass/air boundary at 60° ($\vartheta_i > \vartheta_c$), we find that $\sinh\gamma = 1.64$. Hence, the amplitude of the wave in the second medium is reduced by a factor of 5.4×10^{-3} in a distance of only one wavelength, the latter being of order 1μm. The wave is called an 'evanescent' wave. Even though the evanescent wave is propagating in the second medium, it transports no light energy in a direction normal to the boundary. All the light is totally internally reflected at the boundary. The fields that exist in the second medium give a Poynting vector that averages to zero in this direction, over one oscillation period of the light wave. All the energy in the evanescent wave is transported parallel to the boundary between the two media. The totally internally reflected wave now suffers a phase change that depends both on the angle of incidence and on the polarization. This can readily be derived from Fresnel's equations. Taking Equation (2.12a), we have the following for the TIR case where E lies in the plane of incidence:

$$\frac{E_r}{E_i} = \frac{n_2\cos\vartheta_i - in_1\sinh\gamma}{n_2\cos\vartheta_i + in_1\sinh\gamma}$$

This complex number provides the phase change on TIR as δ_p where

$$(E_{para}):\quad \tan\left(\frac{1}{2}\delta_p\right) = \frac{n_1\left(n_1^2\sin^2\vartheta_i - n_2^2\right)^{1/2}}{n_2^2\cos\vartheta_i}$$

and for the perpendicular E polarization:

$$(E_{perp}):\quad \tan\left(\frac{1}{2}\delta_s\right) = \frac{\left(n_1^2\sin^2\vartheta_i - n_2^2\right)^{1/2}}{n_1\cos\vartheta_i}$$

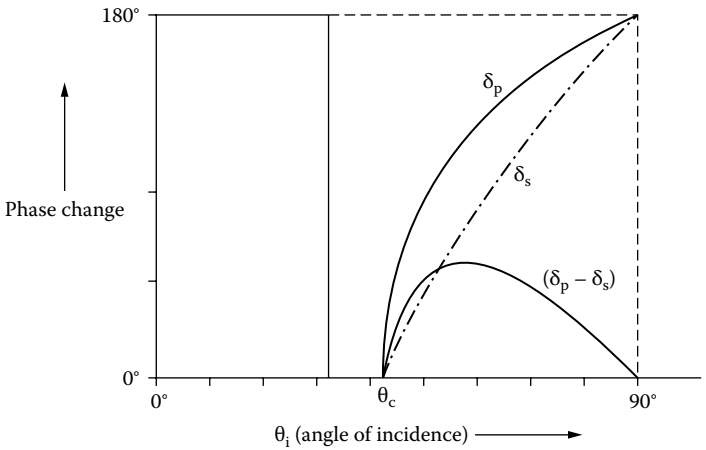


FIGURE 2.10 Phase changes on total internal reflection.

We note also that

$$\tan\left(\frac{1}{2}\delta_p\right) = n_1^2 \tan\left(\frac{1}{2}\delta_s\right)$$

and that

$$\tan\left[\frac{1}{2}(\delta_p - \delta_s)\right] = \frac{\cos\vartheta_i \left(n_1^2 \sin^2\vartheta_i - n_2^2\right)^{1/2}}{n_1 \sin^2\vartheta_i}$$

The variations δ_p, δ_s and $\delta_p - \delta_s$ are shown in Figure 2.10 as a function of δ_i. It is clear that the polarization state of light undergoing TIR will be changed as a result of the differential phase change $\delta_p - \delta_s$. By choosing ϑ_i appropriately and, perhaps, using two TIRs, it is possible to produce any wanted, final polarization state from any given initial state.

It is interesting to note that the reflected ray in TIR appears to originate from a point that is displaced along the boundary from the point of incidence. This is consistent with the incident ray being reflected from a parallel plane that lies a short distance within the second boundary (Figure 2.11). This view is also consistent with the observed phase

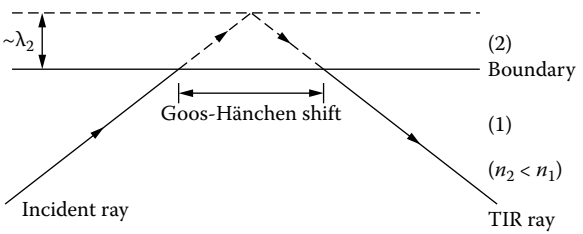


FIGURE 2.11 The Goos-Hänchen shift on total internal reflection.

shift, which now is regarded as being due to the extra optical path travelled by the ray. The displacement is known as the Goos-Hänchen effect and provides an entirely consistent alternative explanation of TIR. This provides food for further interesting thoughts, which we shall not pursue because they are somewhat beyond the scope of this book.

2.7 INTERFERENCE OF LIGHT

We have seen that light consists of oscillating electric and magnetic fields. We know that these fields are vector fields because they represent forces (on unit charge and unit magnetic pole, respectively). The fields will thus add vectorially. Consequently, when two light waves are superimposed on each other, we obtain the resultant by constructing their vector sum at each point in time and space, and this fact has already been used in consideration of the polarization of light (Section 2.4.3).

If two sinusoids are added, the result is another sinusoid. Suppose that two light waves given, via their electric fields, as

$$e_1 = E_1 \cos(\omega t + \varphi_1)$$

$$e_2 = E_2 \cos(\omega t + \varphi_2)$$

have the same polarization and are superimposed at a point in space. We know that the resultant field at the point will be given, using elementary trigonometry or by the complex exponential methods described in Section 2.3, by

$$e_t = E_T \cos(\omega t + \varphi_T)$$

where

$$E_T^2 = E_1^2 + E_2^2 + 2E_1E_2 \cos(\varphi_2 - \varphi_1)$$

and

$$\tan \vartheta_T = \frac{E_1 \sin \varphi_1 + E_2 \sin \varphi_2}{E_1 \cos \varphi_1 + E_2 \cos \varphi_2}$$

For the important case where $E_1 = E_2 = E$, say, we have

$$E_T^2 = 4E^2 \cos^2 \frac{1}{2}(\varphi_2 - \varphi_1) \tag{2.17}$$

and

$$\tan \phi_T = \tan \frac{1}{2}(\varphi_2 + \varphi_1)$$

The intensity of the wave will be proportional to E_T^2 so that, from Equation (2.17) it can be seen to vary from $4E^2$ to 0, as $(\varphi_2 - \varphi_1)/2$ varies from 0 to $\pi/2$.

