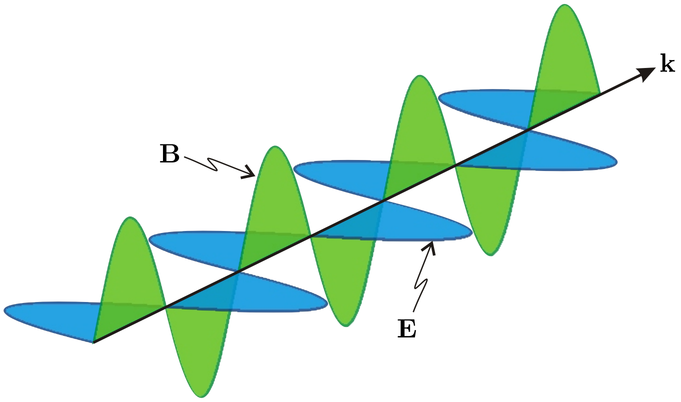
Physics of Light and Optics



Justin Peatross
Michael Ware
Brigham Young University

August 14, 2008

Preface

This book provides an introduction to the field of optics from a physics perspective. It focuses primarily on the wave and ray descriptions of light, but also includes a brief introduction to the quantum description of light. Topics covered include reflection and transmission at boundaries, dispersion, polarization effects, diffraction, coherence, ray optics and imaging, the propagation of light in matter, and the quantum nature of light.

The text is designed for upper-level undergraduate students with a physics background. It assumes that the student already has a basic background with complex numbers, vector calculus, and Fourier transforms, but a brief review of some of these mathematical tools is provided in Chapter 0. The main development of the book begins in Chapter 1 with Maxwell's equations. Subsequent chapters build on this foundation to develop the wave and ray descriptions of classical optics. The final two chapters of the book demonstrate the incomplete nature of classical optics and provide a brief introduction to quantum optics. A collection of electronic material related to the text is available at optics.byu.edu, including videos of students performing the lab assignments found in the book.

This curriculum was developed for a senior-level optics course at Brigham Young University. While the authors retain the copyright, we have made the book available electronically (at no cost) at optics.byu.edu. This site also provides a link to purchase a bound copy of the book for the cost of printing. The authors may be contacted via e-mail at opticsbook@byu.edu. We enjoy hearing reports of how the book is used, and welcome constructive feedback. The text is revised regularly, and the title page indicates the date of the last revision.

We would like to thank all those who have helped improve this material. We especially thank John Colton, Bret Hess, and Harold Stokes for their careful review and extensive suggestions. This curriculum benefits from a CCLI grant from the National Science Foundation Division of Undergraduate Education (DUE-9952773).

Contents

Chapter 0

Mathematical Tools

Optics is an exciting area of study, but (as with most areas of physics) it requires a variety of mathematical tools to be fully appreciated. Before embarking on our study of optics, we take a moment to review a few of the needed mathematical skills. This is not a comprehensive review. We assume that the student already has a basic understanding of differentiation, integration, and standard trigonometric and algebraic manipulation. Section 0.1 reviews complex arithmetic, and students *need to know this material by heart.* Section 0.2 is an overview of vector calculus and related theorems, which are used extensively in electromagnetic theory. It is not essential to be well versed in all of the material presented in section 0.2 (since it is only occasionally needed in homework problems). However, vector calculus is invoked frequently throughout this book, and students will more fully appreciate the connection between electromagnetic principles and optical phenomena when they are comfortable with vector calculus. Section 0.3 is an introduction to Fourier theory. Fourier transforms are used extensively in this course beginning with chapter 7. The presentation below is sufficiently comprehensive for the student who encounters Fourier transforms here for the first time, and such a student is strongly advised to study this section before starting chapter 7.

0.1 Complex Numbers

In optics, it is often convenient to represent electromagnetic wave phenomena as a superposition of sinusoidal functions having the form $A\cos(x+\alpha)$, where x represents a variable, and A and α represent parameters. The sine function is intrinsically present in this formula through the identity

$$\cos(x+\alpha) = \cos x \cos\alpha - \sin x \sin\alpha \tag{0.1}$$

The student of optics should retain this formula in memory, as well as the frequently used identity

$$\sin(x+\alpha) = \sin x \cos\alpha + \sin\alpha \cos x \tag{0.2}$$

With a basic familiarity with trigonometry, one can approach many optical problems including those involving the addition of multiple waves. However, the manipulation of trigonometric functions via identities (0.1) and (0.2) is often cumbersome and tedious. Fortunately, complex notation offers an equivalent approach with far less busy work. One could avoid using complex notation in the study of optics, and this may seem appealing to the student who is unfamiliar with its use. Such a student might opt to pursue all problems using sines, cosines, and real exponents, together with large quantities of trigonometric identities. This, however, would be far more effort than the modest investment needed to become comfortable with the use of complex notation. Optics problems can become cumbersome enough even with the complex notation, so keep in mind that it could be far more messy!

The convenience of complex notation has its origins in Euler's formula:

$$e^{i\phi} = \cos\phi + i\sin\phi \tag{0.3}$$

where $i = \sqrt{-1}$. Euler's formula can be proven using Taylor's expansion:

$$f(x) = f(x_0) + \frac{1}{1!}(x - x_0)\left.\frac{df}{dx}\right|_{x=x_0} + \frac{1}{2!}(x - x_0)^2\left.\frac{d^2f}{dx^2}\right|_{x=x_0} + \cdots \tag{0.4}$$

By expanding each function appearing in (0.3) in a Taylor's series about the origin we obtain

$$\begin{aligned} \cos\phi &= 1 - \frac{\phi^2}{2!} + \frac{\phi^4}{4!} - \cdots \\ i\sin\phi &= i\phi - i\frac{\phi^3}{3!} + i\frac{\phi^5}{5!} - \cdots \\ e^{i\phi} &= 1 + i\phi - \frac{\phi^2}{2!} - i\frac{\phi^3}{3!} + \frac{\phi^4}{4!} + i\frac{\phi^5}{5!} - \cdots \end{aligned} \tag{0.5}$$

The last line of (0.5) is seen to be the sum of the first two lines, from which Euler's formula directly follows.

By inverting Euler's formula (0.3) we can obtain the following representation of the cosine and sine functions:

$$\begin{aligned} \cos\phi &= \frac{e^{i\phi} + e^{-i\phi}}{2}, \\ \sin\phi &= \frac{e^{i\phi} - e^{-i\phi}}{2i} \end{aligned} \tag{0.6}$$

This representation shows how ordinary sines and cosines are intimately related to hyperbolic cosines and hyperbolic sines. If ϕ happens to be imaginary such that $\phi = i\gamma$ where γ is real, then we have

$$\begin{aligned} \sin i\gamma &= \frac{e^{-\gamma} - e^{\gamma}}{2i} = i\sinh\gamma \\ \cos i\gamma &= \frac{e^{-\gamma} + e^{\gamma}}{2} = \cosh\gamma \end{aligned} \tag{0.7}$$

There are several situations in optics where one is interested in a complex angle, $\phi = \beta + i\gamma$ where β and γ are real numbers. For example, the solution to the wave equation when absorption or amplification takes place contains an exponential with a complex argument. In this case, the imaginary part of ϕ introduces exponential decay or growth as is apparent upon examination of (0.6). Another important situation occurs when one attempts to calculate the transmission angle for light incident upon a surface beyond the critical angle for total internal reflection. In this case, it is necessary to compute the arcsine of a number greater than one in an effort to satisfy Snell's law. Even though such an angle does not exist in the usual sense, a complex value for ϕ can be found which satisfies (0.6). The complex value for the angle is useful in computing the characteristics of the evanescent wave on the transmitted side of the surface.

As was mentioned previously, we will be interested in waves of the form $A\cos(x+\alpha)$. We can use complex notation to represent this wave simply by writing

$$A\cos(x+\alpha) = \mathrm{Re}\left\{\tilde{A}e^{ix}\right\} \tag{0.8}$$

where the phase α is conveniently contained within the complex factor $\tilde{A} \equiv Ae^{i\alpha}$. The operation $\mathrm{Re}\{\}$ means to retain only the real part of the argument without regard for the imaginary part. As an example, we have $\mathrm{Re}\{1+2i\} = 1$. The expression (0.8) is a direct result of Euler's equation (0.3).

It is conventional in the study of optics to omit the explicit writing of $\mathrm{Re}\{\}$. Thus, physicists agree that $\tilde{A}e^{ix}$ actually means $A\cos(x+\alpha)$ (or $A\cos\alpha\cos x - A\sin\alpha\sin x$ via (0.1)). This laziness is permissible because it is possible to perform linear operations on $\mathrm{Re}\{f\}$ such as addition, differentiation, or integration while procrastinating the taking of the real part until the end:

$$\begin{gathered}
\mathrm{Re}\{f\} + \mathrm{Re}\{g\} = \mathrm{Re}\{f+g\} \\
\frac{d}{dx}\mathrm{Re}\{f\} = \mathrm{Re}\left\{\frac{df}{dx}\right\} \\
\int \mathrm{Re}\{f\}\,dx = \mathrm{Re}\left\{\int f dx\right\}
\end{gathered} \tag{0.9}$$

As an example, note that $\mathrm{Re}\{1+2i\} + \mathrm{Re}\{3+4i\} = \mathrm{Re}\{(1+2i)+(3+4i)\} = 4$. However, one must be careful when performing other operations such as multiplication. In this case, it is essential to take the real parts before performing the operation. Notice that

$$\mathrm{Re}\{f\} \times \mathrm{Re}\{g\} \neq \mathrm{Re}\{f \times g\} \tag{0.10}$$

As an example, we see $\mathrm{Re}\{1+2i\} \times \mathrm{Re}\{3+4i\} = 3$, but $\mathrm{Re}\{(1+2i)(3+4i)\} = -5$.

When dealing with complex numbers it is often advantageous to transform between a Cartesian representation and a polar representation. With the aid of Euler's formula, it is possible to transform any complex number $a+ib$ into the form $\mathbb{r}$ $\rho e^{i\phi}$, where a, b, ρ, and ϕ are real. From (0.3), the required connection between (ρ,ϕ) and (a,b) is

$$\rho e^{i\phi} = \rho\cos\phi + i\rho\sin\phi = a+ib \tag{0.11}$$

The real and imaginary parts of this equation must separately be equal. Thus, we have

$$\begin{aligned}
a =& \quad \rho\cos\phi \\
b =& \quad \rho\sin\phi
\end{aligned} \tag{0.12}$$

These equations can be inverted to yield

$$\begin{aligned}
\rho &= \sqrt{a^2+b^2} \\
\phi &= \tan^{-1}\frac{b}{a} \qquad (a>0)
\end{aligned} \tag{0.13}$$

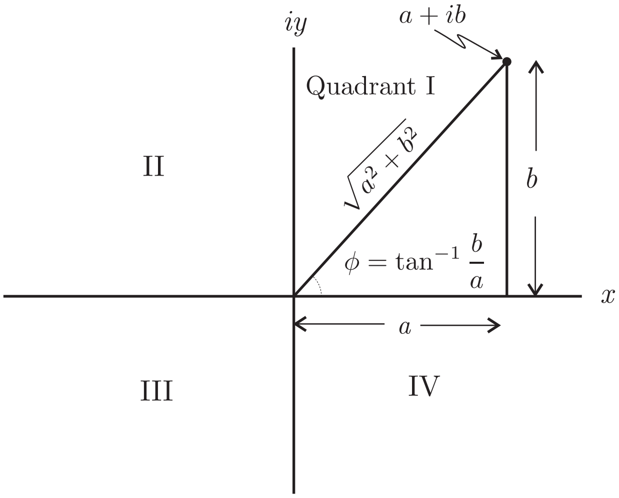


Figure 1 A number in the complex plane can be represented either by Cartesian or polar coordinates.

When $a < 0$, we must adjust ϕ by π since the arctangent has a range only from $-\pi/2$ to $\pi/2$.

The transformations in (0.12) and (0.13) have a clear geometrical interpretation in the complex plane, and this makes it easier to remember them. They are just the usual connections between Cartesian and polar coordinates. As seen in Fig. 1, ρ is the hypotenuse of a right triangle having legs with lengths a and b, and ϕ is the angle that the hypotenuse makes with the x-axis. Again, students should be careful when a is negative since the arctangent is defined in quadrants I and IV. An easy way to deal with the situation of a negative a is to factor the minus sign out before proceeding (i.e. $a + ib = -(-a - ib)$). Then the transformation is made on $-a - ib$ where $-a$ is positive. The minus sign out in front is just carried along unaffected and can be factored back in at the end. Notice that $-\rho e^{i\phi}$ is the same as $\rho e^{i(\phi\pm\pi)}$.

Finally, we consider the concept of a complex conjugate. The conjugate of a complex number $z = a + ib$ is denoted with an asterisk and amounts to changing the sign on the imaginary part of the number:

$$z^* = (a+ib)^* \equiv a - ib \tag{0.14}$$

The complex conjugate is useful when computing the magnitude ρ as defined in (0.13):

$$|z| = \sqrt{z^* z} = \sqrt{(a-ib)(a+ib)} = \sqrt{a^2+b^2} = \rho \tag{0.15}$$

The complex conjugate is also useful for eliminating complex numbers from the denominator of expressions:

$$\frac{a+ib}{c+id} = \frac{(a+ib)}{(c+id)}\frac{(c-id)}{(c-id)} = \frac{ac+bd+i(bc-ad)}{c^2+d^2} \tag{0.16}$$

No matter how complicated an expression, the complex conjugate is calculated by simply inserting a minus sign in front of all occurrences of i in the expression, and placing an asterisk on all complex variables in the expression. For example, the complex conjugate of $\rho e^{i\phi}$ is $\rho e^{-i\phi}$, as can be seen from Euler's formula (0.3). As another example consider $[E\exp\{i(\kappa z-\omega t)\}]^* = E^*\exp\{-i(\kappa^* z-\omega t)\}$, assuming z, ω, and t are real, but E and κ are complex.

A common way of obtaining the real part of an expression is simply by adding the complex conjugate and dividing the result by 2:

$$\mathrm{Re}\{z\} = \frac{1}{2}(z+z^*) \tag{0.17}$$

Notice that the expression for $\cos\phi$ in (0.6) is an example of this formula. Sometimes when a complicated expression is added to its complex conjugate, we let "C.C." represent the complex conjugate in order to avoid writing the expression twice.

0.2 Vector Calculus

In optics we are concerned primarily with electromagnetic fields that are defined throughout space. Each position in space corresponds to a unique vector $\mathbf{r} \equiv x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$, where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are unit vectors of length one, pointing along their respective axes. Electric and magnetic fields are vectors whose magnitude and direction can depend on position, as denoted by $\mathbf{E}(\mathbf{r})$ or $\mathbf{B}(\mathbf{r})$. An example of such a field is $\mathbf{E}(\mathbf{r}) = q(\mathbf{r}-\mathbf{r}_0)\Big/4\pi\epsilon_0|\mathbf{r}-\mathbf{r}_0|^3$, which is the static electric field surrounding a point charge located at position $\mathbf{r}_0$. The absolute value brackets indicate the magnitude (length) of the vector given by

$$\begin{aligned}|\mathbf{r}-\mathbf{r}_0| &= |(x-x_0)\hat{\mathbf{x}} + (y-y_0)\hat{\mathbf{y}} + (z-z_0)\hat{\mathbf{z}}|\\ &= \sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}\end{aligned} \tag{0.18}$$

In addition to space, the electric and magnetic fields almost always depend on time in optics. For example, a time-dependent field common in optics is $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0\exp\{i(\mathbf{k}\cdot\mathbf{r}-\omega t)\}$, where (as discussed above) physicists have the agreement in advance that only the real part of this expression corresponds to the actual field.

The dot product $\mathbf{k}\cdot\mathbf{r}$ is an example of vector multiplication, and signifies the following operation:

$$\begin{aligned}\mathbf{k}\cdot\mathbf{r} &= (k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}} + k_z\hat{\mathbf{z}})\cdot(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}})\\ &= k_x x + k_y y + k_z z\\ &= |\mathbf{k}||\mathbf{r}|\cos\phi\end{aligned} \tag{0.19}$$

where ϕ is the angle between the vectors $\mathbf{k}$ and $\mathbf{r}$. Another type of vector multiplication is the cross product, which is accomplished in the following manner:

$$\begin{aligned}\mathbf{E}\times\mathbf{B} &= \begin{vmatrix}\hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}}\\ E_x & E_y & E_z\\ B_x & B_y & B_z\end{vmatrix}\\ &= (E_yB_z - E_zB_y)\hat{\mathbf{x}} - (E_xB_z - E_zB_x)\hat{\mathbf{y}} + (E_xB_y - E_yB_x)\hat{\mathbf{z}}\end{aligned} \tag{0.20}$$

Note that the cross product results in a vector, whereas the dot product results in a scalar.