Consider now the arrangement shown in Figure 2.12. Here two slits, separated by a distance p, are illuminated by a plane wave with wavelength λ. The portions of the wave which pass through the slits will interfere on the screen S, a distance d away.

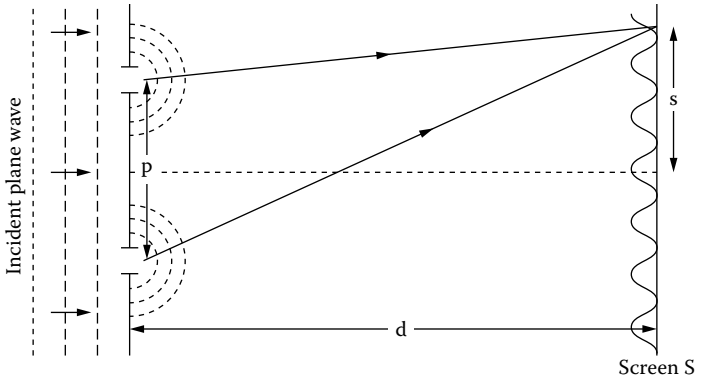


FIGURE 2.12 'Young's slits' interference.

Now each of the slits will act as a source of cylindrical waves, from Huygens' principle. Moreover, because they originate from the same plane wave, they will start in phase. On a line displaced a distance s from the line of symmetry on the screen, the waves from the two slits will differ in phase by

$$\delta = \frac{2\pi}{\lambda}\frac{sp}{d} \quad (d >> s, p)$$

Thus, as s increases, the intensity will vary between a maximum and zero, in accordance with Equation (2.17). These variations will be viewed as fringes (i.e., lines of constant intensity parallel with the slits). They are known as Young's fringes, after their discoverer, and are the simplest example of light interference. We shall now consider some important (and more complex) examples of light interference in action.

2.8 LIGHT WAVEGUIDING

Consider, first, the symmetrical dielectric structure shown in Figure 2.13. Here we have an infinite (in width and length) dielectric slab of refractive index n_1, sandwiched between two other infinite slabs each of refractive index n_2.

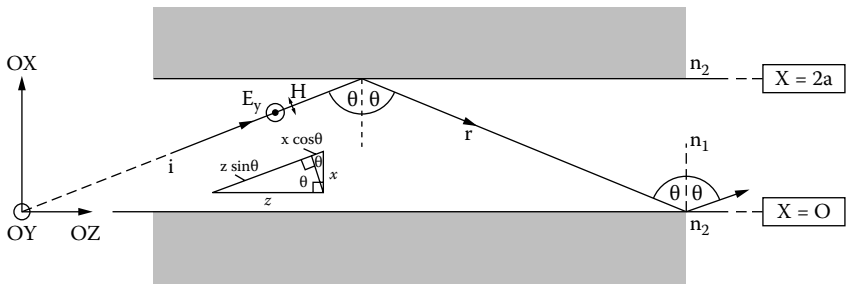


FIGURE 2.13 The dielectric-slab waveguide.

Using the Cartesian axes defined in Figure 2.13, let us consider a light ray starting at the origin of axes and propagating within the first medium at an angle ϑ. If ϑ is greater than the critical angle (ϑ_c), the light will bounce down the first medium by means of a series of total internal reflections at the boundaries with the other media. Because the wave is thus confined to the first medium, it is said to be 'guided' by the structure, which is consequently called a 'waveguide'. Let us, first, consider guided light that is linearly polarized normal to the plane of incidence. The electric field of the wave represented by ray i (see Figure 2.13) can be written

$$E_i = E_0 \exp\left(i\omega t - kn_1 x \cos\vartheta - ikn_1 z \sin\vartheta\right)$$

whilst that represented by *r*, the ray reflected from the first boundary, can be written

$$E_r = E_0 \exp\left(i\omega t + kn_1 x \cos\vartheta - ikn_1 z \sin\vartheta + i\delta_s\right)$$

where δ_s is the phase change at TIR for this polarization. These two waves will be superimposed on each other and will thus interfere. The interference pattern is obtained by adding them:

$$E_T = E_i + E_r = E_0 \exp\left(i\omega t - ikn_1 z \sin\vartheta + i\frac{1}{2}\delta_s\right) 2\cos\left(kn_1 x \cos\vartheta + \frac{1}{2}\delta_s\right) \quad (2.18)$$

This is a wave propagating in the direction Oz with wavenumber $kn_1 \sin\vartheta$, and it is amplitude-modulated in the Ox direction according to

$$\cos\left(kn_1 x \cos\vartheta + \frac{1}{2}\delta_s\right)$$

Now if the wave propagating in the Oz direction is to be a stable, symmetrical entity resulting from a self-reproducing interference pattern, the intensity of the wave must be the same at each of the two boundaries. This requires that it is the same for $x = 0$ as for $x = 2a$. That is,

$$\cos^2\left(\frac{1}{2}\delta_s\right) = \cos^2\left(kn_1 2a\cos\vartheta + \frac{1}{2}\delta_s\right) \quad (2.19)$$

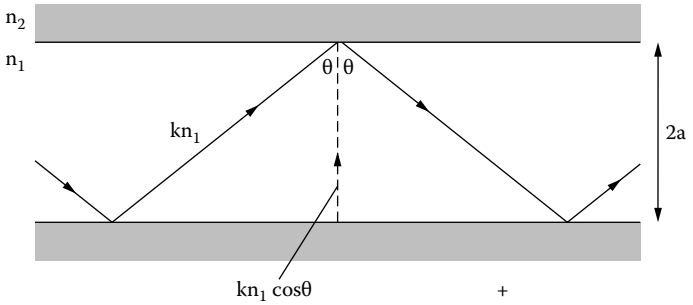
The general solution of this equation is

$$\frac{1}{2}\delta_s = m\pi \pm \left(2akn_1 \cos\vartheta + \frac{1}{2}\delta_s\right)$$

where *m* is any integer (positive or negative).

Hence, either

$$2akn_1 \cos\vartheta + \delta_s = m\pi \,(-) \quad (2.20a)$$



Transverse wave number = $kn_1 \cos\theta$
∴ Phase change across guide, width $2a = 2akn_1 \cos\theta$
Phase change on reflection = δ_s
∴ $2akn_1 \cos\theta + \delta_s = m\pi$, for reinforcement

FIGURE 2.14 Transverse resonance condition.

or

$$2akn_1 \cos \vartheta = -m\pi (+) \tag{2.20b}$$

However, there is another condition to impose. If the interference pattern is to self-reproduce in a consistent way as it propagates down the guide, the phase change experienced by a ray executing one complete 'bounce' down the guide must be an integer times 2π. If this were not so, the waves would not retain mutual phase coherence and the interference pattern would self-destruct. This can be seen from the geometry in Figure 2.14. The wavefronts resulting from ray reflections at all points along the guide can only be in phase provided that

$$2akn_1 \cos \vartheta + \delta_s = m\pi$$

which corresponds to Equation (2.20a). Equation (2.20b) does not satisfy the condition on wavefronts and is, therefore, invalid. Equation (2.20a) is sometimes known as the 'transverse resonance condition' because it corresponds essentially to the condition that, when resolving the wave vector into directions transverse and parallel to the guide axis, the transverse component has just one half cycle, or an integer multiple thereof ($m\pi$), fitting into the guide width. This is a 'resonance' in the sense that a string stretched between two points resonates, when plucked, at frequencies conditioned in just the same way.

Now because δ_s depends only on ϑ (see Fresnel's equations in Section 2.5), it follows that the condition

$$2akn_1 \cos \vartheta + \delta_s = m\pi$$

is a condition on ϑ. The condition tells us that ϑ can have only certain discrete values if the interference pattern is to remain constant in form along the length of the fibre.

Each form of interference pattern is, therefore, characterized by a particular value of m which then provides a corresponding value for ϑ. The allowed interference patterns are called the 'modes' of the waveguide, for they are determined by the properties (geometrical and physical) of the guide.

If we now turn to the progression of the wave along the guide (i.e., along the Oz axis), we see from Equation (2.18) that this is characterized by a wavenumber of value

$$n_1 k \sin \vartheta = \beta(\text{say})$$

Furthermore, because the TIR condition requires that

$$\sin \vartheta \geq \frac{n_2}{n_1}$$

it follows that

$$n_1 k \geq \beta \geq n_2 k$$

so that the longitudinal wavenumber always lies between those of the two guiding media.

Thus we see that waveguiding essentially is a wave interference phenomenon, and we shall leave the subject there for the moment. The subject is an extremely important one and there are many other aspects to be considered. Consequently, we shall return to it in more detail in Chapter 8.

2.9 INTERFEROMETERS

In Section 2.7, the essentials of dual-beam interference were discussed. Although very simple in concept, the phenomenon is extremely useful in practice. The reason for this is that the maxima of the resulting fringe pattern appear where the phase difference between the interfering light beams is a multiple of 2π. Any quite small perturbation in the phase of one of the beams will thus cause a transverse shift in the position of the fringe pattern, which, using photonic techniques, is readily observed to about 10^{-4} of the fringe spacing. Such a shift is caused by, for example, an increase in path length of one of the beams by one hundredth of a wavelength, or about 5×10^{-9} m for visible light. This means that differential distances of this order can be measured, leading to obvious applications in, for example, strain monitoring on mechanical structures.

Another example of a dual-beam interferometer is shown in Figure 2.15. Here the beams are produced from the partial reflection and transmission at a dielectric, or partially silvered, mirror M_1. Another such mirror, M_4, recombines the two beams after their separate passages. Such an arrangement is known as a Mach-Zehnder interferometer and is used extensively to monitor changes in the phase differences between two optical paths. An optical-fibre version of a Mach-Zehnder interferometer is shown in Figure 2.16. In this case the 'mirrors' are optical couplings

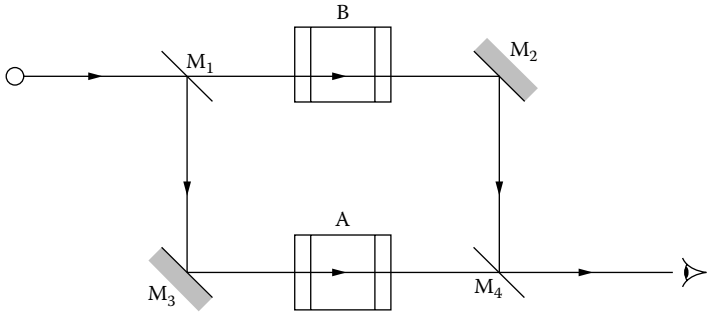


FIGURE 2.15 Basic Mach-Zehnder interferometer.

between the cores of the two fibres. The 'fringe pattern' consists effectively of just one fringe, because the fibre core acts as an efficient spatial filter. But the light that emerges from the fibre end (E) clearly will depend on the phase relationship between the two optical paths when the light beams recombine at R, and thus it will depend critically on propagation conditions within the two arms. If one of the arms varies in temperature, strain, or density compared with the other, then the light output will also vary. Hence, the latter can be used as a sensitive measure of any physical parameters capable of modifying the phase propagation properties of the fibre.

Finally, Figure 2.17a shows another, rather more sophisticated variation of the Mach-Zehnder idea. In this case the beams are again separated by means of a beam-splitting mirror, but are returned to the same point by fully silvered mirrors placed at the ends of the two respective optical paths. (The plate P is necessary to provide equal optical paths for the two beams in the absence of any perturbation.) This arrangement is called the Michelson interferometer, after the experimenter who in the late nineteenth century used optical interferometry with great skill to make many physical advances. His interferometer (not to be confused with his 'stellar' interferometer, which will be discussed later) allows for a greater accuracy of fine adjustment via control of the reflecting mirrors, but uses, of course, just the same basic