We will encounter several multidimensional derivatives in our study: the gradient, the divergence, the curl, and the Laplacian. In Cartesian coordinates, the gradient is given by

$$\nabla f\left(x,y,z\right)=\frac{\partial f}{\partial x}\hat{\mathbf{x}}+\frac{\partial f}{\partial y}\hat{\mathbf{y}}+\frac{\partial f}{\partial z}\hat{\mathbf{z}} \tag{0.21}$$

the divergence is given by

$$\nabla\cdot\mathbf{E}=\frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z} \tag{0.22}$$

the curl is given by

$$\begin{aligned}\nabla\times\mathbf{E}&=\begin{vmatrix}\hat{\mathbf{x}}&\hat{\mathbf{y}}&\hat{\mathbf{z}}\\ \partial/\partial x&\partial/\partial y&\partial/\partial z\\ E_x&E_y&E_z\end{vmatrix}\\ &=\left(\frac{\partial E_z}{\partial y}-\frac{\partial E_y}{\partial z}\right)\hat{\mathbf{x}}-\left(\frac{\partial E_z}{\partial x}-\frac{\partial E_x}{\partial z}\right)\hat{\mathbf{y}}+\left(\frac{\partial E_y}{\partial x}-\frac{\partial E_x}{\partial y}\right)\hat{\mathbf{z}}\end{aligned} \tag{0.23}$$

and the Laplacian is given by

$$\nabla^2 f\left(x,y,z\right)\equiv\nabla\cdot\left[\nabla f\left(x,y,z\right)\right]=\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2} \tag{0.24}$$

You will also encounter the vector Laplacian given by

$$\begin{aligned}\boldsymbol{\nabla}^2\mathbf{E}&\equiv\nabla(\nabla\cdot\mathbf{E})-\nabla\times(\nabla\times\mathbf{E})\\ &=\left(\frac{\partial^2 E_x}{\partial x^2}+\frac{\partial^2 E_x}{\partial y^2}+\frac{\partial^2 E_x}{\partial z^2}\right)\hat{\mathbf{x}}+\left(\frac{\partial^2 E_y}{\partial x^2}+\frac{\partial^2 E_y}{\partial y^2}+\frac{\partial^2 E_y}{\partial z^2}\right)\hat{\mathbf{y}}\\ &\quad+\left(\frac{\partial^2 E_z}{\partial x^2}+\frac{\partial^2 E_z}{\partial y^2}+\frac{\partial^2 E_z}{\partial z^2}\right)\hat{\mathbf{z}}\end{aligned} \tag{0.25}$$

All of these multidimensional derivatives take on a more complicated form in non-cartesian coordinates.

We will also encounter several integral theorems involving vector functions in the course of this book. The divergence theorem for a vector function $\mathbf{f}$ is

$$\oint_S \mathbf{f}\cdot\hat{\mathbf{n}}\,da=\int_V \nabla\cdot\mathbf{f}\,dv \tag{0.26}$$

The integration on the left-hand side is over the closed surface S, which contains the volume V associated with the integration on the right hand side. The unit vector $\hat{\mathbf{n}}$ points normal to the surface. The divergence theorem is especially useful in connection with Gauss's law, where the left hand side is interpreted as the number of field lines exiting a closed surface.

Another important theorem is Stokes' theorem:

$$\int_S \nabla\times\mathbf{f}\cdot\hat{\mathbf{n}}\,da=\oint_C \mathbf{f}\cdot d\ell \tag{0.27}$$

The integration on the left hand side is over an open surface S (not enclosing a volume). The integration on the right hand side is around the edge of the surface. Again, $\hat{\mathbf{n}}$ is a unit vector that always points normal to the surface. The vector $d\ell$ points along the curve C that bounds the surface S. If the fingers of your right hand point in the direction of integration around C, then your thumb points in the direction of $\hat{\mathbf{n}}$. Stokes' theorem is especially useful in connection with Ampere's law and Faraday's law. The right-hand side is an integration of a field around a loop.

The following vector integral theorem will also be useful:

$$\int\limits_V \left[\mathbf{f}\left(\nabla\cdot\mathbf{g}\right)+\left(\mathbf{g}\cdot\nabla\right)\mathbf{f}\right]dv=\oint\limits_S \mathbf{f}\left(\mathbf{g}\cdot\hat{\mathbf{n}}\right)da \tag{0.28}$$

0.3 Fourier Theory

Fourier analysis is an important part of optics. We often decompose complicated light fields into a superposition of pure sinusoidal waves. This enables us to consider the behavior of the individual frequency components one at a time (important since, for example, the optical index is different for different frequencies). After determining how individual sine waves move through an optical system (say a piece of glass), we can reassemble the sinusoidal waves to see the effect of the system on the overall waveform. Fourier transforms are used for this purpose. In fact, it will be possible to work simultaneously with infinitely many sinusoidal waves, where the frequencies comprising a light field are spread over a continuous range. Fourier transforms are also used in diffraction problems where a single frequency is associated with a superposition of many plane waves propagating in different directions.

We begin with a derivation of the Fourier integral theorem. A *periodic* function can be represented in terms of the sine and the cosine in the following manner:

$$f\left(t\right)=\sum_{n=0}^{\infty}a_n\cos\left(n\Delta\omega t\right)+b_n\sin\left(n\Delta\omega t\right) \tag{0.29}$$

This is called a Fourier expansion. It is similar in idea to a Taylor's series (0.4), which rewrites a function as a polynomial. In both cases, the goal is to represent one function in terms of a linear combination of other functions (requiring a complete basis set). In a Taylor's series the basis functions are polynomials and in a Fourier expansion the basis functions are sines and cosines with different frequencies.

The expansion (0.29) is possible even if $f(t)$ is complex (requiring a_n and b_n to be complex). By inspection, we see that all terms in (0.29) repeat with a maximum period of $2\pi/\Delta\omega$. This is why the expansion is limited in its use to periodic functions. The period of the function by such an expansion is such that $f(t)=f(t+2\pi/\Delta\omega)$.

We can rewrite the sines and cosines in the expansion (0.29) using (0.6) as follows:

$$\begin{aligned}f(t)&=\sum_{n=0}^{\infty}a_n\frac{e^{in\Delta\omega t}+e^{-in\Delta\omega t}}{2}+b_n\frac{e^{in\Delta\omega t}-e^{-in\Delta\omega t}}{2i}\\&=a_0+\sum_{n=1}^{\infty}\frac{a_n-ib_n}{2}e^{in\Delta\omega t}+\sum_{n=1}^{\infty}\frac{a_n+ib_n}{2}e^{-in\Delta\omega t}\end{aligned} \tag{0.30}$$

Thus, we can rewrite (0.29) as

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{-in\Delta\omega t} \tag{0.31}$$

where

$$\begin{aligned} c_{n<0} &\equiv \frac{a_{-n} - ib_{-n}}{2} \\ c_{n>0} &\equiv \frac{a_n + ib_n}{2} \\ c_0 &\equiv a_0 \end{aligned} \tag{0.32}$$

Notice that if $c_{-n} = c_n^*$ for all n, then $f(t)$ is real (i.e. real a_n and b_n); otherwise $f(t)$ is complex. The real parts of the c_n coefficients are connected with the cosine terms in (0.29), and the imaginary parts of the c_n coefficients are connected with the sine terms in (0.29).

Given a known function $f(t)$, we can compute the various coefficients c_n. There is a trick for doing this. We multiply both sides of (0.31) by $e^{im\Delta\omega t}$, where m is an integer, and integrate over the function period $2\pi/\Delta\omega$:

$$\begin{aligned} \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t)e^{im\Delta\omega t}dt &= \sum_{n=-\infty}^{\infty} c_n \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} e^{i(m-n)\Delta\omega t}dt \\ &= \sum_{n=-\infty}^{\infty} c_n \left[\frac{e^{i(m-n)\Delta\omega t}}{i\,(m-n)\,\Delta\omega} \right]_{-\pi/\Delta\omega}^{\pi/\Delta\omega} \\ &= \sum_{n=-\infty}^{\infty} \frac{2\pi c_n}{\Delta\omega} \left[\frac{e^{i(m-n)\pi} - e^{-i(m-n)\pi}}{2i\,(m-n)\,\pi} \right] \\ &= \sum_{n=-\infty}^{\infty} \frac{2\pi c_n}{\Delta\omega} \frac{\sin\left[(m-n)\,\pi\right]}{(m-n)\,\pi} \end{aligned} \tag{0.33}$$

The function $\sin\left[(m-n)\,\pi\right] / \left[(m-n)\,\pi\right]$ is equal to zero for all $n \neq m$, and it is equal to one when $n = m$ (to see this, use L'Hospital's rule on the zero-over-zero situation). Thus, only one term contributes to the summation in (0.33). We now have

$$c_m = \frac{\Delta\omega}{2\pi} \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t)e^{im\Delta\omega t}dt \tag{0.34}$$

from which the coefficients c_n can be computed, given a function $f(t)$. (Note that m is a dummy index so we can change it back to n if we like.)

This completes the circle. If we know the function $f(t)$, we can find the coefficients c_n via (0.34), and, if we know the coefficients c_n, we can generate the function $f(t)$ via (0.31). If we are feeling a bit silly, we can combine these into a single identity:

$$f(t) = \sum_{n=-\infty}^{\infty} \left[\frac{\Delta\omega}{2\pi} \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t)e^{in\Delta\omega t}dt \right] e^{-in\Delta\omega t} \tag{0.35}$$

We start with a function $f(t)$ followed by a lot of computation and obtain the function back again! (This is not quite as foolish as it first appears, as we will see later.)

As mentioned above, Fourier expansions represent functions $f(t)$ that are periodic over the interval $2\pi/\Delta\omega$. This is disappointing since many optical waveforms do not repeat (e.g. a single short laser pulse). Nevertheless, we can represent a function $f(t)$ that is not periodic if we let the period $2\pi/\Delta\omega$ become infinitely long. In other words, we can accommodate non-periodic functions if we take the limit as $\Delta\omega$ goes to zero so that the spacing of terms in the series becomes very fine. Applying this limit to (0.35) we obtain

$$f(t) = \frac{1}{2\pi} \lim_{\Delta\omega \to 0} \sum_{n=-\infty}^{\infty} \left[e^{-in\Delta\omega t} \int_{-\infty}^{\infty} f\left(t'\right) e^{in\Delta\omega t'} dt' \right] \Delta\omega \tag{0.36}$$

At this point, a brief review of the definition of an integral is helpful to better understand the next step that we administer to (0.36). Recall that an integral is really a summation of rectangles under a curve with finely spaced steps:

$$\int_a^b g(\omega)\, d\omega \equiv \lim_{\Delta\omega \to 0} \sum_{n=0}^{\frac{b-a}{\Delta\omega}} g(a + n\Delta\omega)\, \Delta\omega = \lim_{\Delta\omega \to 0} \sum_{n=-\frac{b-a}{2\Delta\omega}}^{\frac{b-a}{2\Delta\omega}} g\left(\frac{a+b}{2} + n\Delta\omega\right) \Delta\omega \tag{0.37}$$

The final expression has been manipulated so that the index ranges through both negative and positive numbers. If we set $a = -b$ and take the limit $b \to \infty$, then (0.37) becomes

$$\int_{-\infty}^{\infty} g(\omega)\; d\omega = \lim_{\Delta\omega \to 0} \sum_{n=-\infty}^{\infty} g(n\Delta\omega)\, \Delta\omega \tag{0.38}$$

This concludes our short review of calculus.

We can use (0.38) in connection with (0.36) (where $g(n\Delta\omega)$ represents everything in the square brackets). The result is the Fourier integral theorem:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega t} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f\left(t'\right) e^{i\omega t'} dt' \right] d\omega \tag{0.39}$$

The piece in brackets is called the Fourier transform, and the rest of the operation is called the inverse Fourier transform. The Fourier integral theorem (0.39) is often written with the following (potentially confusing) notation:

$$\begin{aligned} f(\omega) &\equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{i\omega t}\; dt \\ f(t) &\equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\omega)\, e^{-i\omega t}\; d\omega \end{aligned} \tag{0.40}$$

The transform and inverse transform are also sometimes written as $f(\omega) \equiv \mathcal{F}\{f(t)\}$ and $f(t) \equiv \mathcal{F}^{-1}\{f(\omega)\}$. Note that the functions $f(t)$ and $f(\omega)$ are entirely different, even taking on different units (e.g. the latter having extra units of per frequency). The two functions are distinguished by their arguments, which also have different units (e.g. time vs. frequency). Nevertheless, it is customary to use the same letter to denote either function since they form a transform pair.

You should be aware that it is arbitrary which of the expressions in (0.40) is called the transform and which is called the inverse transform. In other words, the signs in the exponents of (0.40) may be interchanged. The convention varies in published works. Also, the factor 2π may be placed on either the transform or the inverse transform, or divided equally between the two as has been done here.

As was previously mentioned, it would seem rather pointless to perform a Fourier transform on the function $f(t)$ followed by an inverse Fourier transform, just to end up with $f(t)$ again. Nevertheless, we are interested in this because we want to know the effect of an optical system on a waveform (represented by $f(t)$). It turns out that in many cases, the effect of the optical system can only be applied to $f(\omega)$ (if the effect is frequency dependent). Thus, we perform a Fourier transform on $f(t)$, then apply the frequency-dependent effect on $f(\omega)$, and finally perform an inverse Fourier transform on the result. The final function will be different from $f(t)$. Keep in mind that $f(\omega)$ is the continuous analog of the discrete coefficients c_n (or the a_n and b_n). The real part of $f(\omega)$ indicates the amplitudes of the cosine waves necessary to construct the function $f(t)$. The imaginary part of $f(\omega)$ indicates the amplitudes of the sine waves necessary to construct the function $f(t)$.

Finally, we note that a remarkable attribute of the delta function can be seen from the Fourier integral theorem. The delta function $\delta(t'-t)$ is defined indirectly through

$$f(t) = \int_{-\infty}^{\infty} f(t')\,\delta(t'-t)\,dt' \tag{0.41}$$

The delta function $\delta(t'-t)$ is zero everywhere except at $t' = t$, since the result of the integration only pays attention to the value of $f(t')$ at that point. At $t' = t$, the delta function is infinite in such a way as to make the integral take on the value of the function $f(t)$. (One can consider $\delta(t'-t)\,dt'$ with $t' = t$ to be the dimensions of an infinitely tall and infinitely thin rectangle with an area unity.) After rearranging the order of integration, the Fourier integral theorem (0.39) can be written as

$$f(t) = \int_{-\infty}^{\infty} f(t')\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega(t'-t)}d\omega\right]dt' \tag{0.42}$$

A comparison of (0.41) and (0.42) reveals the delta function to be a uniform superposition of all frequency components:

$$\delta(t'-t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega(t'-t)}\,d\omega \tag{0.43}$$

This representation of the delta function comes in handy when proving Parseval's theorem (see P 0.31), which is used extensively in the study of light and optics.

0.4 Linear Algebra and Sylvester's Theorem

In this section we outline two useful results from linear algebra. The first result states that the inverse of a 2 × 2 matrix is given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \frac{1}{AD - BC}\begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \tag{0.44}$$

This can be proven by direct substitution:

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \frac{1}{AD - BC}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \\ &= \frac{1}{AD - BC}\begin{bmatrix} AD - BC & 0 \\ 0 & AD - BC \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{0.45}$$

The next result is Sylvester's Theorem, which is useful when a 2 × 2 matrix (with a determinate of unity) is raised to a high power. This situation occurs for modeling periodic multilayer mirror coatings or for light rays trapped in a laser cavity. Sylvester's Theorem states that if the determinant of a 2 × 2 matrix is one, i.e.

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = AD - BC = 1 \tag{0.46}$$

then the following holds:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{N} = \frac{1}{\sin\theta}\begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix} \tag{0.47}$$

where

$$\cos\theta = \frac{1}{2}(A + D) \tag{0.48}$$

We prove the theorem by induction. When $N = 1$, the equation is seen to be correct by direct substitution. Next we assume that the theorem holds for arbitrary N, and we check to see if it holds for $N + 1$:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{N+1} = \frac{1}{\sin\theta}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix} \tag{0.49}$$

Now we inject condition (0.46) on the determinant ($AD - BC = 1$) into the right-hand side of (0.49)

$$\frac{1}{\sin\theta}\begin{bmatrix} \left(A^2 + BC\right)\sin N\theta - A\sin(N-1)\theta & (AB + BD)\sin N\theta - B\sin(N-1)\theta \\ (AC + CD)\sin N\theta - C\sin(N-1)\theta & \left(D^2 + BC\right)\sin N\theta - D\sin(N-1)\theta \end{bmatrix}$$

and rearrange the result to give

$$\frac{1}{\sin\theta}\begin{bmatrix} \left(A^2 + AD - 1\right)\sin N\theta - A\sin(N-1)\theta & B\left[(A + D)\sin N\theta - \sin(N-1)\theta\right] \\ C\left[(A + D)\sin N\theta - \sin(N-1)\theta\right] & \left(D^2 + AD - 1\right)\sin N\theta - D\sin(N-1)\theta \end{bmatrix}$$

and then

$$\frac{1}{\sin\theta}\begin{bmatrix} A\left[(A+D)\sin N\theta - \sin\left(N-1\right)\theta\right] - \sin N\theta & B\left[(A+D)\sin N\theta - \sin\left(N-1\right)\theta\right] \\ C\left[(A+D)\sin N\theta - \sin\left(N-1\right)\theta\right] & D\left[(A+D)\sin N\theta - \sin\left(N-1\right)\theta\right] - \sin N\theta \end{bmatrix}$$

In each matrix element, the expression

$$(A+D)\sin N\theta = 2\cos\theta\sin N\theta = \sin\left(N+1\right)\theta + \sin\left(N-1\right)\theta \tag{0.50}$$

occurs, which we have rearranged using $\cos\theta = \frac{1}{2}\left(A+D\right)$ while twice invoking (0.2). The result is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{N+1} = \frac{1}{\sin\theta}\begin{bmatrix} A\sin\left(N+1\right)\theta - \sin N\theta & B\sin\left(N+1\right)\theta \\ C\sin\left(N+1\right)\theta & D\sin\left(N+1\right)\theta - \sin N\theta \end{bmatrix} \tag{0.51}$$

which completes the proof.