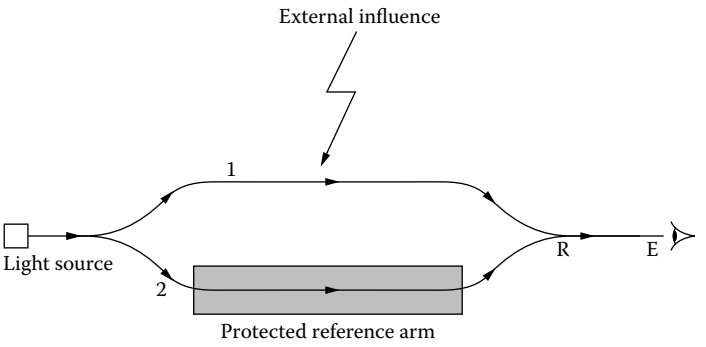
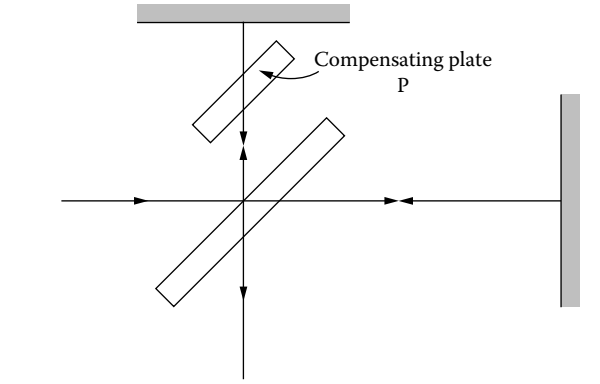
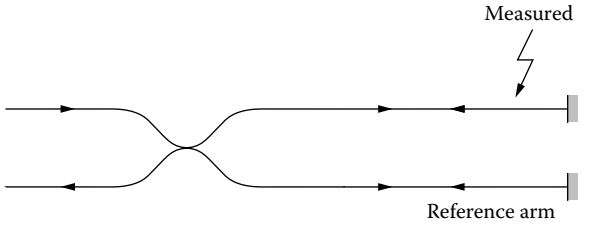


FIGURE 2.16 An optical-fibre Mach-Zehnder interferometer.



(a) Bulk version



(b) Optical-fibre version

FIGURE 2.17 Michelson interferometers.

interferometric principles as before. The optical-fibre version of this device is shown in Figure 2.17(b).

For completeness, and because of its historical importance, mention must be made of the use of Michelson's interferometer in the famous Michelson-Morley experiment of 1887. This demonstrated that light travelled with the same velocity in each of two orthogonal paths, no matter what was the orientation of the interferometer with respect to the earth's proper motion through space. This result was crucial to Einstein's formulation of special relativity in 1905, and thus is certainly one of the most important results in the history of experimental physics.

Valuable as dual-beam interferometry is, it suffers from the limitation that its accuracy depends upon the location of the maxima (or minima) of a sinusoidal variation. For very accurate work, such as precision spectroscopy, this limitation is severe. By using the interference amongst many beams, rather than just two, we find that we can improve the accuracy considerably. We can see this by considering the arrangement of Figure 2.18. Light from a single source gives a large number of phase-related, separate beams by means of multiple reflections and transmissions within a dielectric (e.g., glass) plate. For a given angle of incidence (ϑ) there will be fixed values for the transmission (T,T') and reflection (R) coefficients, as shown. If we start with a wave of amplitude a, the waves on successive reflections will suffer attenuation by a constant factor and will increase in phase by a constant amount. If

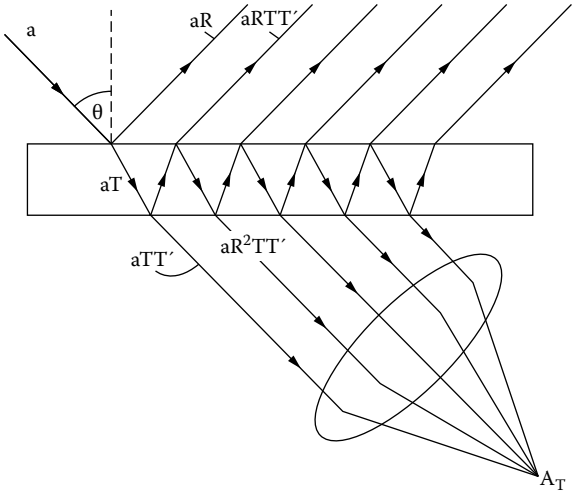


FIGURE 2.18 Multiple interference.

we consider the transmitted light only, then the total amplitude that arrives at the focus of the lens L is given by the sum

$$A_T = aTT' \exp(i\omega t) + aTT'R^2 \exp(i\omega t - iks) + aTT'R^4 \exp(i\omega t - 2iks) + \cdots\cdots$$

where s is the optical path difference between successive reflections at the lower surface (including the phase changes on reflection and transmission). The sum can be expressed as

$$A_T = aTT' \sum_{p=0}^{\infty} R^{2p} \exp(i\omega t - ipks)$$

which is a geometric series whose sum value is

$$A_T = \frac{aTT' \exp(i\omega t)}{1 - R^2 \exp(-iks)}$$

Hence, the intensity I of the light is proportional to $\left|A_T\right|^2$, that is,

$$I \propto \left|A_T\right|^2 = \frac{(aTT')^2}{1 + R^4 - 2R^2 \cos ks} \tag{2.21}$$

We note from this equation that the ratio of maximum and minimum intensities

$$\frac{I_{max}}{I_{min}} = \frac{(1+R^2)^2}{(1-R^2)^2}$$

so that the fringe contrast increases with R. However, as R increases, so does the attenuation between the successive reflections. Hence, the total transmitted light power will fall.

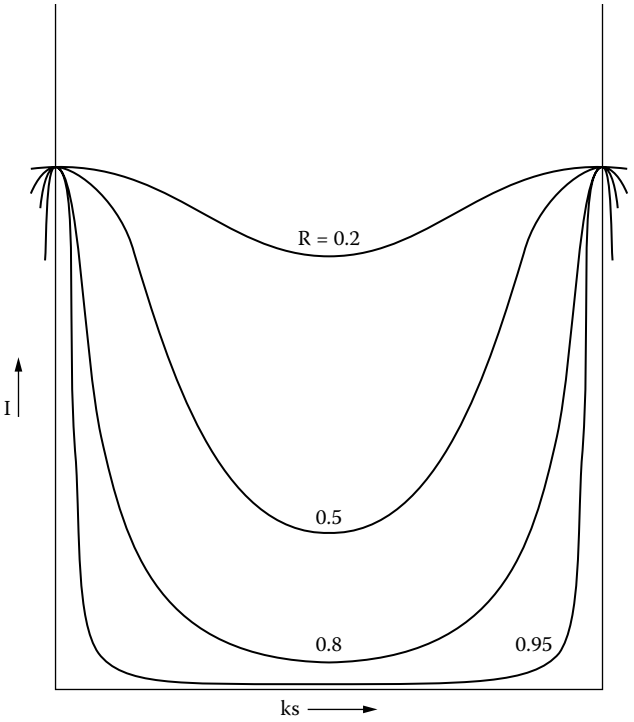


FIGURE 2.19 Variation of intensity with optical path, for various reflectivities, in a multiple interference plate.