Appendix 0.A Integral and Sum Table

The following table of formulas are useful for various problems encountered in the text.

$$\int_{-\infty}^{\infty} e^{-ax^2+bx+c}\, dx = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}+c} \qquad \mathrm{Re}\left\{a\right\} > 0 \tag{0.52}$$

$$\int_{0}^{\infty} \frac{e^{iax}}{1+x^2/b^2}\, dx = \frac{\pi\left|b\right|}{2}e^{-\left|ab\right|} \qquad b > 0 \tag{0.53}$$

$$\int_{0}^{2\pi} e^{\pm ia\cos(\theta-\theta')}\, d\theta = 2\pi J_0\left(a\right) \tag{0.54}$$

$$\int_{0}^{a} J_0\left(bx\right) x\, dx = \frac{a}{b}J_1\left(ab\right) \tag{0.55}$$

$$\int_{0}^{\infty} e^{-ax^2} J_0\left(bx\right) x\, dx = \frac{e^{-b^2/4a}}{2a} \tag{0.56}$$

$$\int_{0}^{\infty} \frac{\sin^2(ax)}{\left(ax\right)^2}\, dx = \frac{\pi}{2a} \tag{0.57}$$

$$\int_{0}^{\pi} \sin(ax)\sin(bx)\, dx = \int_{0}^{\pi} \cos(ax)\cos(bx)\, dx = \frac{1}{2}\delta_{ab} \qquad (a, b \text{ integer}) \tag{0.58}$$

$$\sum_{n=1}^{N} ar^n = a\frac{1-r^N}{1-r} \tag{0.59}$$

$$\sum_{n=1}^{\infty} ar^n = \frac{a}{1-r} \qquad (r < 1) \tag{0.60}$$

Exercises

0.1 Complex Numbers

P0.1 Do the following complex arithmetic problems using real arithmetic functions along with the fundamentals of complex numbers (i.e. don't use your calculator's complex arithmetic abilities):

(a) For $z_1 = 2+3i$ and $z_2 = 3-5i$, calculate z_1+z_2 and $z_1 \times z_2$ in both rectangular and polar form.

(b) For $z_1 = 1-i$ and $z_2 = 3+4i$, calculate $z_1 - z_2$ and z_1/z_2 in both rectangular and polar form.

P0.2 Show that $-3+4i$ can be written as $5\exp\left\{-i\tan^{-1}4/3 + i\pi\right\}$.

P0.3 Show $(a-ib)/(a+ib) = \exp\left\{-2i\tan^{-1}b/a\right\}$ regardless of the sign of a, assuming a and b are real.

P0.4 Invert (0.3) to get both formulas in (0.6).

P0.5 Show $\operatorname{Re}\{A\} \times \operatorname{Re}\{B\} = (AB + A^*B)/4 + C.C.$

P0.6 If $E = |E|\, e^{i\alpha_E}$ and $B = |B|\, e^{i\alpha_B}$, and if k, z, ω, and t are all real, prove

$$\begin{aligned}\operatorname{Re}\left\{Ee^{i(kz-\omega t)}\right\}\operatorname{Re}\left\{Be^{i(kz-\omega t)}\right\} &= \frac{1}{4}\left(E^*B + EB^*\right)\\ &+ \frac{1}{2}|E|\,|B|\cos\left[2\left(kz-\omega t\right) + \alpha_E + \alpha_B\right]\end{aligned}$$

P0.7 (a) If $\sin\phi = 2$, show that $\cos\phi = i\sqrt{3}$. HINT: Use $\sin^2\phi + \cos^2\phi = 1$.

(b) Show that the angle ϕ in (a) is $\pi/2 - i\ln(2+\sqrt{3})$.

P0.8 Use the techniques/principles of complex numbers to write the following as simple phase-shifted cosine waves (i.e. find the amplitude and phase of the resultant cosine waves):

(a) $5\cos(4t) + 5\sin(4t)$

(b) $3\cos(5t) + 10\sin(5t + 0.4)$

0.2 Vector Calculus

P0.9 Let $\mathbf{r} = (\hat{\mathbf{x}} + 2\hat{\mathbf{y}} - 3\hat{\mathbf{z}})$ m and $\mathbf{r}_0 = (-\hat{\mathbf{x}} + 3\hat{\mathbf{y}} + 2\hat{\mathbf{z}})$ m.

(a) Find the magnitude of $\mathbf{r}$.

(b) Find $\mathbf{r} - \mathbf{r}_0$.

(c) Find the angle between $\mathbf{r}$ and $\mathbf{r}_0$.

Answer: (a) $r = \sqrt{14}$ m; (c) 94°.

P0.10 Prove that the dot product between two vectors is the product of the magnitudes of the two vectors multiplied by the cosine of the angle between them.

Solution: Consider the plane containing the two vectors in (0.19). Call it the xy-plane. In this coordinate system, the two vectors can be written as $\mathbf{k} = k\cos\theta\hat{\mathbf{x}} + k\sin\theta\hat{\mathbf{y}}$ and $\mathbf{r} = r\cos\alpha\hat{\mathbf{x}} + r\sin\alpha\hat{\mathbf{y}}$, where θ and α are the respective angles that the two vectors make with the x-axis. The dot product gives $\mathbf{k}\cdot\mathbf{r} = kr\left(\cos\theta\cos\alpha + \sin\theta\sin\alpha\right)$. From (0.1) we have $\mathbf{k}\cdot\mathbf{r} = kr\cos\left(\theta-\alpha\right)$, which shows that $\theta-\alpha$ is the angle between the vectors.

P0.11 Prove that the cross product between two vectors is the product of the magnitudes of the two vectors multiplied by the sine of the angle between them. The result is a vector directed perpendicular to the plane containing the original two vectors in accordance with the right hand rule.

P0.12 Verify the "BAC-CAB" rule: $\mathbf{A}\times(\mathbf{B}\times\mathbf{C}) = \mathbf{B}\left(\mathbf{A}\cdot\mathbf{C}\right) - \mathbf{C}\left(\mathbf{A}\cdot\mathbf{B}\right)$.

P0.13 Prove the following identity:

$$\nabla_{\mathbf{r}}\frac{1}{|\mathbf{r}-\mathbf{r}'|} = -\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3},$$

where $\nabla_{\mathbf{r}}$ operates only on $\mathbf{r}$, treating $\mathbf{r}'$ as a constant vector.

P0.14 Prove that $\nabla_{\mathbf{r}}\cdot\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}$ is zero, except at $\mathbf{r}=\mathbf{r}'$ where a singularity situation occurs.

P0.15 Verify $\nabla\cdot(\nabla\times\mathbf{f}) = 0$ for any vector function $\mathbf{f}$.

P0.16 Verify $\nabla\times(\nabla\times\mathbf{f}) = \nabla\left(\nabla\cdot\mathbf{f}\right) - \nabla^2\mathbf{f}$

Solution: From (0.23), we have

$$\nabla\times\mathbf{f} = \left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z}\right)\hat{\mathbf{x}} - \left(\frac{\partial f_z}{\partial x} - \frac{\partial f_x}{\partial z}\right)\hat{\mathbf{y}} + \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y}\right)\hat{\mathbf{z}}$$

and

$$\begin{aligned}\nabla\times(\nabla\times\mathbf{f}) &= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ \left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z}\right) & -\left(\frac{\partial f_z}{\partial x} - \frac{\partial f_x}{\partial z}\right) & \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y}\right) \end{vmatrix} \\ &= \left[\frac{\partial}{\partial y}\left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial f_z}{\partial x} - \frac{\partial f_x}{\partial z}\right)\right]\hat{\mathbf{x}} - \left[\frac{\partial}{\partial x}\left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z}\right)\right]\hat{\mathbf{y}} \\ &\quad + \left[-\frac{\partial}{\partial x}\left(\frac{\partial f_z}{\partial x} - \frac{\partial f_x}{\partial z}\right) - \frac{\partial}{\partial y}\left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z}\right)\right]\hat{\mathbf{z}}\end{aligned}$$

After rearranging, we get

$$\begin{aligned}\nabla\times(\nabla\times\mathbf{f}) &= \left[\frac{\partial^2 f_x}{\partial x^2} + \frac{\partial^2 f_y}{\partial x\partial y} + \frac{\partial^2 f_z}{\partial x\partial z}\right]\hat{\mathbf{x}} + \left[\frac{\partial^2 f_x}{\partial x\partial y} + \frac{\partial^2 f_y}{\partial y^2} + \frac{\partial^2 f_z}{\partial y\partial z}\right]\hat{\mathbf{y}} + \left[\frac{\partial^2 f_x}{\partial x\partial z} + \frac{\partial^2 f_y}{\partial y\partial z} + \frac{\partial^2 f_z}{\partial z^2}\right]\hat{\mathbf{z}} \\ &\quad - \left[\frac{\partial^2 f_x}{\partial x^2} + \frac{\partial^2 f_x}{\partial y^2} + \frac{\partial^2 f_x}{\partial z^2}\right]\hat{\mathbf{x}} - \left[\frac{\partial^2 f_y}{\partial x^2} + \frac{\partial^2 f_y}{\partial y^2} + \frac{\partial^2 f_y}{\partial z^2}\right]\hat{\mathbf{y}} - \left[\frac{\partial^2 f_z}{\partial x^2} + \frac{\partial^2 f_z}{\partial y^2} + \frac{\partial^2 f_z}{\partial z^2}\right]\hat{\mathbf{z}}\end{aligned}$$

where we have added and subtracted $\frac{\partial^2 f_x}{\partial x^2} + \frac{\partial^2 f_y}{\partial y^2} + \frac{\partial^2 f_z}{\partial z^2}$. After some factorization, we obtain

$$\begin{aligned}\nabla \times (\nabla \times \mathbf{f}) &= \left[\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right]\left[\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z}\right] - \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right][f_x\hat{\mathbf{x}} + f_y\hat{\mathbf{y}} + f_z\hat{\mathbf{z}}] \\ &= \nabla(\nabla \cdot \mathbf{f}) - \nabla^2\mathbf{f}\end{aligned}$$

where on the final line we invoked (0.21), (0.22), and (0.24).

P0.17 Verify $\nabla \times (\mathbf{f} \times \mathbf{g}) = \mathbf{f}(\nabla \cdot \mathbf{g}) - \mathbf{g}(\nabla \cdot \mathbf{f}) + (\mathbf{g} \cdot \nabla)\mathbf{f} - (\mathbf{f} \cdot \nabla)\mathbf{g}$.

P0.18 Verify $\nabla \cdot (\mathbf{f} \times \mathbf{g}) = \mathbf{g} \cdot (\nabla \times \mathbf{f}) - \mathbf{f} \cdot (\nabla \times \mathbf{g})$.

P0.19 Verify $\nabla \cdot (g\mathbf{f}) = \mathbf{f} \cdot \nabla g + g\nabla \cdot \mathbf{f}$.

P0.20 Verify $\nabla \times (g\mathbf{f}) = (\nabla g) \times \mathbf{f} + g\nabla \times \mathbf{f}$.

P0.21 Verify the divergence theorem (0.26) for $\mathbf{f}(x, y, z) = y^2\hat{\mathbf{x}} + xy\hat{\mathbf{y}} + x^2z\hat{\mathbf{z}}$. Take as the volume a cube contained by the six planes $|x| = \pm 1$, $|y| = \pm 1$, and $|z| = \pm 1$.

Solution:

$$\begin{aligned}\oint_S \mathbf{f} \cdot \hat{\mathbf{n}} da &= \int_{-1}^{1}\int_{-1}^{1} dxdy\,(x^2z)_{z=1} - \int_{-1}^{1}\int_{-1}^{1} dxdy\,(x^2z)_{z=-1} + \int_{-1}^{1}\int_{-1}^{1} dxdz\,(xy)_{y=1} - \\ &\quad - \int_{-1}^{1}\int_{-1}^{1} dxdz\,(xy)_{y=-1} + \int_{-1}^{1}\int_{-1}^{1} dydz\,(y^2)_{x=1} - \int_{-1}^{1}\int_{-1}^{1} dydz\,(y^2)_{x=-1} \\ &= 2\int_{-1}^{1}\int_{-1}^{1} dxdyx^2 + 2\int_{-1}^{1}\int_{-1}^{1} dxdzx = 4\left.\frac{x^3}{3}\right|_{-1}^{1} + 4\left.\frac{x^2}{2}\right|_{-1}^{1} = \frac{8}{3}. \\ \int_V \nabla \cdot \mathbf{f} dv &= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} dxdydz\,[x + x^2] = 4\int_{-1}^{1} dx\,[x + x^2] = 4\left[\frac{x^2}{2} + \frac{x^3}{3}\right]_{-1}^{1} = \frac{8}{3}.\end{aligned}$$

P0.22 Verify Stokes' theorem (0.27) for the function given in P 0.21. Take the surface to be a square in the xy-plane contained by $|x| = \pm 1$ and $|y| = \pm 1$.

P0.23 Use the divergence theorem to show that the function in P 0.14 is 4π times the three-dimensional delta function.

Solution: We have by the divergence theorem

$$\oint_S \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \cdot \hat{\mathbf{n}} da = \int_V \nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} dv$$

From P 0.14, the argument in the integral on the right-hand side is zero except at $\mathbf{r} = \mathbf{r}'$. Therefore, if the volume V does not contain the point $\mathbf{r} = \mathbf{r}'$, then the result of both integrals must be zero. Let us construct a volume between an arbitrary surface S_1 containing $\mathbf{r} = \mathbf{r}'$and S_2, the surface of a tiny sphere centered on $\mathbf{r} = \mathbf{r}'$. Since the point $\mathbf{r} = \mathbf{r}'$ is excluded by the tiny sphere, the result of either integral in the divergence theorem is still zero. However, we have on the tiny sphere

$$\oint_{S_2} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \cdot \hat{\mathbf{n}} da = -\int_0^{2\pi}\int_0^{\pi} \left(\frac{1}{r_\epsilon^2}\right) r_\epsilon^2 \sin\phi d\phi d\alpha = -4\pi$$

Therefore, for the outer surface S_1 (containing $\mathbf{r} = \mathbf{r}'$) we must have the equal and opposite result:

$$\oint_{S_1} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \cdot \hat{\mathbf{n}} da = 4\pi$$

This implies

$$\int_V \nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} dv = \begin{cases} 4\pi \text{ if } V \text{ contains } \mathbf{r}' \\ 0 \text{ otherwise} \end{cases}$$

The argument of this integral exhibits the same characteristics as the delta function $\delta^3(\mathbf{r}' - \mathbf{r}) \equiv \delta(x' - x)\,\delta(y' - y)\,\delta(z' - z)$. Namely,

$$\int_V \delta^3(\mathbf{r}' - \mathbf{r})\, dv = \begin{cases} 1 \text{ if } V \text{ contains } \mathbf{r}' \\ 0 \text{ otherwise} \end{cases}$$

Therefore, $\nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} = 4\pi\delta^3(\mathbf{r} - \mathbf{r}')$. The delta function is defined in (0.41)

0.3 Fourier Theory

P0.24 Prove linear superposition of Fourier Transforms:

$$\mathcal{F}\{ag(t) + bh(t)\} = ag(\omega) + bh(\omega)$$

where $g(\omega) \equiv \mathcal{F}\{g(t)\}$ and $h(\omega) \equiv \mathcal{F}\{h(t)\}$.

P0.25 Prove $\mathcal{F}\{g(at)\} = \frac{1}{|a|} g\left(\frac{\omega}{a}\right)$.

P0.26 Prove $\mathcal{F}\{g(t - \tau)\} = g(\omega)e^{i\omega\tau}$.

P0.27 Show that the Fourier transform of $E(t) = E_0 e^{-(t/\tau)^2} \cos\omega_0 t$ is

$$E(\omega) = \frac{\tau E_0}{2\sqrt{2}} \left(e^{-\frac{(\omega+\omega_0)^2}{4/\tau^2}} + e^{-\frac{(\omega-\omega_0)^2}{4/\tau^2}} \right)$$

P0.28 Take the inverse Fourier transform of the result in P 0.27. Check that it returns exactly the original function.

P0.29 The following operation is referred to as the *convolution* of the functions g and h:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(t)h(\tau - t)\, dt$$

A convolution measures the overlap of g and a reversed h as a function of the offset τ.

(a) Prove the convolution theorem:

$$\mathcal{F}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(t)h(\tau-t)\ dt\right\} = g(\omega)h(\omega)$$

(b) Prove this related form of the convolution theorem:

$$\mathcal{F}\left\{g(t)h(t)\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(\omega')h(\omega-\omega')\ d\omega'$$

Solution: Part (a)

$$\begin{aligned}
\mathcal{F}\left\{\int_{-\infty}^{\infty} g(t)h(\tau-t)\ dt\right\} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} g\left(t\right)h\left(\tau-t\right)dt\right\}e^{i\omega\tau}d\tau \qquad (\text{Let } \tau = t'+t)\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} g\left(t\right)h\left(t'\right)dt\right\}e^{i\omega\left(t'+t\right)}dt'\\
&= \sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g\left(t\right)e^{i\omega t}dt\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h\left(t'\right)e^{i\omega t'}dt'\\
&= \sqrt{2\pi}g\left(\omega\right)h\left(\omega\right)
\end{aligned}$$

P0.30 Prove the autocorrelation theorem:

$$\mathcal{F}\left\{\int_{-\infty}^{\infty} h(t)h^*(t-\tau)dt\right\} = \sqrt{2\pi}\left|h(\omega)\right|^2$$

P0.31 Prove Parseval's theorem:

$$\int_{-\infty}^{\infty}\left|f\left(\omega\right)\right|^2 d\omega = \int_{-\infty}^{\infty}\left|f(t)\right|^2 dt$$

P0.32 (a) Compute the Fourier transform of a Gaussian function, $f_1(t) = e^{-t^2/2\tau^2}$. Do the integral by hand using the table in Appendix 0.A.

(b) Compute the Fourier transform of a sine function, $f_2(t) = \sin\omega_0 t$. Don't use a computer to do the integral—use the fact that $\sin(x) = \frac{1}{2i}(e^{ix} - e^{-ix})$, combined with the integral formula (0.43).

(c) Use your results to parts (a) and (b) and a convolution theorem from P 0.29 to evaluate the Fourier transform of $g(t) = e^{-t^2/2\tau^2}\sin\omega_0 t$. (The answer should be similar to 0.27).

(d) Plot $g(t)$ and the imaginary part of its Fourier transform for the parameters $\omega_0 = 1$ and $\tau = 8$.

P0.33 Use your results from P 0.32, along with a convolution theorem from P 0.29, to evaluate the Fourier transform of

$$h(t) = e^{-(t-t_0)^2/2\tau^2} \sin \omega_0 t + e^{-t^2/2\tau^2} \sin \omega_0 t + e^{-(t+t_0)^2/2\tau^2} \sin \omega_0 t$$

which consists of the sum of three Gaussian pulses, each separated by a time t_0.

HINT: The three-pulse function $h(t)$ is a convolution of $e^{-t^2/2\tau^2} \sin \omega_0 t$ with three delta functions. Here is a good check for your final answer: if you set $t_0 = 0$, the three pulses are on top of each other, so you should get three times the answer to problem P 0.32(c).

(b) Plot $h(t)$ and the imaginary part of its Fourier transform for the parameters $\omega_0 = 1$, $\tau = 8$, and $t_0 = 30$.

(c) This $h(t)$ is "longer" than the single pulse in problem P 0.32(c). Should its Fourier transform be broader or narrower than in P 0.32(c)? Comment on what you see in the plots.

Chapter 1

Electromagnetic Phenomena

1.1 Introduction

In the mid 1800's James Maxwell assembled the various known relationships of electricity and magnetism into a concise set of equations:[1]

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \qquad \text{(Gauss's Law from Coulomb's Law)} \tag{1.1}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(Gauss's Law for magnetism from Biot-Savart)} \tag{1.2}$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \qquad \text{(Faraday's Law)} \tag{1.3}$$

$$\nabla \times \frac{\mathbf{B}}{\mu_0} - \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \mathbf{J} \qquad \text{(Ampere's Law revised by Maxwell)} \tag{1.4}$$

Here **E** and **B** represent electric and magnetic fields, respectively. The charge density ρ describes the *charge per volume* distributed through space. The current density **J** describes the *motion* of charge density (in units of ρ times velocity). The constant $\epsilon_0 = 8.854 \times 10^{-12}\ \mathrm{C}^2 \big/ \mathrm{N} \cdot \mathrm{m}^2$ is called the permittivity, and the constant $\mu_0 = 4\pi \times 10^{-7} \mathrm{T} \cdot \mathrm{m} \big/ \mathrm{A}$ (same as $\mathrm{kg} \cdot \mathrm{m} \big/ \mathrm{C}^2$) is called the permeability.

After introducing a key component into Ampere's law, Maxwell realized that together these equations comprise a complete self-consistent theory of electromagnetic phenomena. Moreover, the equations imply the existence of electromagnetic waves, which travel at the speed of light. Since the speed of light had been measured before Maxwell's time, it was immediately apparent (as was already suspected) that light is a high-frequency manifestation of the same phenomena that govern the influence of currents and charges upon each other. Previously, optics was considered to be a topic quite separate from electricity and magnetism.

In this chapter, we review the physical principles associated with each of Maxwell's equations. The main intent is to help students appreciate the connection between electromagnetic phenomena and light. While students need to understand and be able to use Maxwell's equations, many of the details presented in this chapter are not directly used in the study of optics.

[1] In Maxwell's original notation these equations were not so concise, and would have been hard fit onto a T-shirt. Lacking the convenience of modern vector notation, he wrote them as 20 equations in 20 variables.

James Clerk Maxwell

(1831–1879, Scottish)

Maxwell is best known for his fundamental contributions to electricity and magnetism and the kinetic theory of gases. He studied numerous other subjects, including the human perception of color and color-blindness, and is credited with producing the first color photograph. He originally postulated that electromagnetic waves propagated in a mechanical "luminiferous ether," but subsequent experiments have found this model untenable. He founded the Cavendish laboratory at Cambridge in 1874, which has produced 28 Nobel prizes to date.

1.2 Coulomb's and Gauss's Laws

The force on charge q located at $\mathbf{r}$ exerted by charge q' located at $\mathbf{r}'$ is

$$\mathbf{F} = q\mathbf{E} \tag{1.5}$$

where

$$\mathbf{E}(\mathbf{r}) = \frac{q'}{4\pi\epsilon_0} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \tag{1.6}$$

This relationship is known as Coulomb's law. The force is directed along the vector $\mathbf{r} - \mathbf{r}'$, which points from charge q' to q as seen in Fig. 1.1. The length or *magnitude* of this vector is given by $|\mathbf{r} - \mathbf{r}'|$ (i.e. the distance between q' and q). The familiar inverse square law can be seen by noting that $(\mathbf{r} - \mathbf{r}')/|\mathbf{r} - \mathbf{r}'|$ is a unit vector. We have written the force in terms of an electric field $\mathbf{E}(\mathbf{r})$, which is defined throughout space (regardless of whether

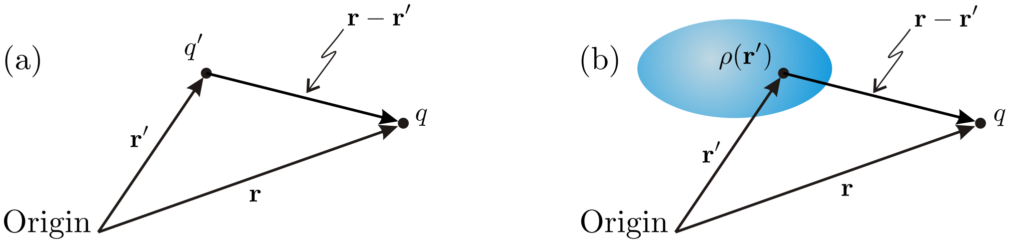


Figure 1.1 The geometry of Coulomb's law for (a) a point charge and (b) a charge distribution.

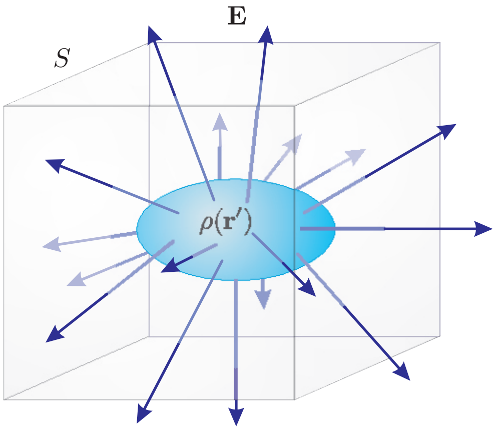


Figure 1.2 Gauss's law.

the second charge q is actually present). The permittivity ϵ_0 amounts to a proportionality constant.

The total force from a collection of charges is found by summing expression (1.5) over all charges q'_n associated with their specific locations $\mathbf{r}'_n$. If the charges are distributed continuously throughout space, having density $\rho(\mathbf{r}')$ (units of charge per volume), the summation for finding the net field at $\mathbf{r}$ becomes an integral:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int_V \rho(\mathbf{r}')\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\,dv' \tag{1.7}$$

This three-dimensional integral gives the net electric field produced by the charge density ρ distributed throughout the volume V.

Gauss's law follows directly from (1.7). By performing some mathematical operations on (1.7), we can demonstrate that the electric field uniquely satisfies the differential equation

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0} \tag{1.8}$$

(see appendix 1.A for details). No new physical phenomenon is introduced by writing Gauss's law. It is simply a mathematical interpretation of Coulomb's law.

The (perhaps more familiar) integral form of Gauss's law can be obtained by integrating (1.8) over a volume V and applying the divergence theorem (0.26) to the left-hand side:

$$\oint_S \mathbf{E}(\mathbf{r})\cdot\hat{\mathbf{n}}\,da = \frac{1}{\epsilon_0}\int_V \rho(\mathbf{r})\;dv \tag{1.9}$$

This form of Gauss's law shows that the total electric field flux extruding through a closed surface S (i.e. the integral on the left side) is proportional to the net charge contained within it (i.e. within volume V contained by S).

1.3 Biot-Savart and Ampere's Laws

The Biot-Savart law describes the force on a charged particle that comes about from a magnetic field. In this case, the charge q must move with a velocity (call it $\mathbf{v}$) in order to experience the force. The magnetic field arises itself from charges that are in motion. We consider a distribution of moving charges that form a current density throughout space. The moving charge distribution is described by a continuous current density $\mathbf{J}(\mathbf{r}')$ in units of charge times velocity per volume (or equivalently, current per cross sectional area). Analogous to (1.5) and (1.7), the Biot-Savart law is

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \tag{1.10}$$

where

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_V \mathbf{J}(\mathbf{r}') \times \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}\, dv' \tag{1.11}$$

(The latter equation is referred to as the Biot-Savart law; the first equation is known as the Lorentz force for a magnetic field.) The permeability μ_0 dictates the strength of the force, given the current distribution.

As before, we can apply mathematics to the Biot-Savart law to obtain another of Maxwell's equations. Nevertheless, the essential physics is already inherent in the Biot-Savart law. With the result from P 0.13, we can rewrite (1.11) as

$$\mathbf{B}(\mathbf{r}) = -\frac{\mu_0}{4\pi} \int_V \mathbf{J}(\mathbf{r}') \times \nabla_{\mathbf{r}} \frac{1}{|\mathbf{r} - \mathbf{r}'|}\, dv' = \frac{\mu_0}{4\pi} \nabla \times \int_V \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, dv' \tag{1.12}$$

Taking the divergence of this expression gives (see P 0.15)

$$\nabla \cdot \mathbf{B} = 0 \tag{1.13}$$

since the divergence of a curl is identically zero. This is another of Maxwell's equations (two down; two to go). The similarity between this equation and Gauss's law for electric fields (1.8) is apparent. In fact, (1.13) is known as Gauss's law for magnetic fields. In integral form, Gauss's law for magnetic fields looks like that for electric fields (1.9), only with zero on the right hand side. The law implies that the total magnetic flux extruding through any closed surface is zero (i.e. there will be as many field lines pointing inwards as pointing outwards). If one were to imagine the existence of magnetic "charges" (monopoles with either a north or south "charge"), then the right-hand side would not be zero. However, since magnetic charges have yet to be discovered, there is no point in introducing them.

It is interesting to show that the Biot-Savart law implies Ampere's law. Ampere's law is obtained by inverting the Biot-Savart law (1.11) so that $\mathbf{J}$ appears by itself, unfettered by integrals or the like. This is accomplished through mathematics, so again no new physical phenomenon is introduced, only a new interpretation. The mathematics for inverting (1.10) is given in Appendix 1.B. The result is

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \tag{1.14}$$

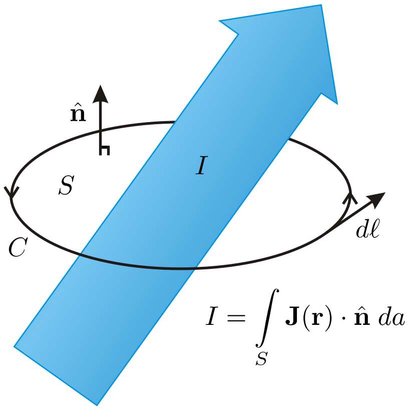

$\hat{\mathbf{n}}$
S
I
C
$d\ell$
$I = \int_S \mathbf{J}(\mathbf{r}) \cdot \hat{\mathbf{n}}\, da$

Figure 1.3 Ampere's law.

which is the differential form of Ampere's law. It is important to note that Ampere's law is valid only if the current density $\mathbf{J}$ does not vary rapidly in time. Specifically, to obtain (1.14) one must make the approximation

$$\nabla \cdot \mathbf{J} \cong 0 \qquad \text{(steady-state approximation)} \tag{1.15}$$

which in general is not true, especially for optical phenomena. We will discuss this further in section 1.4.

The (perhaps more familiar) integral form of Ampere's law can be obtained by integrating both sides of (1.14) over an open surface S, contained by contour C. Stokes' theorem (0.27) is applied to the left-hand side to get

$$\oint_C \mathbf{B}(\mathbf{r}) \cdot d\ell = \mu_0 \int_S \mathbf{J}(\mathbf{r}) \cdot \hat{\mathbf{n}}\, da \equiv \mu_0 I \tag{1.16}$$

This law says that the line integral of $\mathbf{B}$ around a closed loop C is proportional to the total current flowing through the loop (see Fig. 1.3). Recall that the units of $\mathbf{J}$ are current per area, so the surface integral containing $\mathbf{J}$ yields the current I in units of charge per time. In summary, the physics in Ampere's law is present in the Biot-Savart law. The laws are connected through mathematics.

1.4 Maxwell's Adjustment to Ampere's Law

Let's continue our discussion of Ampere's law and take up the possibility of a current density $\mathbf{J}$ that varies dynamically in time. Consider a volume of space enclosed by a surface S through which current is flowing. The total current exiting the volume is

$$I = \oint_S \mathbf{J} \cdot \hat{\mathbf{n}}\, da \tag{1.17}$$

The units on this equation are that of current, or charge per time, leaving the volume.

Since we have considered a *closed* surface S, the net current leaving the enclosed volume V must be the same as the rate at which charge within the volume vanishes:

$$I = -\frac{\partial}{\partial t} \int\limits_V \rho \, dv \tag{1.18}$$

Upon equating these two expressions for current, as well as applying the divergence theorem (0.26) to the former, we get

$$\int\limits_V \nabla \cdot \mathbf{J} \, dv = -\int\limits_V \frac{\partial \rho}{\partial t} \, dv \tag{1.19}$$

or

$$\int\limits_V \left(\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} \right) \, dv = 0 \tag{1.20}$$

which implies

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} \tag{1.21}$$

This is called a continuity equation. It is a statement of the conservation of charge as it flows. We derived it from the simple principle that the charge in a volume must decrease in time if we are to have a net current flowing out. This is not a concern in the steady-state situation (where Ampere's law applies) since in that case $\partial \rho / \partial t = 0$; a steady current has equal amounts of charge flowing both into and out of any particular volume.

Maxwell's main contribution (aside from organizing other people's formulas) was the injection of the continuity equation (1.21) into the derivation of Ampere's law to make it applicable to dynamical situations. As outlined in Appendix 1.B, the revised law becomes

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{1.22}$$

The final term is known as the *displacement current* (density), which exists even in the absence of any actual charge density ρ. A changing electric field behaves like a current in the sense that it produces magnetic fields. Notice the similarity to Faraday's law (1.26), which no doubt in part helped motivate Maxwell's work.

1.5 Faraday's Law

Michael Faraday discovered and characterized the relationship between changing magnetic fluxes and induced electric fields. Faraday showed that a change in magnetic flux through the area of a circuit loop (see Fig. 1.4) induces an electromotive force in the loop according to

$$\oint\limits_C \mathbf{E} \cdot d\boldsymbol{\ell} = -\frac{\partial}{\partial t} \int\limits_S \mathbf{B} \cdot \hat{\mathbf{n}} \, da \tag{1.23}$$

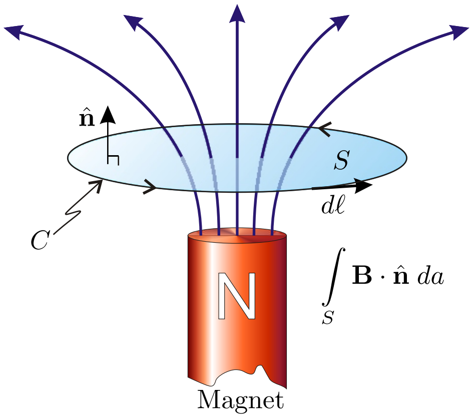


Figure 1.4 Faraday's law.

Faraday's law is one of Maxwell's equations. However, in (1.3) it is written in differential form. To obtain the differential form, we apply Stokes' theorem to the left-hand side and obtain

$$\int_S \nabla\times\mathbf{E}\cdot\hat{\mathbf{n}}\,da = -\frac{\partial}{\partial t}\int_S \mathbf{B}\cdot\hat{\mathbf{n}}\,da \tag{1.24}$$

or

$$\int_S \left(\nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t}\right)\cdot\hat{\mathbf{n}}\,da = 0 \tag{1.25}$$

This implies

$$\nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t} = 0 \tag{1.26}$$

which is the differential form of Faraday's law.

1.6 Polarization of Materials

We are essentially finished with our analysis of Maxwell's equations except for a brief examination of current density $\mathbf{J}$ and charge density ρ. The current density can be decomposed into three categories. The first category is associated with charges that are free to move, such as electrons in a metal. We will denote this type of current density by $\mathbf{J}_{\text{free}}$. The second category is associated with effective currents inside individual atoms that give rise to paramagnetic and diamagnetic effects. These are seldom important in optics problems, and so we will ignore these types of currents. The third type of current occurs when molecules in a material become polarized (i.e. elongate or orient as dipoles) in response to an applied electric field. We denote this type of current by $\mathbf{J}_{\text{p}}$ to distinguish it from free currents. The total current (ignoring magnetic effects) is then

$$\mathbf{J} = \mathbf{J}_{\text{free}} + \mathbf{J}_{\text{p}} \tag{1.27}$$

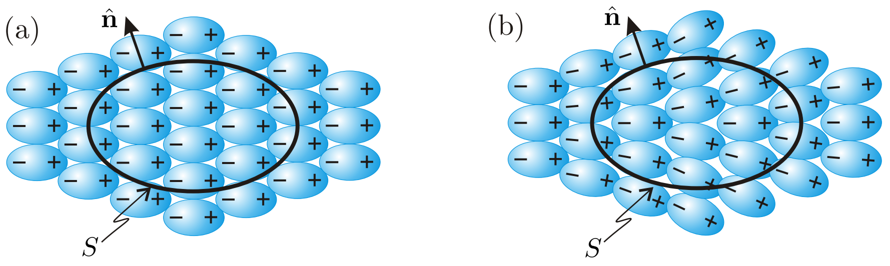


Figure 1.5 A polarized medium with (a) $\nabla \cdot \mathbf{P} = 0$ and with (b) $\nabla \cdot \mathbf{P} \neq 0$.