Figure 2.19 shows how I varies with ks for different values of R. We note that the fringes become very sharp for large values of R. Hence, the position of the maxima may now be accurately determined. Further, because the spacing of the maxima specifies ks, this information can be used to determine either k or s, if the other is known. Consequently, multiple interference may be used either to select (or measure) a specific wavelength, or to measure very small changes in optical path length.

The physical reason for the sharpening of the fringes as the reflectivity increases is indicated in Figure 2.20. The addition of the multiplicity of waves is equivalent to the addition of vectors with progressively decreasing amplitude, and increasing relative phase. For small reflectivity (Figure 2.20a), the wave amplitudes decrease rapidly, so that the phase increase has a relatively small effect on the resultant wave amplitude. In the case of high reflectivity (Figure 2.20b), the reverse is the case and a small successive phase change rapidly reduces the resultant.

Two important devices based on these ideas of multiple reflection are the Fabry-Perot interferometer and the Fabry-Perot etalon. In the former case, the distance between the two surfaces is finely variable for fringe control; in the case of the etalon, the surfaces are fixed. In both cases, the flatness and parallelism of the surfaces must be accurate to $\sim\lambda/100$ for good quality fringes. This is difficult to achieve in a variable device, and the etalon is preferred for most practical purposes.

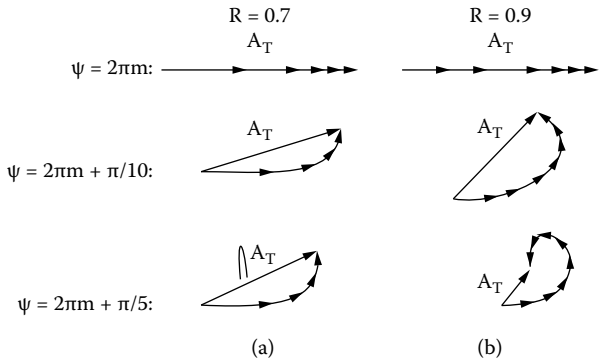


FIGURE 2.20 The dependence of fringe sharpness on reflectivity (R).

The Fabry-Perot interferometer is extremely important in photonics. We already noted its wavelength selectivity, but we should also note its ability to store optical energy by continually bouncing light between two parallel mirrors. For this reason it is often called a Fabry-Perot 'cavity' and is, roughly speaking, the optical equivalent of an electronic oscillator. The optical term is 'resonator', and it is this property that makes it an integral feature in all lasers.

Because of its importance, we must further understand in more detail the parameters that characterize the performance of the Fabry-Perot resonator: there are three main parameters.

These parameters relate, as is to be expected, to the instrument's ability to separate closely spaced optical wavelengths. The first is a measure of the sharpness of the fringes. This measure is normalized to the separation of the fringes for a single wavelength, because, clearly, there is no advantage in having narrow fringes if they are all crowded together, so that the orders of different wavelengths overlap. We hence define a quantity:

$$\Phi = \frac{\text{separation of successive fringes}}{\text{width at half maximum of a single fringe}}$$

which is called the 'finesse' and is roughly equivalent to the Q ('quality' factor measuring the sharpness of the resonance) of an electronic oscillator.

It is easy to derive an expression for Φ from Equation (2.21) as follows.

The equation may be written in the following form:

$$I = \frac{I_{max}}{1 + F \sin^2\left(\frac{1}{2}\Psi\right)} \tag{2.22}$$

where

$$F = \frac{4R^2}{(1-R^2)^2}$$

and

$$\psi = ks$$

From this it is clear that $I = I_{max}/2$ when

$$\Psi_h = \frac{2}{\sqrt{F}}$$

Hence, the width at half maximum = $2\Psi_h = 4/\sqrt{F}$.

The 'ψ distance' between successive maxima is just 2π, and thus the finesse is given by

$$\Phi = \frac{2\pi}{2\Psi_h} = \frac{\pi\sqrt{F}}{2} = \frac{\pi R}{(1-R^2)}$$

This quantity has a value of 2 for a dual-beam interferometer. For a Fabry-Perot etalon with $R = 0.9$, its value is 15. Clearly, the higher the value, the sharper are the fringes for a given fringe separation and the more wavelength selective is the device.

The next quantity we need to look at is the resolving power. This is a measure of the smallest detectable wavelength separation $\delta\lambda$ at a given wavelength λ and is defined as

$$\rho = \frac{\lambda}{\delta\lambda}$$

If we take λ to be that which corresponds to a ψ difference equal to the width of the half maximum, we have

$$\rho = \frac{\lambda}{\delta\lambda} = \frac{\Psi}{2\Psi_h} = \frac{2\pi p}{4/\sqrt{F}} = \frac{\pi p}{2}\sqrt{F} = p \times \text{finesse}$$

That is,

$$\rho = p\Phi$$

where p is the 'order' of the maximum (i.e., the number of maxima from the one at $\psi = 0$). If the etalon is being viewed close to normal incidence, then p will be effectively just the number of wavelengths in a double passage across the etalon. If the etalon has optical thickness t, we have $p = 2t/\lambda$ and thus

$$\rho = \pi t \frac{\sqrt{F}}{\lambda}$$

This is typically of the order of 10^6, compared with a figure ~10^4 for a dual-beam interferometer such as the Michelson. The ratio of these figures thus represents the improvement in accuracy afforded by multiple-beam interferometry over dual-beam techniques.

Finally, we define a quantity concerned with the overlapping of orders. If the range of wavelengths ($\Delta\gamma$) under investigation is such that the $(p + 1)$th maximum of γ is to coincide with the pth maximum of $(\gamma + \Delta\gamma)$, then, clearly, there is an unresolvable confusion. For this just to be so,

$$(p + 1)\, k = p\, (k + \Delta k)$$

so that

$$\frac{\Delta k}{k} = \frac{\Delta\lambda}{\lambda} = \frac{1}{p}$$

Again, close to normal incidence, we may write, with $p = 2t/\gamma$:

$$\Delta\lambda = \frac{\lambda}{p} = \frac{\lambda^2}{2t}$$

$\Delta\gamma$ is called the 'free spectral range' of the etalon and represents the maximum usable wavelength range without recourse to prior separation of the confusable wavelengths.