The polarization current $\mathbf{J}_\mathrm{p}$ is associated with a dipole distribution function $\mathbf{P}(\mathbf{r})$, called the *polarization* (in units of dipoles per volume, or charge times length per volume). Physically, if the dipoles (depicted in Fig. 1.5) change their orientation as a function of time in some coordinated fashion, an effective current density results. Since the time-derivative of dipole moments renders charge times velocity, a distribution of "sloshing" dipoles gives a current density equal to

$$\mathbf{J}_\mathrm{p} = \frac{\partial \mathbf{P}}{\partial t} \tag{1.28}$$

With this, Maxwell's equation (1.22) becomes

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \mathbf{J}_\mathrm{free} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} \tag{1.29}$$

Note that the combination $\mathbf{B}/\mu_0$ is sometimes written as $\mathbf{H}$.[2]

In the study of light and optics, we seldom consider the propagation of electromagnetic waveforms through electrically charged materials. In the case of no net charge, one might be tempted to set the right-hand side of Gauss's law (1.1) to zero. However, this would be wrong because neutral materials can become polarized, as described by $\mathbf{P}(\mathbf{r})$. The polarization can vary within a material, leading to local concentrations of positive or negative charge even though on average the material is neutral. This local buildup of charge due to the polarization current obeys the continuity equation (1.21):

$$\nabla \cdot \mathbf{J}_\mathrm{p} = -\frac{\partial \rho_\mathrm{p}}{\partial t} \tag{1.30}$$

Substitution of (1.28) into this equation yields an expression for the resulting charge density ρ_p:

$$\rho_\mathrm{p} = -\nabla \cdot \mathbf{P} \tag{1.31}$$

[2] This identification is only valid in non-magnetic materials—in magnetic materials $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$ where $\mathbf{M}$ is the material's magnetization.

To further appreciate local charge variation due to medium polarization, consider the divergence theorem (0.26) applied to $\mathbf{P}(\mathbf{r})$ in a neutral medium:

$$-\oint_S \mathbf{P}(\mathbf{r}) \cdot \hat{\mathbf{n}}\, da = -\int_V \nabla \cdot \mathbf{P}(\mathbf{r})\, dv \tag{1.32}$$

The left-hand side of (1.32) is a surface integral, which after integrating gives units of charge. Physically, it is the sum of the charges touching the inside of surface S (multiplied by a minus since dipole vectors point from the negatively charged end of a molecule to the positively charged end). The situation is depicted in Fig. 1.5. Keep in mind that $\mathbf{P}(\mathbf{r})$ is a continuous function so that Fig. 1.5 depicts crudely an enormous number of very tiny dipoles (no fair drawing a surface that avoids cutting the dipoles; cut through them at random). When $\nabla \cdot \mathbf{P}$ is zero, there are equal numbers of positive and negative charges touching S from within. When $\nabla \cdot \mathbf{P}$ is not zero, the positive and negative charges touching S are not balanced. Essentially, excess charge ends up within the volume because the non-uniform alignment of dipoles causes them to be cut preferentially at the surface. Since either side of (1.32) is equal to the excess charge inside the volume, $-\nabla \cdot \mathbf{P}$ may be interpreted as a charge density (it certainly has the right units—charge per volume), in agreement with (1.31). Again, the negative sign occurs since when $\mathbf{P}$ points out of the surface S, negative charges are left inside.

The total charge density thus can be written as

$$\rho = \rho_{\text{free}} + \rho_{\text{p}} \tag{1.33}$$

With (1.31), Gauss's law (1.8) becomes

$$\nabla \cdot (\epsilon_0 \mathbf{E} + \mathbf{P}) = \rho_{\text{free}} \tag{1.34}$$

where the combination $\epsilon_0 \mathbf{E} + \mathbf{P}$ is often called the displacement field, denoted by $\mathbf{D}$. For typical optics problems (involving neutral materials), we have $\rho_{\text{free}} = 0$.

1.7 The Macroscopic Maxwell Equations

In summary, in electrically neutral non-magnetic materials, Maxwell's equations are

$$\nabla \cdot \mathbf{E} = -\frac{\nabla \cdot \mathbf{P}}{\epsilon_0} \qquad \text{(Coulomb's law} \Rightarrow \text{Gauss's law)} \tag{1.35}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(Biot-Savart law} \Rightarrow \text{Gauss's law for magnetism)} \tag{1.36}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's law)} \tag{1.37}$$

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} + \mathbf{J}_{\text{free}} \qquad \text{(Ampere's law; fixed by Maxwell)} \tag{1.38}$$

Notice that we have dismissed the possibility of a free charge density ρ_{free} while we have retained the possibility of free current density $\mathbf{J}_{\text{free}}$. This is not a contradiction. In a neutral material, some charges may move differently than their oppositely charged counterparts, such as electrons versus ions in a metal. This gives rise to currents without the requirement of a net charge.

1.8 The Wave Equation

When Maxwell unified electromagnetic theory, he immediately noticed that waves are solutions to this set of equations. In fact his desire to find a set of equations that allowed for waves aided his effort to find the correct equations. After all, it was already known that light traveled as waves, Kirchhoff had previously noticed that $1\big/\sqrt{\epsilon_0\mu_0}$ gives the correct speed of light $c = 3.00\times10^8$ m/s (which had already been measured), and Faraday and Kerr had observed that strong magnetic and electric fields affect light propagating in crystals. At first glance, Maxwell's equations might not immediately suggest (to the inexperienced eye) that waves are solutions. However, we can manipulate the equations (first order differential equations coupling **E** and **B**) into the familiar wave equation (second order differential equations for either **E** or **B**, decoupled).

We will derive the wave equation for **E**. The derivation of the wave equation for **B** is very similar (see problem P 1.7). We begin our derivation by taking the curl of (1.37), from which we obtain

$$\nabla\times(\nabla\times\mathbf{E})+\frac{\partial}{\partial t}(\nabla\times\mathbf{B})=0 \tag{1.39}$$

The equation can be simplified with the differential vector identity (see P 0.16):

$$\nabla\times(\nabla\times\mathbf{E})=\nabla(\nabla\cdot\mathbf{E})-\nabla^2\mathbf{E} \tag{1.40}$$

In addition, we can make a substitution for $\nabla\times\mathbf{B}$ from (1.38). Together, these substitutions give

$$\nabla(\nabla\cdot\mathbf{E})-\nabla^2\mathbf{E}+\frac{\partial}{\partial t}\left(\epsilon_0\mu_0\frac{\partial\mathbf{E}}{\partial t}+\mu_0\mathbf{J}_{\text{free}}+\mu_0\frac{\partial\mathbf{P}}{\partial t}\right)=0 \tag{1.41}$$

Applying (1.35) to the first term, and after rearranging, we get

$$\nabla^2\mathbf{E}-\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2}=\mu_0\frac{\partial\mathbf{J}_{\text{free}}}{\partial t}+\mu_0\frac{\partial^2\mathbf{P}}{\partial t^2}-\frac{1}{\epsilon_0}\nabla(\nabla\cdot\mathbf{P}) \tag{1.42}$$

The left-hand side of (1.42) is the familiar wave equation. However, the right-hand side contains a number of source terms, which arise when various currents and polarizations are present. The first term on the right-hand side of (1.42) describes electric currents, which are important for determining the reflection of light from a metallic surface or for determining the propagation of light within a plasma. The second term on the right-hand side of (1.42) describes dipole oscillations, which behave similar to currents. In a non-conducting optical material such as glass, the free current is zero, but $\partial^2\mathbf{P}\big/\partial t^2$ is not zero, as the medium polarization responds to the light field. This polarization current determines the refractive index of the material (discussed in chapter 2). The final term on the right-hand side of (1.42) is important in non-isotropic media such as a crystal. In this case, the polarization **P** responds to the electric field along a direction not necessarily parallel to **E**, due to the influence of the crystal lattice (addressed in chapter 5). For most problems in optics, some of the terms on the right-hand side of (1.42) are zero. However, usually at least one of the terms must be retained when considering propagation in a medium other than vacuum.

In vacuum all of the terms on the right-hand side in (1.42) are zero, in which case the equation reduces to

$$\nabla^2\mathbf{E}-\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2}=0 \qquad \text{(vacuum)} \tag{1.43}$$

The solutions to the vacuum wave equation (1.43) propagate with speed

$$c \equiv 1 \big/ \sqrt{\epsilon_0 \mu_0} = 2.9979 \times 10^8 \text{ m/s} \qquad \text{(vacuum)} \tag{1.44}$$

and any function $\mathbf{E}$ is a valid solution as long as it caries the dependence on the argument $\hat{\mathbf{u}} \cdot \mathbf{r} - ct$, where $\hat{\mathbf{u}}$ is a unit vector specifying the direction of propagation. The argument $\hat{\mathbf{u}} \cdot \mathbf{r} - ct$ preserves the shape of the waveform as it propagates in the $\hat{\mathbf{u}}$ direction; features occurring at a given position recur 'downstream' at a distance ct after a time t. By checking this solution in (1.43), one effectively verifies that the speed of propagation is c (see P 1.9). Note that we may add together any combination of solutions (even with differing directions of propagation) to form other valid solutions. In most situations we multiply the argument $\hat{\mathbf{u}} \cdot \mathbf{r} - ct$ by a constant k (known as the wave number) that has units of inverse length to obtain the dimensionless form of the argument:

$$k \left(\hat{\mathbf{u}} \cdot \mathbf{r} - ct \right) = \mathbf{k} \cdot \mathbf{r} - \omega t \tag{1.45}$$

where $\mathbf{k} \equiv k\hat{\mathbf{u}}$ and we have defined the *vacuum dispersion relation*

$$\omega \equiv kc \qquad \text{(vacuum)} \tag{1.46}$$

After solving the wave equation (1.42) for $\mathbf{E}$, one may obtain $\mathbf{B}$ through an application of Faraday's law (1.37). Even though the magnetic field $\mathbf{B}$ satisfies a similar wave equation, decoupled from $\mathbf{E}$ (see P 1.7), the two waves are not independent. The fields for $\mathbf{E}$ and $\mathbf{B}$ must be chosen to be consistent with each other through Maxwell's equations.

Appendix 1.A Derivation of Gauss's Law

To derive Gauss's law, we take the divergence of (1.7):

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int\limits_V \rho\left(\mathbf{r}'\right) \nabla_{\mathbf{r}} \cdot \frac{\left(\mathbf{r} - \mathbf{r}'\right)}{\left|\mathbf{r} - \mathbf{r}'\right|^3} \, dv' \tag{1.47}$$

The subscript on $\nabla_{\mathbf{r}}$ indicates that it operates on $\mathbf{r}$ while treating $\mathbf{r}'$ as a constant. As messy as this integral appears, it contains a remarkable mathematical property that can be exploited, even without specifying the form of the charge distribution $\rho\left(\mathbf{r}'\right)$. In modern mathematical language, the vector expression in the integral is a three-dimensional delta function:

$$\nabla_{\mathbf{r}} \cdot \frac{\left(\mathbf{r} - \mathbf{r}'\right)}{\left|\mathbf{r} - \mathbf{r}'\right|^3} \equiv 4\pi\delta^3\left(\mathbf{r}' - \mathbf{r}\right) \equiv 4\pi\delta\left(x' - x\right)\delta\left(y' - y\right)\delta\left(z' - z\right) \tag{1.48}$$

A derivation of this formula and a description of its properties are addressed in problem P 0.23. The delta function allows the integral in (1.47) to be performed, and the relation becomes simply

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\epsilon_0} \tag{1.49}$$

which is the differential form of Gauss's law.

Appendix 1.B Derivation of Ampere's Law

To obtain Ampere's law from the Biot-Savart law, we take the curl of (1.11):

$$\nabla \times \mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_V \nabla_{\mathbf{r}} \times \left[\mathbf{J}(\mathbf{r}') \times \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \right] dv' \tag{1.50}$$

We next apply the differential vector rule from P 0.17 while noting that $\mathbf{J}(\mathbf{r}')$ does not depend on $\mathbf{r}$ so that only two terms survive. The curl of $\mathbf{B}(\mathbf{r})$ then becomes

$$\nabla \times \mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_V \left(\mathbf{J}(\mathbf{r}') \left[\nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \right] - \left[\mathbf{J}(\mathbf{r}') \cdot \nabla_{\mathbf{r}} \right] \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \right) dv' \tag{1.51}$$

According to (1.48), the first term in the integral is $4\pi \mathbf{J}(\mathbf{r}') \delta^3 (\mathbf{r}' - \mathbf{r})$, which is easily integrated. To make progress on the second term, we observe that the gradient can be changed to operate on the primed variables without affecting the final result (i.e. $\nabla_{\mathbf{r}} \to -\nabla_{\mathbf{r}'}$). In addition, we take advantage of the vector integral theorem (0.28) to arrive at

$$\nabla \times \mathbf{B}(\mathbf{r}) = \mu_0 \mathbf{J}(\mathbf{r}) - \frac{\mu_0}{4\pi} \int_V \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \left[\nabla_{\mathbf{r}'} \cdot \mathbf{J}(\mathbf{r}') \right] dv' + \frac{\mu_0}{4\pi} \oint_S \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \left[\mathbf{J}(\mathbf{r}') \cdot \hat{\mathbf{n}} \right] da' \tag{1.52}$$

The last term in (1.52) vanishes if we assume that the current density $\mathbf{J}$ is completely contained within the volume V so that it is zero at the surface S. Thus, the expression for the curl of $\mathbf{B}(\mathbf{r})$ reduces to

$$\nabla \times \mathbf{B}(\mathbf{r}) = \mu_0 \mathbf{J}(\mathbf{r}) - \frac{\mu_0}{4\pi} \int_V \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \left[\nabla_{\mathbf{r}'} \cdot \mathbf{J}(\mathbf{r}') \right] dv' \tag{1.53}$$

The latter term in (1.53) vanishes if $\nabla \cdot \mathbf{J} \cong 0$, yielding Ampere's law (1.14). Maxwell was the first to realize that this term must be retained in dynamical situations. Injection of the continuity equation (1.21) into (1.53) yields

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \frac{\mu_0}{4\pi} \frac{\partial}{\partial t} \int_V \rho(\mathbf{r}') \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \, dv' \tag{1.54}$$

Finally, substitution of (1.7) into this formula gives

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \tag{1.55}$$

the generalized form of Ampere's law.

Exercises

1.1 Introduction

P1.1 Suppose that an electric field is given by $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 \cos(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi)$, where $\mathbf{k}\perp\mathbf{E}_0$ and ϕ is a constant phase. Show that $\mathbf{B}(\mathbf{r},t) = \frac{\mathbf{k}\times\mathbf{E}_0}{\omega}\cos(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi)$ is consistent with (1.3).

1.4 Maxwell's Adjustment to Ampere's Law

P1.2 (a) Use Gauss's law to find the electric field in the gap shown in Fig. 1.6. Assume that the cross-sectional area of the wire A is much wider than the gap separation d. Let the accumulated charge on the "plates" be Q. HINT: The E-field is essentially zero except in the gap.

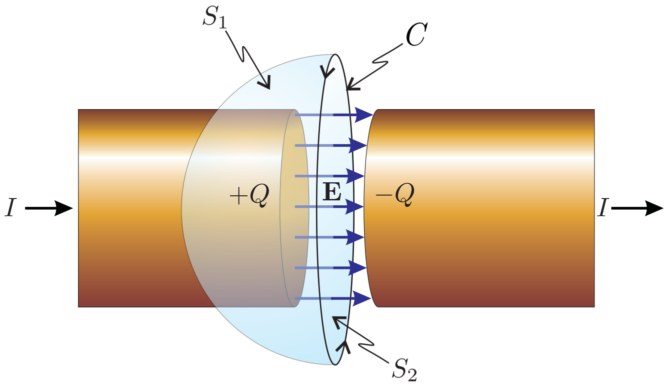


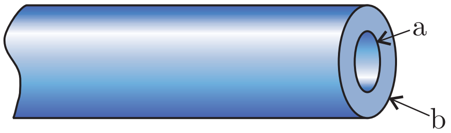
Figure 1.6 Charging capacitor.

(b) Find the strength of the magnetic field on contour C using Ampere's law applied to surface S_1. Let the current in the wire be I.

(c) Show that the displacement current leads to the identical magnetic field when using surface S_2.

HINT: Multiply $\epsilon_0 \partial \mathbf{E}/\partial t$ by the cross-sectional area to obtain a "current". The current in the wire is related to the charge Q through $I = \partial Q/\partial t$.

P1.3 Consider an infinitely long hollow cylinder (inner radius a, outer radius b) which carries a volume charge density $\rho = k/s^2$ for $a < s < b$ and no charge elsewhere, where s is the distance from the axis of the cylinder as shown in Fig. 1.7.



charge is located between a and b

Figure 1.7 A charged cylinder

Use Gauss's Law in integral form to find the electric field produced by this charge for each of the three regions: $s < a$, $a < s < b$, and $s > b$.

HINT: For each region first draw an appropriate "Gaussian surface" and integrate the charge density over the volume to figure out the enclosed charge. Then use Gauss's law in integral form and the symmetry of the problem to solve for the electric field.

P1.4 A conducting cylinder with the same geometry as P 1.3 carries a volume current density $\mathbf{J} = k/s\hat{\mathbf{z}}$ (along the axis of the cylinder) for $a < s < b$. Using Ampere's Law in integral form, find the magnetic field due to this current. Find the field for each of the three regions: $s < a$, $a < s < b$, and $s > b$.

HINT: For each region first draw an appropriate "Amperian loop" and integrate the current density over the surface to figure out how much current passes through the loop. Then use Ampere's law in integral form and the symmetry of the problem to solve for the magnetic field.

1.7 The Macroscopic Maxwell Equations

P1.5 Memorize the macroscopic Maxwell equations and be prepared to reproduce them from memory on an exam. Write them from memory in your homework to indicate that you have completed this problem.

P1.6 For the fields given in P 1.1, what are the implications for $\mathbf{J}_{\text{free}} + \partial \mathbf{P}/\partial t$?

1.8 The Wave Equation

P1.7 Derive the wave equation for the magnetic field $\mathbf{B}$ in vacuum (i.e. $\mathbf{J}_{\text{free}} = 0$ and $\mathbf{P} = 0$).

P1.8 Show that the magnetic field in P 1.1 is consistent with the wave equation derived in P 1.7.

P1.9 Check that $\mathbf{E}\left(\hat{\mathbf{u}} \cdot \mathbf{r} - ct\right)$ satisfies the vacuum wave equation (1.43), where $\mathbf{E}$ is an arbitrary functional form.

P1.10 (a) Show that $\mathbf{E}\left(\mathbf{r}, t\right) = \mathbf{E}_0 \cos\left(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi\right)$ is a solution to (1.43) if the dispersion relation (1.46) holds.

(b) Show that each wave front forms a plane, which is why such solutions are often called 'plane waves'. HINT: A wavefront is a surface in space where the argument of the cosine (i.e., the *phase* of the wave) has a constant value. Set the cosine argument to an arbitrary constant and see what positions are associated with that phase.

(c) Determine the speed $v = \Delta r/\Delta t$ that a wave front moves in the $\mathbf{k}$ direction. HINT: Set the cosine argument to a constant, solve for $\mathbf{r}$, and differentiate.

(d) By analysis, determine the wavelength λ in terms of k and in terms of ω and c. HINT: Find the distance between identical wave fronts by changing the cosine argument by 2π at a given instant in time.

(e) Use (1.35) to show that $\mathbf{E}_0$ and $\mathbf{k}$ must be perpendicular to each other in vacuum.

P1.11 If $\mathbf{E} = (7x^2y^3\hat{x} + 2z^4\hat{y})\cos\omega t$

(a) Find $\rho(x, y, z, t)$

(b) Find $\frac{\partial \mathbf{B}(x,y,z,t)}{\partial t}$

(c) Determine if $\mathbf{E}$ is a solution to the vacuum wave equation, (1.43).

P1.12 Determine the speed of the wave crests of a simple plane wave: $f = \cos(kx - t)$. Do this by figuring out how far a give wave crest has moved between times t and $t + \Delta t$.

L1.13 Measure the speed of light using a rotating mirror. Provide an estimate of the experimental uncertainty in your answer (not the percentage error from the known value).

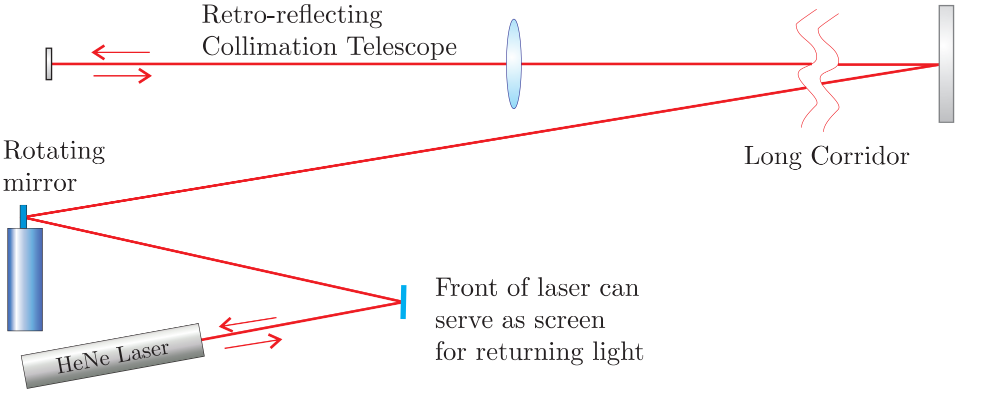


Figure 1.8 A schematic of the setup for lab 1.13.

Figure 1.9 shows a simplified geometry for the optical path for light in this experiment. Laser light from A reflects from a rotating mirror at B towards C. The light returns to B, where the mirror has rotated, sending the light to point D. Notice that a mirror rotation of θ deflects the beam by 2θ.

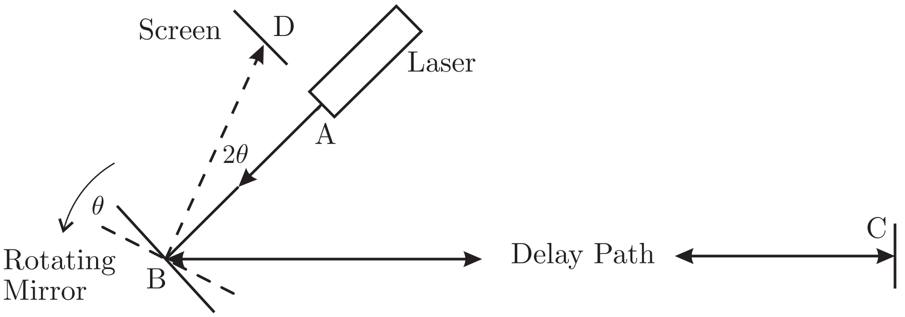
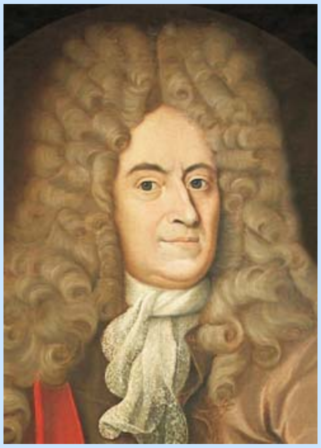


Figure 1.9 Geometry for lab 1.13.



Ole Roemer

(1644–1710, Danish)

Roemer was a man of many interests. In addition to measuring the speed of light, he created a temperature scale which with slight modification became the Fahrenheit scale, introduced a system of standard weights and measures, and was heavily involved in civic affairs (city planning, etc.). Scientists initially became interested in Io's orbit because its eclipse (when it went behind Jupiter) was an event that could be seen from many places on earth. By comparing accurate measurements of the local time when Io was eclipsed by Jupiter at two remote places on earth, scientists in the 1600's were able to determine the longitude difference between the two places.

P1.14 Ole Roemer made the first successful measurement of the speed of light in 1676 by observing the orbital period of Io, a moon of Jupiter with a period of 42.5 hours. When Earth is moving toward Jupiter, the period is measured to be shorter than 42.5 hours because light indicating the end of the moon's orbit travels less distance than light indicating the beginning. When Earth is moving away from Jupiter, the situation is reversed, and the period is measured to be longer than 42.5 hours.

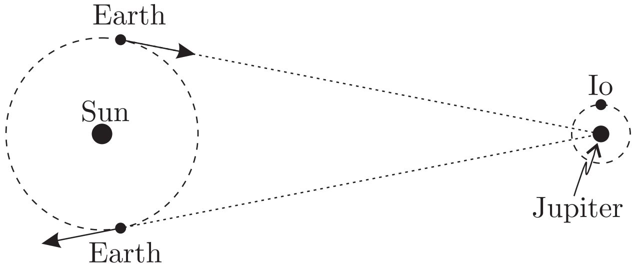


Figure 1.10

(a) If you were to measure the time for 40 observed orbits of Io when Earth is moving directly toward Jupiter and then several months later measure the time for 40 observed orbits when Earth is moving directly away from Jupiter, what would you expect the difference between these two measurements be? Take the Earth's orbital radius to be 1.5×10^{11} m. (To simplify the geometry, just assume that Earth move directly toward or away from Jupiter over the entire 40 orbits.)

(b) Roemer actually did the experiment described in part (a), and experimentally measured a 22 minute difference. What speed of light would one deduce from that value?

P1.15 In an isotropic medium (i.e. $\nabla \cdot \mathbf{P} = 0$), the polarization can often be written as function of the electric field: $\mathbf{P} = \epsilon_0 \chi(E)\,\mathbf{E}$, where $\chi(E) = \chi_1 + \chi_2 E + \chi_3 E^2 \cdots$. The higher order coefficients in the expansion (i.e. χ_2, χ_3, ...) are typically small, so only the first term is important at low intensities. The field of nonlinear optics deals with intense light-matter interactions, where the higher order terms of the expansion are important. This can lead to phenomena such as harmonic generation.

Starting with Maxwell's equations, derive the wave equation for nonlinear optics in an isotropic medium:

$$\nabla^2 \mathbf{E} - \mu_0 \epsilon_0 (1 + \chi_1) \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mu_0 \epsilon_0 \frac{\partial^2 \left(\chi_2 E + \chi_3 E^2 + \cdots\right) \mathbf{E}}{\partial t^2} + \mu_0 \frac{\partial \mathbf{J}}{\partial t}$$

We have retained the possibility of current here since, for example, in a gas some of the molecules might ionize in the presence of a strong field, giving rise to a current.

Chapter 2

Plane Waves and Refractive Index

2.1 Introduction

In this chapter we consider the interaction between matter and sinusoidal waves called plane waves. We also consider the energy carried by such waves. In section 2.6, we introduce Poynting's theorem, which governs the flow of energy carried by electromagnetic fields. This leads to the concept of irradiance (or intensity), which we discuss in the plane-wave context in section 2.7.

We will primarily restrict our attention to sinusoidal solutions to Maxwell's equations. This may seem somewhat limiting at first, since (as mentioned in chapter 1) any waveform can satisfy the wave equation in vacuum (and therefore Maxwell's equations) as long as it travels at c and has appropriate connections between $\mathbf{E}$ and $\mathbf{B}$. It turns out, however, that an arbitrary waveform can be constructed from a linear superposition of sinusoidal waves. Thus, we can model the behavior of more complicated waveforms by considering the behavior of many sinusoidal waves and then summing them to produce the desired waveform. The ability to treat the frequency components of a waveform separately is essential when considering the propagation of light within a material medium, since materials respond differently to different frequencies of light. As a result, a waveform propagating in a material medium invariably changes its shape as it travels (a phenomenon called dispersion) unless that waveform is a pure sinusoidal wave. This is why physicists and engineers choose to work with sinusoidal waves.

When describing light, it is convenient to use complex number notation. This is particularly true for problems involving absorption of light such as what takes place inside metals and, to a lesser degree (usually), inside dielectrics (e.g. glass). In such cases, oscillatory fields decay as they travel, owing to absorption. In chapter 4, we will see that this absorption rate plays an important role in the reflectance of light from metal surfaces. We will introduce complex electric field waves in section 2.2. When the electric field is represented using complex notation, the phase parameter $\mathbf{k} \cdot \mathbf{r}$ also becomes a complex number. The imaginary part controls the rate at which the field decays, while the real part governs the familiar oscillatory behavior. In section 2.3 we introduce the complex index of refraction $\mathcal{N} \equiv n + i\kappa$. The complex index only makes sense when the electric field is also expressed using complex notation. (Don't be alarmed at this point if this seems puzzling.)

To compute the index of refraction in either a dielectric or a conducting material, we require a model to describe the response of electrons in the material to the passing electric field wave. Of course, the model in turn influences how the electric field propagates, which is what influences the material in the first place! The model therefore must be solved together with the propagating field in a self-consistent manner. Henry Lorentz developed a very successful model in the late 1800's, which treats each (active) electron in the medium as a classical particle obeying Newton's second law ($\mathbf{F} = m\mathbf{a}$). In the case of a dielectric medium, electrons are subject to an elastic restoring force (that keeps each electron bound to its respective atom) in addition to a damping force, which dissipates energy and gives rise to absorption. In the case of a conducting medium, electrons are free to move outside of atoms but they are still subject to a damping force (due to collisions), which removes energy and gives rise to absorption.

2.2 Plane Wave Solutions to the Wave Equation

Consider the wave equation for an electric field waveform propagating in vacuum (1.43):

$$\nabla^2 \mathbf{E} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 \tag{2.1}$$

We are interested in solutions to (2.1) that have the functional form (see P 1.10)

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \cos\left(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi\right) \tag{2.2}$$

Here ϕ represents an arbitrary (constant) phase term. The vector $\mathbf{k}$ may be written as

$$\mathbf{k} \equiv \frac{2\pi}{\lambda_{\text{vac}}} \hat{\mathbf{u}} \qquad \text{(vacuum)} \tag{2.3}$$

where $\hat{\mathbf{u}}$ is a unit vector defining the direction of propagation, and λ_{vac} is the length by which $\mathbf{r}$ must vary to cause the cosine to go through a complete cycle. This distance is known as the (vacuum) wavelength. The frequency of oscillation is related to the wavelength via

$$\omega = \frac{2\pi c}{\lambda_{\text{vac}}} \qquad \text{(vacuum)} \tag{2.4}$$

Notice that k and ω are not independent of each other but form a pair. $\mathbf{k}$ is called the wave vector. Typical values for λ_{vac} are given in table 2.1. Sometimes the spatial period of the wave is expressed as $1/\lambda_{\text{vac}}$, in units of cm^{-1}, called the wave number.

A magnetic wave accompanies any electric wave, and it obeys a similar wave equation (see P 1.7). The magnetic wave corresponding to (2.2) is

$$\mathbf{B}(\mathbf{r}, t) = \mathbf{B}_0 \cos\left(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi\right), \tag{2.5}$$

but it is important to note that $\mathbf{B}_0$, $\mathbf{k}$, ω, and ϕ are not independently chosen. In order to satisfy Faraday's law (1.3), the arguments of the cosine in (2.2) and (2.5) must be identical. In addition, Faraday's law requires (see P 1.1)

$$\mathbf{B}_0 = \frac{\mathbf{k} \times \mathbf{E}_0}{\omega} \tag{2.6}$$

	Frequency $\nu = \omega/2\pi$	Wavelength λ_{vac}
AM Radio	10^6 Hz	300 m
FM Radio	10^8 Hz	3 m
Radar	10^{10} Hz	0.03 m
Microwave	$10^9 - 10^{12}$ Hz	0.3 m- 3×10^{-4} m
Infrared	$10^{12} - 4 \times 10^{14}$ Hz	$3 \times 10^{-4} - 7 \times 10^{-7}$ m
Light (red)	4.6×10^{14} Hz	6.5×10^{-7} m
Light (yellow)	5.5×10^{14} Hz	5.5×10^{-7} m
Light (blue)	6.7×10^{14} Hz	4.5×10^{-7} m
Ultraviolet	$10^{15} - 10^{17}$ Hz	$4 \times 10^{-7} - 3 \times 10^{-9}$ m
X-rays	$10^{17} - 10^{20}$ Hz	$3 \times 10^{-9} - 3 \times 10^{-12}$ m
Gamma rays	$10^{20} - 10^{23}$ Hz →	$3 \times 10^{-12} - 3 \times 10^{-15}$ m →

Table 2.1 The electromagnetic spectrum.

In vacuum, the electric and magnetic fields travel in phase. They are directed perpendicular to each other as defined by the cross product in (2.6). Since both fields are also perpendicular to the direction of propagation, given by **k**, the magnitudes of the field vectors are related by $B_0 = kE_0/\omega$ or $B_0 = E_0/c$ in view of (1.46). Although the fields in Fig. 2.1 are drawn like transverse waves on a string, they are actually large planar sheets containing uniform fields (different fields in different planes) that move in the direction of **k**.

The magnetic field can be ignored in most optics problems. The influence of the magnetic field only becomes important (in comparison to the electric field) for charged particles moving near the speed of light. This typically takes place only for extremely intense lasers (intensities above 10^{18} W/cm^2, see P 2.14) where the electric field is sufficiently strong to cause electrons to oscillate with velocities near the speed of light. Throughout the remainder of this book, we will focus our attention mainly on the electric field with the understanding that we can at any time deduce the (less important) magnetic field from the electric field

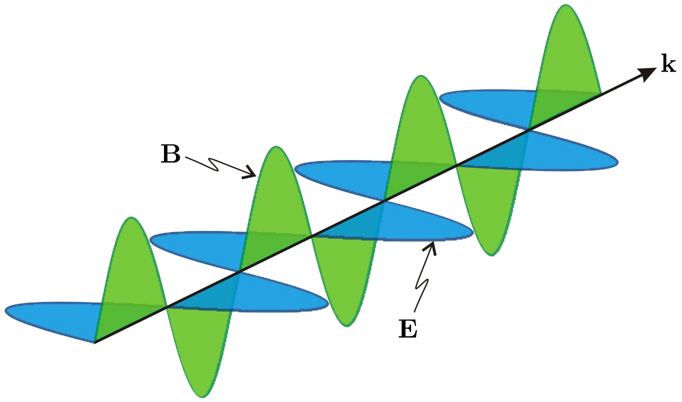


Figure 2.1 Depiction of electric and magnetic fields associated with a plane wave.

via Faraday's law.

We next check our solution (2.2) in the wave equation. First, however, we will adopt complex number notation. (For a review of complex notation, see section 0.1.) Although this change in notation will not make the task at hand any easier, we introduce it here in preparation for sections 2.3 and 2.5 where it will save considerable labor. Using complex notation we rewrite (2.2) as

$$\mathbf{E}(\mathbf{r}, t) = \mathrm{Re}\left\{\tilde{\mathbf{E}}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\right\} \tag{2.7}$$

where we have hidden the phase term ϕ inside of $\tilde{\mathbf{E}}_0$ as follows:

$$\tilde{\mathbf{E}}_0 \equiv \mathbf{E}_0 e^{i\phi} \tag{2.8}$$

The next step we take is to become intentionally sloppy. Physicists throughout the world have conspired to avoid writing $\mathrm{Re}\{\}$ in an effort (or lack thereof if you prefer) to make expressions less cluttered. Nevertheless, only the real part of the field is physically relevant even though expressions and calculations contain both real and imaginary terms. This sloppy notation is okay since the real and imaginary parts of complex numbers never intermingle when adding, subtracting, differentiating, or integrating. We can delay taking the real part of the expression until the end of the calculation. Also, when hiding a phase ϕ inside of the field amplitude as in (2.7), we drop the tilde (might as well since we are already being sloppy); when using complex notation, we will automatically assume that the complex field amplitude contains phase information.