We shall need to return to the Fabry-Perot interferometer later on.

2.10 DIFFRACTION

In Section 2.5 it was noted that each point on a wavefront could be regarded formally and rigorously as a source of spherical waves. In Section 2.7 it was noted that any two waves, when superimposed, will interfere. Consequently, wavefronts can interfere with themselves and with other, separate, wavefronts. To the former usually is attached the name 'diffraction' and to the latter 'interference', but the distinction is somewhat arbitrary and, in several cases, far from clear cut.

Diffraction of light may be regarded as the limiting case of multiple interference as the source spacings become infinitesimally small. Consider the slit aperture in Figure 2.21. This slit is illuminated with a uniform plane wave and the light that passes through the slit is observed on a screen that is sufficiently distant from the slit for the light which falls upon it to be effectively, again, a plane wave. These are the conditions for Fraunhofer diffraction. If source and screen are close enough to the slit for the waves not to be plane, we have a more complex situation, known as Fresnel diffraction. Fraunhofer diffraction is by far the more important of the two and is the only form of diffraction we shall deal with here. Fresnel diffraction usually can be transformed into Fraunhofer diffraction, in any case, by the use of lenses that render the waves effectively plane, even over short distances.

Suppose that in Figure 2.21 the amplitude of the wave at distances between x and $x + \mathrm{d}x$ along the slit is given by the complex quantity $f(x)\mathrm{d}x$, and consider the effect of this at angle ϑ, as shown. (Because each point on the wavefront acts as a source of spherical waves, all angles will, of course, be illuminated by the strip.) The screen, being effectively infinitely distant from the slit, will be illuminated at one point by the light leaving the slit at angles between ϑ and $\vartheta + \mathrm{d}\vartheta$. Taking the bottom of the slit as the phase reference, the light, on arriving at the screen, will lead by a phase:

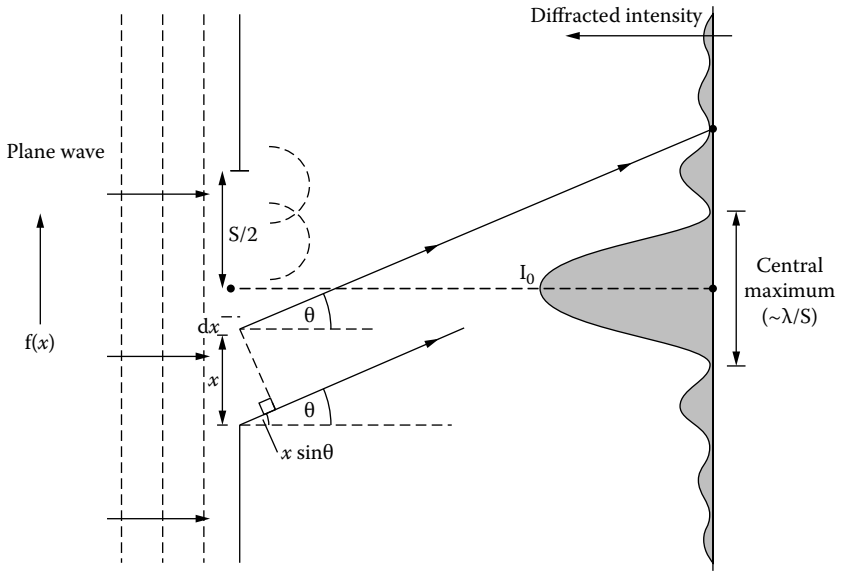


FIGURE 2.21 Diffraction at a slit.

$$\Phi = kx \sin \vartheta$$

and hence, the total amplitude in directions ϑ to $\vartheta + \mathrm{d}\vartheta$ will be given by

$$A(\vartheta) = \int_{-\infty}^{\infty} f(x) \exp(-ikx \sin \vartheta) dx$$

We can also write

$$A(\alpha) = \int_{-\infty}^{\infty} f(x) \exp(-i\alpha x) dx$$

with

$$\alpha = k \sin \vartheta$$

Hence, $A(\alpha)$ and $f(x)$ constitute a reciprocal Fourier transform pair (see Appendix II)—that is, each is the Fourier transform of the other. This is an important result. For small values of ϑ it implies that the angular distribution of the diffracted light is the Fourier transform of the aperture's amplitude distribution.

Let us see how this works for some simple cases. Take first a uniformly illuminated slit of width s. The angular distribution of the diffracted light will now be

$$A(k \sin \vartheta) = \int_{-s/2}^{s/2} a \exp(-ikx \sin \vartheta) dx$$

where a is the (uniform) amplitude at the slit per unit of slit width.

Hence,

$$A(k \sin \vartheta) = a \frac{\sin\left(\frac{1}{2} ks \sin \vartheta\right)}{\left(\frac{1}{2} k \sin \vartheta\right)}$$

Writing, for convenience, $\beta = \frac{1}{2} ks \sin \vartheta$, we find that the intensity in a direction ϑ is given by

$$I(\vartheta) = (as)^2 \frac{\sin^2 \beta}{\beta^2} = I_0 \frac{\sin^2 \beta}{\beta^2} \tag{2.23}$$

where I_0 is intensity at the centre of the diffraction pattern. This variation is shown in Figure 2.21, and as in the case of multiple interference between discrete sources, its shape is a result of the addition of wave vectors with phase increasing steadily with ϑ.

This form of variation occurs frequently in physics across a broad range of applications, and it is instructive to understand why. The function appropriate to the variation is given the name 'sinc' (pronounced 'sink')—that is,

$$\sin c\, \beta = \frac{\sin \beta}{\beta}$$

$$\sin c^2 \beta = \frac{\sin^2 \beta}{\beta^2}$$

Let us examine the physical reason for the sinc function in the case we have been considering—a uniformly illuminated slit. In this case, each infinitesimal element of the slit provides a wave amplitude $a\mathrm{d}x$ and at the centre of the screen all of these elements are in phase, producing a total amplitude, as. Hence, it is possible to represent all these elementary vectors as a straight line (because they are all in phase) of length as (Figure 2.22a). Now consider the situation at angle ϑ to the axis. As already shown, the ray from the bottom of the slit lags that from the top by a phase:

$$\varphi_T = ks \sin \vartheta = 2\beta$$

The result can, therefore, be depicted as in Figure 2.22b. The first and last infinitesimal vectors are inclined at 2β to each other, and the intervening vectors form an arc of a circle which subtends 2β at the circle's centre. The vector addition of all the vectors thus leads to a resultant that is the chord across the arc in Figure 2.22b. Simple geometry gives the length of this chord as

$$A(\theta) = 2r \sin \beta \tag{2.24}$$

where r is the radius of the circle. Now the total length of the arc is the same as that of the straight line when all vectors were in phase (i.e., as); hence,