Our solution (2.2) or (2.7) is written simply as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.9}$$

which is referred to as a *plane wave*. It is possible to construct any electromagnetic disturbance from a linear superposition of such waves. The name *plane wave* is given since the argument in (2.7) at any moment is constant (and hence the electric field is uniform) across planes that are perpendicular to $\mathbf{k}$. A plane wave fills all space and may be thought of as a series of infinite sheets of uniform electric field moving in the $\mathbf{k}$ direction.

Finally, we verify (2.9) as a solution to the wave equation (2.1). The first term gives

$$\begin{aligned}\nabla^2 \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} &= \mathbf{E}_0 \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right] e^{i(k_x x + k_y y + k_z z - \omega t)} \\ &= -\mathbf{E}_0 \left(k_x^2 + k_y^2 + k_z^2\right) e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ &= -k^2 \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\end{aligned} \tag{2.10}$$

and the second term gives

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left(\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\right) = -\frac{\omega^2}{c^2}\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.11}$$

Upon insertion into (2.1) we obtain the vacuum dispersion relation (1.46), which specifies the connection between the wavenumber k and the frequency ω. While the vacuum dispersion relation is simple, it emphasizes that k and ω cannot be independently chosen (as we saw in (2.3) and (2.4)).

2.3 Index of Refraction in Dielectrics

Let's take a look at how plane waves behave in dielectric media (e.g. glass). We assume an isotropic, homogeneous, and non-conducting medium (i.e. $\mathbf{J}_{\text{free}} = 0$). In this case, we expect $\mathbf{E}$ and $\mathbf{P}$ to be parallel to each other so $\nabla \cdot \mathbf{P} = 0$ from (1.35). The general wave equation (1.42) for the electric field reduces in this case to

$$\nabla^2 \mathbf{E} - \epsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mu_0 \frac{\partial^2 \mathbf{P}}{\partial t^2} \tag{2.12}$$

Since we are considering sinusoidal waves, we consider solutions of the form

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \\ \mathbf{P} &= \mathbf{P}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \end{aligned} \tag{2.13}$$

By writing this, we are making the (reasonable) assumption that if an electric field stimulates a medium at frequency ω, then the polarization in the medium also oscillates at frequency ω. This assumption is typically rather good except when extreme electric fields are used (see P 1.15). Substitution of the trial solutions (2.13) into (2.12) yields

$$-k^2 \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} + \epsilon_0 \mu_0 \omega^2 \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} = -\mu_0 \omega^2 \mathbf{P}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \tag{2.14}$$

In a *linear* medium (essentially any material if the electric field strength is not extreme), the polarization amplitude is proportional to the strength of the applied electric field:

$$\mathbf{P}_0(\omega) = \epsilon_0 \chi(\omega) \mathbf{E}_0(\omega) \tag{2.15}$$

We have introduced a dimensionless proportionality factor $\chi(\omega)$ called the susceptibility, which depends on the frequency of the field. With this, we can obtain the dispersion relation in dielectrics from (2.14):

$$k^2 = \epsilon_0 \mu_0 \left[1 + \chi(\omega)\right] \omega^2 \tag{2.16}$$

or

$$k = \frac{\omega}{c} \sqrt{1 + \chi(\omega)} \tag{2.17}$$

where we have used $c \equiv 1/\sqrt{\epsilon_0 \mu_0}$. By direct comparison with vacuum case (**??**), we see that the speed of the sinusoidal wave in the material is

$$v = c/n(\omega) \tag{2.18}$$

where

$$n(\omega) \cong \sqrt{1 + \chi(\omega)} \qquad \text{(negligible absorption)} \tag{2.19}$$

The dimensionless quantity $n(\omega)$, called the *index of refraction*, is the ratio of the speed of the light in vacuum to the speed of the wave in the material. (Note that the wave speed v is a function of frequency.) The index of refraction is a function of the material and of the frequency of the light.

In general the susceptibility $\chi(\omega)$ is a complex number, which allows $\mathbf{P}_0$ to have a different phase from $\mathbf{E}_0$ in (2.15). When absorption is small we can neglect the imaginary

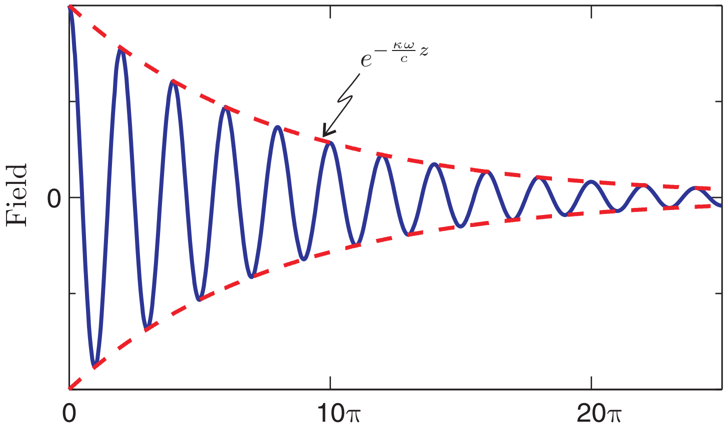


Figure 2.2 Electric field of a decaying plane wave.

part of $\chi(\omega)$, as we have done in (2.19). However, in cases where absorption plays a role, we must use the complex index of refraction, defined by

$$\mathcal{N} \equiv (n + i\kappa) = \sqrt{1 + \chi(\omega)} \tag{2.20}$$

where n and κ are respectively the real and imaginary parts of the index. (Note that κ is not k.) According to (2.17), the magnitude of the wave vector is

$$k = \frac{\mathcal{N}\omega}{c} = \frac{(n + i\kappa)\,\omega}{c} \tag{2.21}$$

which is a complex value. The complex index $\mathcal{N}$ takes account of absorption as well as the usual oscillatory behavior of the wave. We see this by explicitly placing (2.21) into (2.13):

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{-\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} e^{i\left(\frac{n\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r} - \omega t\right)} \tag{2.22}$$

As before, here $\hat{\mathbf{u}}$ is a *real* unit vector specifying the direction of $\mathbf{k}$.

As a reminder, when looking at (2.22), by special agreement in advance, we should just think of the real part, namely

$$\begin{aligned} \mathbf{E}(\mathbf{r}, t) &= \tilde{\mathbf{E}}_0 e^{-\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} \cos\left(\frac{n\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r} - \omega t\right) \\ &= \mathbf{E}_0 e^{-\mathrm{Im}\{\mathbf{k}\}\cdot\mathbf{r}} \cos\left(\mathrm{Re}\left\{\mathbf{k}\right\}\cdot\mathbf{r} - \omega t + \phi\right) \end{aligned} \tag{2.23}$$

where the phase ϕ was formerly held in the complex vector $\tilde{\mathbf{E}}_0$ (where the tilde had been suppressed). Fig. 2.2 shows a graph of the exponent and cosine factor in (2.22). For convenience in plotting, the direction of propagation is chosen to be in the z direction (i.e. $\hat{\mathbf{u}} = \hat{\mathbf{z}}$). The imaginary part of the index κ causes the wave to decay as it travels. The real part of the index n is associated with the oscillations of the wave.

In a dielectric, the vacuum relations (2.3) and (2.4) are modified to read

$$\mathrm{Re}\left\{\mathbf{k}\right\} \equiv \frac{2\pi}{\lambda}\hat{\mathbf{u}}, \tag{2.24}$$

and

$$\omega = \frac{2\pi c}{\lambda n}, \tag{2.25}$$

where

$$\lambda \equiv \lambda_{\text{vac}}/n. \tag{2.26}$$

While the frequency ω is the same, whether in a material or in vacuum, the wavelength λ is different as indicated by (2.26).

As a final note, for the sake of simplicity in writing (2.23) we assumed linearly polarized light. That is, all vector components of $\mathbf{E}_0$ were assumed to have the same complex phase ϕ. The expression would be somewhat more complicated, for example, in the case of circularly polarized light (described in chapter 4).

2.4 The Lorentz Model of Dielectrics

In this section, we develop a simple linear model for describing refractive index. The model determines the susceptibility $\chi(\omega)$, the connection between the electric field $\mathbf{E}_0$ and the polarization $\mathbf{P}_0$. Lorentz introduced this model well before the development of quantum mechanics. Even though the model pays no attention to quantum physics, it works surprisingly well for describing frequency-dependent optical index and absorption of light. As it turns out, the Schroedinger equation applied to two levels in an atom reduces in mathematical form to the Lorentz model in the limit of low-intensity light. Quantum mechanics also explains a fudge factor (called the oscillator strength) in the Lorentz model, which before the development of quantum mechanics had to be inserted *ad hoc* to make the model agree with experiments.

We assume (for simplicity) that all atoms (or molecules) in the medium are identical, each with one (or a few) active electrons responding to the external field. The atoms are uniformly distributed throughout space with N identical active electrons per volume (units of number per volume). The polarization of the material is then

$$\mathbf{P} = q_e N \mathbf{r}_{\text{micro}} \tag{2.27}$$

Recall that polarization has units of dipoles per volume. Each dipole has strength $q_e\mathbf{r}_{\text{micro}}$, where $\mathbf{r}_{\text{micro}}$ is a microscopic displacement of the electron from equilibrium. In our modern quantum-mechanical viewpoint, $\mathbf{r}_{\text{micro}}$ corresponds to an average displacement of the electronic cloud, which surrounds the nucleus (see Fig. 2.3). (At the time of Lorentz, atoms were thought to be clouds of positive charge wherein point-like electrons sit at rest unless stimulated by an applied electric field.)

The displacement $\mathbf{r}_{\text{micro}}$ of the electron charge in an individual atom depends on the *local* strength of the applied electric field $\mathbf{E}$. By *local*, we mean the position of the atom. Since the diameter of the electronic cloud is tiny compared to a wavelength of (visible) light, we may consider the electric field to be uniform across any individual atom.

The Lorentz model uses Newton's equation of motion to describe an electron displacement from equilibrium within an atom. In accordance with the classical laws of motion, the electron mass m_e times its acceleration is equal to the sum of the forces on the electron:

$$m_e\ddot{\mathbf{r}}_{\text{micro}} = q_e\mathbf{E} - m_e\gamma\dot{\mathbf{r}}_{\text{micro}} - k_{\text{Hooke}}\mathbf{r}_{\text{micro}} \tag{2.28}$$

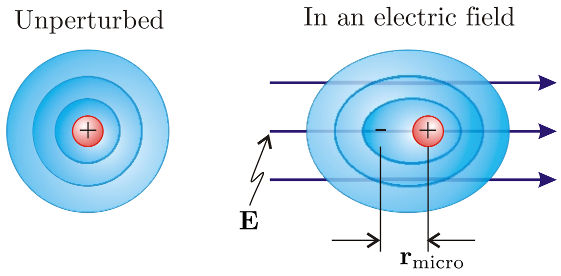


Figure 2.3 A distorted electronic cloud becomes a dipole.

The electric field pulls on the electron with force $q_e\mathbf{E}$.[1] A dragging force $-m_e\gamma\dot{\mathbf{r}}_{\text{micro}}$ opposes the electron motion and accounts for absorption of energy. Without this term, it is only possible to describe optical index at frequencies away from where absorption takes place. Finally, $-k_{\text{Hooke}}\mathbf{r}_{\text{micro}}$ is a force accounting for the fact that the electron is bound to the nucleus. This restoring force can be thought of as an effective spring that pulls the displaced electron back towards equilibrium with a force proportional to the amount of displacement. To a good approximation, this term resembles the familiar Hooke's law.

With some rearranging, (2.28) can be written as

$$\ddot{\mathbf{r}}_{\text{micro}} + \gamma\dot{\mathbf{r}}_{\text{micro}} + \omega_0^2\mathbf{r}_{\text{micro}} = \frac{q_e}{m_e}\mathbf{E} \tag{2.29}$$

where $\omega_0 \equiv \sqrt{k_{\text{Hooke}}/m_e}$ is the natural oscillation frequency (or resonant frequency) associated with the electron mass and the "spring constant."

In accordance with our examination of a single sinusoidal wave, we insert (2.13) into (2.29) and obtain

$$\ddot{\mathbf{r}}_{\text{micro}} + \gamma\dot{\mathbf{r}}_{\text{micro}} + \omega_0^2\mathbf{r}_{\text{micro}} = \frac{q_e}{m_e}\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.30}$$

Note that within a given atom the excursions of $\mathbf{r}_{\text{micro}}$ are so small that $\mathbf{k}\cdot\mathbf{r}$ remains essentially constant, since $\mathbf{k}\cdot\mathbf{r}$ varies with displacements on the scale of an optical wavelength, which is huge compared to the size of an atom. The inhomogeneous solution to (2.30) is (see P 2.1)

$$\mathbf{r}_{\text{micro}} = \left(\frac{q_e}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.31}$$

The electron position $\mathbf{r}_{\text{micro}}$ oscillates (not surprisingly) with the same frequency ω as the driving electric field. This solution illustrates the convenience of the complex notation. The imaginary part in the denominator implies that the electron oscillates with a different phase from the electric field oscillations; the damping term γ (the imaginary part in the denominator) causes the two to be out of phase somewhat. The complex algebra in (2.31) accomplishes what would otherwise be cumbersome and require trigonometric manipulations.

[1] The electron also experiences a force due to the magnetic field of the light, $\mathbf{F} = q_e\mathbf{v}_{\text{micro}} \times \mathbf{B}$, but this force is tiny for typical optical fields.

Hendrik A. Lorentz

(1853–1928, Dutch)

Lorentz extended Maxwell's work in electromagnetic theory and used it to explain the reflection and refraction of light. He developed a simple and useful model for dielectric media and correctly hypothesized that the atoms were composed of charged particles, and that their movement was the source of light. He won the Nobel prize in 1902 for his contributions to electromagnetic theory.

We are now able to write the polarization in terms of the electric field. By substituting (2.31) into (2.27), we obtain

$$\mathbf{P} = \left(\frac{Nq_e^2}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.32}$$

A comparison with (2.15) in view of (2.13) reveals the (complex) susceptibility:

$$\chi(\omega) = \frac{\omega_{\mathrm{p}}^2}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.33}$$

where the *plasma frequency* ω_{p} is

$$\omega_{\mathrm{p}} = \sqrt{\frac{Nq_e^2}{\epsilon_0 m_e}} \tag{2.34}$$

In terms of the susceptibility, the index of refraction according to (2.19) is

$$\mathcal{N}^2 \equiv 1 + \chi \tag{2.35}$$

The real and imaginary parts of the index are solved by equating separately the real and imaginary parts of (2.20), namely

$$(n + i\kappa)^2 = 1 + \chi(\omega) = 1 + \frac{\omega_{\mathrm{p}}^2}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.36}$$

A graph of n and κ is given in Fig. 2.4(a). In actuality, materials usually have more than one species of active electron, and different active electrons behave differently. The generalization of (2.36) in this case is

$$(n + i\kappa)^2 = 1 + \chi(\omega) = 1 + \sum_j \frac{f_j \omega_{\mathrm{p}j}^2}{\omega_{0j}^2 - i\omega\gamma_j - \omega^2} \tag{2.37}$$

where f_j is the aptly named *oscillator strength* for the j^{th} species of active electron.

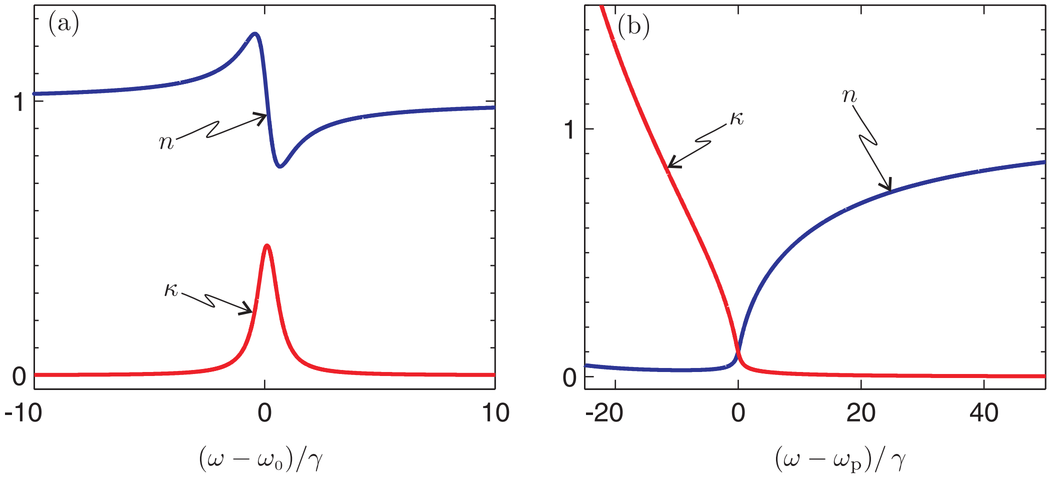


Figure 2.4 (a) Real and imaginary parts of the index for a single Lorentz oscillator dielectric with $\omega_p = 10\gamma$. (b) Real and imaginary parts of the index for conductor with $\omega_p = 50\gamma$.