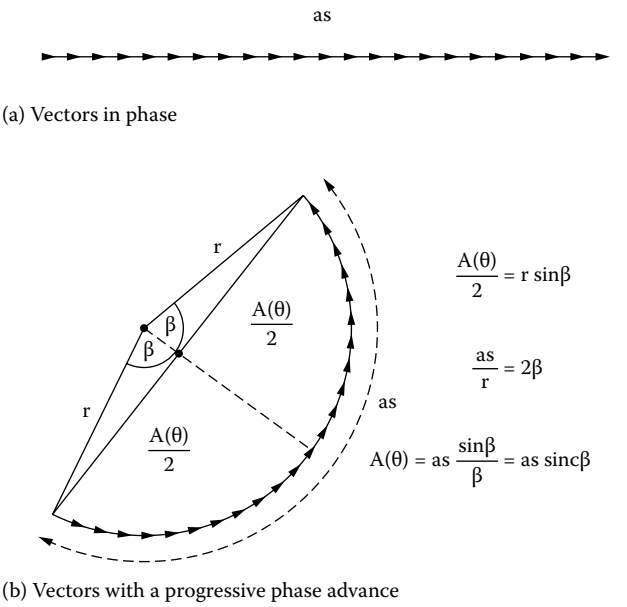


FIGURE 2.22 The 'sinc' function.

$$\frac{as}{t} = 2\beta$$

and thus, substituting for r in Equation (2.24), we have

$$A(\vartheta) = as\frac{\sin\beta}{\beta}$$

Hence, the resultant intensity at angle ϑ will be

$$I(\vartheta) = (as)^2\frac{\sin^2\beta}{\beta^2} = I_0\frac{\sin^2\beta}{\beta^2}$$

as in Equation (2.23).

The reason for the ubiquity of this variation in physics can now be seen to be due to the fact that one very often encounters situations where there is a systematically increasing phase difference amongst a large number of infinitesimal vector quantities: optical interference, electron interference, mass spectrometer energies, particle scattering, and so forth.

The principles that lead to the sinc function are the same and are those that have just been described.

Let us return now to the intensity diffraction pattern for a slit:

$$I(\vartheta) = I_0\frac{\sin^2\beta}{\beta^2}$$

An important feature of this variation is the scale of the angular divergence. The two minima immediately on either side of the principal maximum (at $\vartheta = 0$) occur when

$$\beta = \frac{1}{2}(ks)\sin\vartheta = \pm\pi$$

giving

$$\sin\vartheta = \pm\frac{\lambda}{s}$$

so that if ϑ is small, the width of the central maximum is given by

$$\theta_w = 2\vartheta = \pm\frac{2\lambda}{s}$$

Thus, the smaller s is for a given wavelength, the more quickly does the light energy diverge, and vice versa. This is an important determinant of general behaviour in optical systems.

As a second example, consider a sinusoidal variation of amplitude over the aperture. The Fourier transform of a sinusoid consists of one positive and one negative 'frequency' equally spaced around the origin. Thus, the diffraction pattern consists of just two lines of intensity equally spaced about the centre position of the observing screen (Figure 2.23). Those two lines of intensity could themselves be photographed to provide a 'two-slit' aperture plate that would then provide a sinusoidal diffraction (interference?) pattern. This latter pattern will be viewed as an 'intensity' pattern, however, not an 'amplitude' pattern. Consequently, it will not comprise the original aperture, which must have positive and negative amplitude in order to yield just two lines in its diffraction pattern. Thus, whilst this example illustrates well the strong relationship that exists between the two functions, it also serves to emphasize that the relationship is between the *amplitude* functions, while the observed diffraction pattern is (in the absence of special arrangements) the *intensity* function.

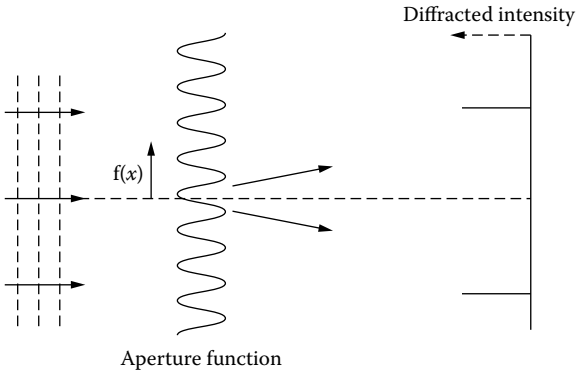


FIGURE 2.23 Sinusoidal diffracting aperture.

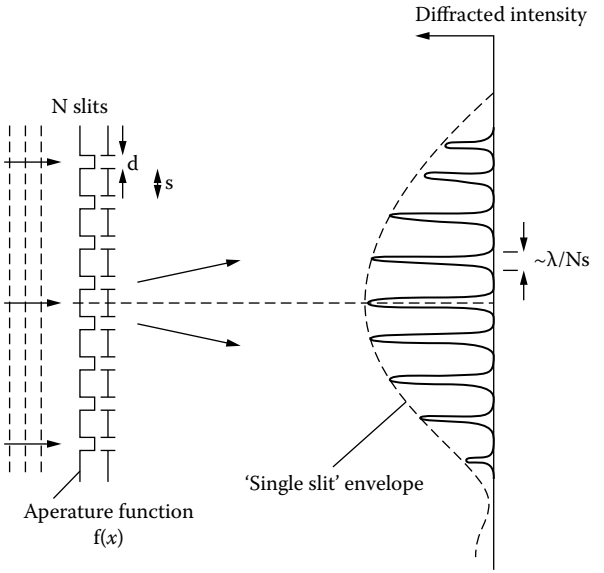


FIGURE 2.24 Diffraction grating.

Finally, we consider one of the most important examples of all: a rectangular-wave aperture amplitude function. The function is shown in Figure 2.24. This is equivalent to a set of narrow slits—that is, to a diffraction grating. The Fourier transform (and hence the Fraunhofer diffraction pattern) will be a set of discrete lines of intensity, spaced uniformly to accord with the 'fundamental' frequency of the aperture function, and enveloped by the Fourier transform of one slit. If the aperture function extended to infinity in each direction, then the individual lines would be infinitely narrow (delta functions), but, as it cannot do so in practice, their width is inversely proportional to the total width of the grating (i.e., the intensity distribution across one line is essentially the Fourier transform of the envelope function for the rectangular wave).

To fix these ideas, consider a grating of N slits, each of width d, and separated by distance s. The diffracted intensity pattern is now given by

$$I(\vartheta) = I_0 \frac{\sin^2 \beta}{\beta^2} \cdot \frac{\sin^2 N\gamma}{\sin^2 \gamma}$$

where

$$\beta = \frac{1}{2}(kd)\sin\vartheta$$

$$\gamma = \frac{1}{2}(ks)\sin\vartheta$$

The pattern is shown in Figure 2.24. It is similar in many ways to that of the Fabry-Perot etalon, as we would expect, because it is a case of multiple interference. Clearly, each wavelength present in the light incident on a diffraction grating will produce its own separate diffraction pattern. This fact is used to analyse the

spectrum of incident light, and also to select and measure specific component wavelengths. Its ability to perform these tasks is most readily characterized by means of its resolving power, which is defined as it was for the Fabry-Perot etalon:

$$\rho = \frac{\lambda}{\delta\lambda}$$

where $\delta\lambda$ is the smallest resolvable wavelength difference. If we take λ to be that wavelength difference that causes the pattern from $\lambda + \mathrm{d}\lambda$ to produce a maximum, of order p, which falls on the first minimum of λ at that same order, then we have

$$pN\lambda + \lambda = pN(\lambda + \delta\lambda)$$

and thus,

$$\rho = \frac{\lambda}{\delta\lambda} = pN$$

Gratings are ruled either on glass (transmission) or on mirrors (reflection) with ~10^5 'lines' (slits) in a distance ~150 mm. The first-order resolving power is thus ~10^5, which is an order down on that for a Fabry-Perot etalon. However, the grating is less demanding of optical and mechanical tolerances in production and use, and it is thus cheaper and less prone to degradation with time.

2.11 GAUSSIAN BEAMS AND STABLE OPTICAL RESONATORS

In the discussions of Fabry-Perot etalons, the reflecting surfaces were parallel and plane. The more recent discussions on diffraction provide further insights into the detailed behaviour of such an arrangement, for we have assumed that the light incident on the mirrors is a plane wave, with uniform amplitude and phase across its aperture. For a circular mirror of diameter d, our considerations of Fraunhofer diffraction have indicated that such an aperture will yield a reflected beam that diverges at angle ~λ/d. Hence, if the mirrors are a distance D apart, and $D >> d$, only a fraction ~ d^4/λ^2D^2 of the light power will be interrupted by the second mirror, and this loss will be sustained for each mirror-to-mirror passage. How can this loss be reduced? To answer this question it is reasonable first to look for a *stable* solution to the problem (i.e., one that does not involve an additional loss for each pass between mirrors). To find this, we may employ our knowledge of diffraction theory to ask the subsidiary question: what aperture amplitude distribution is stable in the face of aperture diffraction effects? Because we know that the far field diffraction pattern is the Fourier transform of the aperture distribution function, it is clear that we are asking, effectively, which function Fourier transforms into itself, or in more mathematical language: which function is invariant under Fourier transformation? There is only one such function—the Gaussian function, of the form

$$f(r) = A\exp\left(-\frac{r^2}{\sigma^2}\right)$$

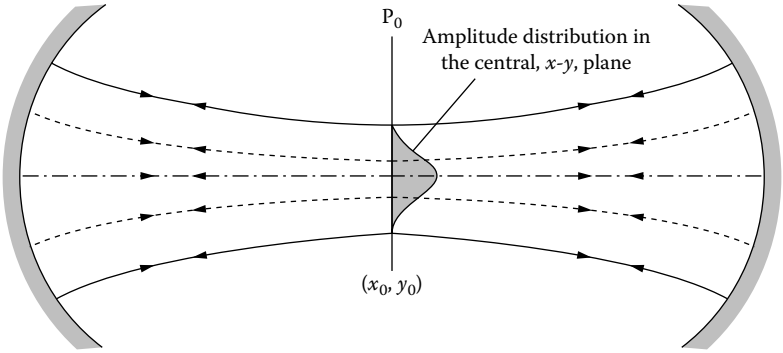


FIGURE 2.25 Gaussian stable resonator.

where r is the radial dimension in the aperture plane and σ is a constant known, in this context, as the 'spot size'.

Suppose, then, that we consider a wave with uniform phase in the plane (x_0y_0) at P_0 as shown in Figure 2.25, and that this wave has a Gaussian amplitude distribution:

$$f(r) = A\exp\left(-\frac{x^2+y^2}{\sigma^2}\right)$$

In the far field this will diffract into a spherical wave with the same form of amplitude distribution; thus, if we place in that field a perfectly reflecting spherical mirror whose diameter is much greater than the spot size at that distance from P_0, essentially all the light will be returned along its incident path (Figure 2.25) (99% of the light will be reflected for a mirror diameter three times the spot size). If another such mirror is also placed on the opposite side of P_0, the light continues to bounce between the spherical mirrors with very little loss. Such an arrangement is known as a 'stable resonator', and it is clear that the light within it is in the form of a 'Gaussian beam'. Such arrangements are preferred for laser structures, because the losses are minimized and the laser becomes more efficient. It also follows that the light that emerges from the partially silvered mirror of the laser source will possess a Gaussian intensity distribution. The condition on the size of the mirror will be satisfied automatically because the settling position for the resonance will be that which minimizes the losses. One may readily see, also, that if a plane mirror is placed at the central plane that contains (x_0y_0), the optical situation is essentially unchanged, so that a spherical mirror can also be used with a plane mirror to create a stable resonator. This configuration is, indeed, sometimes used in laser design.

As the radii of curvature of the mirrors increase, so their diameters must also increase, for a given spacing, in order to obtain a stable Gaussian resonator mode. In the limit as the radius tends to infinity, and the two mirrors thus become plane, the configuration is right on the limit of stability. The diffraction approximations, in fact, break down, and other methods must be used to obtain the aperture intensity distribution, which is now critically dependent on mirror alignment and surface finish.