2.5 Conductor Model of Refractive Index and Absorption

The details of the conductor model are very similar to those of the dielectric model in the previous section. We will go through the derivation quickly since the procedure so closely parallels the previous section. In this model, we will ignore polarization (i.e. $\mathbf{P} = 0$), but take the current density $\mathbf{J}_{\text{free}}$ to be non-zero. The wave equation then becomes

$$\nabla^2 \mathbf{E} - \epsilon_0 \mu_0 \frac{\partial^2}{\partial t^2} \mathbf{E} = \mu_0 \frac{\partial}{\partial t} \mathbf{J}_{\text{free}} \tag{2.38}$$

In a manner similar to (2.13), we assume sinusoidal solutions:

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \\ \mathbf{J}_{\text{free}} &= \mathbf{J}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \end{aligned} \tag{2.39}$$

In a manner similar to (2.13), we assume that the current is made up of individual electrons traveling with velocity $\mathbf{v}_{\text{micro}}$:

$$\mathbf{J}_{\text{free}} = q_e N \mathbf{v}_{\text{micro}} \tag{2.40}$$

Again, N is the number density of free electrons (in units of number per volume). Recall that current density $\mathbf{J}_{\text{free}}$ has units of charge times velocity per volume (or current per cross sectional area), so (2.40) may be thought of as a definition of current density in a fundamental sense.

As before, we use Newton's equation of motion on a representative electron. Mass times acceleration equals the sum of the forces on the electron:

$$m_e \dot{\mathbf{v}}_{\text{micro}} = q_e \mathbf{E} - m_e \gamma \mathbf{v}_{\text{micro}} \tag{2.41}$$

The electric field pulls on the electron with force $q_e \mathbf{E}$. A dragging force $-m_e \gamma \mathbf{v}_{\text{micro}}$ opposes the motion in proportion to the speed (identical to the dielectric model, see (2.28)).

Physically, the dragging term arises due to collisions between electrons and lattice sites in a metal. Such collisions give rise to resistance in a conductor.

When a DC field is applied, the electrons initially accelerate, but soon reach a terminal velocity as the drag force kicks in. In the steady state, we may thus take the acceleration to be zero where the other two forces balance (i.e. $\dot{\mathbf{v}} = 0$). Then by combining (2.40) and (2.41) we get Ohm's law $\mathbf{J} = \sigma\mathbf{E}$, where $\sigma = Nq_\mathrm{e}^2/m_\mathrm{e}\gamma$ is the conductivity. Although our model relates the dragging term γ to the DC conductivity σ, the connection matches poorly with experimental observations made for visible frequencies. This is because the collision rate actually varies somewhat with frequency. Nevertheless, the qualitative behavior of the model is useful.

Upon substitution of (2.39) into (2.41) we get

$$\dot{\mathbf{v}}_\mathrm{micro} + \gamma\mathbf{v}_\mathrm{micro} = \frac{q_e}{m_e}\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.42}$$

The solution to this equation is (see P 2.5)

$$\mathbf{v}_\mathrm{micro} = \frac{q_e}{m_e}\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\gamma - i\omega} \tag{2.43}$$

We are now able to find an expression for the current density (2.40) in terms of the electric field:

$$\mathbf{J}_\mathrm{free} = \left(\frac{Nq_e^2}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\gamma - i\omega} \tag{2.44}$$

We substitute this expression together with (2.39) back into (2.38) and obtain

$$-k^2\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \frac{\omega^2}{c^2}\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = -i\omega\left(\frac{\mu_0 Nq_e^2}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\gamma - i\omega} \tag{2.45}$$

The solutions (2.39) then require the following relation to hold:

$$k^2 = \frac{\omega^2}{c^2} - \left(\frac{\mu_0 Nq_e^2}{m_e}\right)\frac{\omega}{i\gamma + \omega} \tag{2.46}$$

Using (2.21) with (2.46), we find that the complex index of refraction for the conductor model is given by

$$(n + i\kappa)^2 = 1 - \frac{\omega_\mathrm{p}^2}{i\gamma\omega + \omega^2} \tag{2.47}$$

A graph of n and κ in the conductor model is given in Fig. 2.4(b).

Here we have introduced a complex refractive index for the conductor model just as we did for the dielectric model. Equations (2.22) through (2.26) also apply to the conductor model. The similarity is not surprising since both models include oscillating electrons. In the one case the electrons are free, and in the other case they are tethered to their atoms. In either model, the damping term removes energy from the electron oscillations. In the complex notation for the field, the damping term gives rise to an imaginary part of the index. Again, the imaginary part of the index causes an exponential attenuation of the plane wave as it propagates.

2.6 Poynting's Theorem

We next turn our attention to the detection and measurement of light. Until now, we have described light as the propagation of an electromagnetic disturbance. However, we typically observe light by detecting absorbed energy rather than the field amplitude directly. In this section we examine the connection between propagating electromagnetic fields (such as the plane waves discussed above) and the energy transported by such fields.

John Henry Poynting (1852-1914) developed (from Maxwell's equations) the theoretical foundation that describes light energy transport. In this section we examine its development, which is surprisingly concise. Students should concentrate mainly on the ideas involved (rather than the details of the derivation), especially the definition and meaning of the Poynting vector, describing energy flow in an electromagnetic field.

Poynting's theorem derives from just two of Maxwell's Equations: (1.37) and (1.38). We take the dot product of $\mathbf{B}/\mu_0$ with the first equation and the dot product of $\mathbf{E}$ with the second equation. Then by subtracting the second equation from the first we obtain

$$\frac{\mathbf{B}}{\mu_0}\cdot(\nabla\times\mathbf{E})-\mathbf{E}\cdot\left(\nabla\times\frac{\mathbf{B}}{\mu_0}\right)+\epsilon_0\mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t}+\frac{\mathbf{B}}{\mu_0}\cdot\frac{\partial\mathbf{B}}{\partial t}=-\mathbf{E}\cdot\left(\mathbf{J}_{\text{free}}+\frac{\partial\mathbf{P}}{\partial t}\right)\tag{2.48}$$

The first two terms can be simplified using the vector identity P 0.18. The next two terms are the time derivatives of $\epsilon_0E^2/2$ and $B^2/2\mu_0$, respectively. The relation (2.48) then becomes

$$\nabla\cdot\left(\mathbf{E}\times\frac{\mathbf{B}}{\mu_0}\right)+\frac{\partial}{\partial t}\left(\frac{\epsilon_0E^2}{2}+\frac{B^2}{2\mu_0}\right)=-\mathbf{E}\cdot\left(\mathbf{J}_{\text{free}}+\frac{\partial\mathbf{P}}{\partial t}\right)\tag{2.49}$$

This is Poynting's theorem. Each term in this equation has units of power per volume.

The conventional way of writing Poynting's theorem is as follows:

$$\nabla\cdot\mathbf{S}+\frac{\partial u_{\text{field}}}{\partial t}=-\frac{\partial u_{\text{medium}}}{\partial t}\tag{2.50}$$

where

$$\mathbf{S}\equiv\mathbf{E}\times\frac{\mathbf{B}}{\mu_0}\tag{2.51}$$

$$u_{\text{field}}\equiv\frac{\epsilon_0E^2}{2}+\frac{B^2}{2\mu_0},\tag{2.52}$$

and

$$\frac{\partial u_{\text{medium}}}{\partial t}\equiv\mathbf{E}\cdot\left(\mathbf{J}_{\text{free}}+\frac{\partial\mathbf{P}}{\partial t}\right).\tag{2.53}$$

$\mathbf{S}$ is called the Poynting vector and has units of power per area, called irradiance. The quantity u_{field} is the energy per volume stored in the electric and magnetic fields. Derivations of the electric field energy density and the magnetic field energy density are given in Appendices 2.A and 2.B. (See (2.68) and (2.75).) The term $\partial u_{\text{medium}}/\partial t$ is the power per volume delivered to the medium. Equation (2.53) is reminiscent of the familiar circuit power law, *Power* = *Voltage* × *Current*. Power is delivered when a charged particle traverses a distance while experiencing a force. This happens when currents flow in the presence of electric fields. Recall that $\partial\mathbf{P}/\partial t$ is a current density similar to $\mathbf{J}_{\text{free}}$, with units of charge times velocity per volume.

The interpretation of the Poynting vector is straightforward when we recognize Poynting's theorem as a statement of the conservation of energy. **S** describes the flow of energy. To see this more clearly, consider Poynting's theorem (2.50) integrated over a volume V (enclosed by surface S). If we also apply the divergence theorem (0.26) to the term involving $\nabla \cdot \mathbf{S}$ we obtain

$$\oint_S \mathbf{S} \cdot \hat{\mathbf{n}}\, da = -\frac{\partial}{\partial t} \int_V (u_{\text{field}} + u_{\text{medium}})\, dv \tag{2.54}$$

Notice that the volume integral over energy densities u_{field} and u_{medium} gives the total energy stored in V, whether in the form of electromagnetic field energy density or as energy density that has been given to the medium. The integration of the Poynting vector over the surface gives the net Poynting vector flux directed outward. Equation (2.54) indicates that the outward Poynting vector flux matches the rate that total energy disappears from the interior of V. Conversely, if the Poynting vector is directed inward (negative), then the net inward flux matches the rate that energy increases within V. *The vector* **S** *defines the flow of energy through space.* Its units of *power per area* are just what are needed to describe the brightness of light impinging on a surface.

2.7 Irradiance of a Plane Wave

Consider the electric field wave described by (2.9). The magnetic field that accompanies this electric field can be found from Maxwell's equation (1.37), and it turns out to be

$$\mathbf{B}(\mathbf{r}, t) = \frac{\mathbf{k} \times \mathbf{E}_0}{\omega} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \tag{2.55}$$

When **k** is complex, **B** is out of phase with **E**, and this occurs when absorption takes place. When there is no absorption, then **k** is real, and **B** and **E** carry the same complex phase.

Before computing the Poynting vector (2.51), which involves multiplication, we must remember our unspoken agreement that only the real parts of the fields are relevant. We necessarily remove the imaginary parts before multiplying (see (0.10)). We could rewrite **B** and **E** like in (2.22), imposing the assumption that the complex phase for each vector component of $\mathbf{E}_0$ is the same. However, we can defer making this assumption by taking the real parts of the field in the following manner: Obtain the real parts of the fields by adding their respective complex conjugates and dividing the result by 2 (see (0.17)). The real field associated with (2.9) is

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{2}\left[\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} + \mathbf{E}_0^* e^{-i(\mathbf{k}^*\cdot\mathbf{r} - \omega t)}\right] \tag{2.56}$$

and the real field associated with (2.55) is

$$\mathbf{B}(\mathbf{r}, t) = \frac{1}{2}\left[\frac{\mathbf{k} \times \mathbf{E}_0}{\omega} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} + \frac{\mathbf{k}^* \times \mathbf{E}_0^*}{\omega} e^{-i(\mathbf{k}^*\cdot\mathbf{r} - \omega t)}\right] \tag{2.57}$$

By writing (2.56) and (2.57), we have merely exercised our previous agreement that only the real parts of (2.36) and (2.55) are to be retained.

The Poynting vector (2.51) associated with the plane wave is then computed as follows:

$$\begin{aligned}
\mathbf{S} &\equiv \mathbf{E} \times \frac{\mathbf{B}}{\mu_0} \\
&= \frac{1}{2}\left[\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \mathbf{E}_0^* e^{-i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\right] \times \frac{1}{2\mu_0}\left[\frac{\mathbf{k}\times\mathbf{E}_0}{\omega} e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \frac{\mathbf{k}^*\times\mathbf{E}_0^*}{\omega} e^{-i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\right] \\
&= \frac{1}{4\mu_0}\left[\begin{array}{l} \frac{\mathbf{E}_0\times(\mathbf{k}\times\mathbf{E}_0)}{\omega} e^{2i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \frac{\mathbf{E}_0^*\times(\mathbf{k}\times\mathbf{E}_0)}{\omega} e^{i(\mathbf{k}-\mathbf{k}^*)\cdot\mathbf{r}} \\ + \frac{\mathbf{E}_0\times\left(\mathbf{k}^*\times\mathbf{E}_0^*\right)}{\omega} e^{i(\mathbf{k}-\mathbf{k}^*)\cdot\mathbf{r}} + \frac{\mathbf{E}_0^*\times\left(\mathbf{k}^*\times\mathbf{E}_0^*\right)}{\omega} e^{-2i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)} \end{array}\right] \\
&= \frac{1}{4\mu_0}\left[\frac{k}{\omega}\mathbf{E}_0 \times (\hat{\mathbf{u}}\times\mathbf{E}_0)\, e^{2i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \frac{k}{\omega}\mathbf{E}_0^* \times (\hat{\mathbf{u}}\times\mathbf{E}_0)\, e^{-2\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} + \text{C.C.}\right]
\end{aligned} \tag{2.58}$$

The letters "C.C." stand for the complex conjugate of what precedes. The direction of $\mathbf{k}$ is specified with the real unit vector $\hat{\mathbf{u}}$. We have also used (2.21) to rewrite $i\,(\mathbf{k}-\mathbf{k}^*)$ as $-2\,(\kappa\omega/c)\,\hat{\mathbf{u}}$.

In an isotropic medium (not a crystal) we have from Maxwell's equations the requirement $\nabla\cdot\mathbf{E}\,(\mathbf{r},t)=0$ (see (1.35)), or in other words $\hat{\mathbf{u}}\cdot\mathbf{E}_0=0$. We can use this fact together with the BAC-CAB rule P 0.12 to replace the above expression with

$$\mathbf{S} = \frac{\hat{\mathbf{u}}}{4\mu_0}\left[\frac{k}{\omega}\left(\mathbf{E}_0\cdot\mathbf{E}_0\right) e^{2i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \frac{k}{\omega}\left(\mathbf{E}_0\cdot\mathbf{E}_0^*\right) e^{-2\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} + \text{C.C.}\right] \tag{2.59}$$

This expression shows that in an isotropic medium the flow of energy is in the direction of $\hat{\mathbf{u}}$ (or $\mathbf{k}$). This agrees with our intuition that energy flows in the direction that the wave propagates.

Very often, we are interested in the time-average of the Poynting vector, denoted by $\langle\mathbf{S}\rangle_t$. Under the time averaging, the first term in (2.59) vanishes since it oscillates positive and negative by the same amount. Note that k is the only factor in the second term that is (potentially) not real. The time-averaged Poynting vector becomes

$$\begin{aligned}
\langle\mathbf{S}\rangle_t &= \frac{\hat{\mathbf{u}}}{4\mu_0}\frac{k+k^*}{\omega}\left(\mathbf{E}_0\cdot\mathbf{E}_0^*\right) e^{-2\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} \\
&= \hat{\mathbf{u}}\frac{n\epsilon_0 c}{2}\left(\left|E_{0x}\right|^2 + \left|E_{0y}\right|^2 + \left|E_{0z}\right|^2\right) e^{-2\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}}
\end{aligned} \tag{2.60}$$

We have used (2.21) to rewrite $k+k^*$ as $2\,(n\omega/c)$. We have also used (1.44) to rewrite $1/\mu_0 c$ as $\epsilon_0 c$.

The expression (2.60) is called irradiance (with the direction $\hat{\mathbf{u}}$ included). However, we often speak of the intensity of a field I, which amounts to the same thing, but without regard for the direction $\hat{\mathbf{u}}$. The definition of intensity is thus less specific, and it can be applied, for example, to standing waves where the net irradiance is technically zero (i.e. counter-propagating plane waves with zero net energy flow). Nevertheless, atoms in standing waves "feel" the oscillating field. In general, the intensity is written as

$$I = \frac{n\epsilon_0 c}{2}\mathbf{E}_0\cdot\mathbf{E}_0^* = \frac{n\epsilon_0 c}{2}\left(\left|E_{0x}\right|^2 + \left|E_{0y}\right|^2 + \left|E_{0z}\right|^2\right) \tag{2.61}$$

where in this case we have ignored absorption (i.e. $\kappa \cong 0$), or, alternatively, we could have considered $\left|E_{0x}\right|^2$, $\left|E_{0y}\right|^2$, and $\left|E_{0z}\right|^2$ to possess the factor $\exp\left\{-2\,(\kappa\omega/c)\,\hat{\mathbf{u}}\cdot\mathbf{r}\right\}$ already.

Appendix 2.A Energy Density of Electric Fields

In this appendix and the next, we prove that the term $\epsilon_0 E^2/2$ in (2.52) corresponds to the energy density of an electric field. The electric potential $\phi(\mathbf{r})$ (in units of energy per charge, or in other words volts) describes each point of an electric field in terms of the potential energy that a charge would experience if placed in that field. The electric field and the potential are connected through

$$\mathbf{E}(\mathbf{r}) = -\nabla\phi(\mathbf{r}) \tag{2.62}$$

The energy U necessary to assemble a distribution of charges (owing to attraction or repulsion) can be written in terms of a summation over all of the charges (or charge density $\rho(\mathbf{r})$) located within the potential:

$$U = \frac{1}{2}\int\limits_V \phi(\mathbf{r})\,\rho(\mathbf{r})\,dv \tag{2.63}$$

The factor 1/2 is necessary to avoid double counting. To appreciate this factor consider two charges: We need only count the energy due to one charge in the presence of the other's potential to obtain the energy required to bring the charges together.

A substitution of (1.8) for $\rho(\mathbf{r})$ into (2.63) gives

$$U = \frac{\epsilon_0}{2}\int\limits_V \phi(\mathbf{r})\,\nabla\cdot\mathbf{E}(\mathbf{r})\,dv \tag{2.64}$$

Next, we use the vector identity in P 0.19 and get

$$U = \frac{\epsilon_0}{2}\int\limits_V \nabla\cdot[\phi(\mathbf{r})\,\mathbf{E}(\mathbf{r})]\,dv - \frac{\epsilon_0}{2}\int\limits_V \mathbf{E}(\mathbf{r})\cdot\nabla\phi(\mathbf{r})\,dv \tag{2.65}$$

An application of the Divergence theorem (0.26) on the first integral and a substitution of (2.62) into the second integral yields

$$U = \frac{\epsilon_0}{2}\oint\limits_S \phi(\mathbf{r})\,\mathbf{E}(\mathbf{r})\cdot\hat{\mathbf{n}}da + \frac{\epsilon_0}{2}\int\limits_V \mathbf{E}(\mathbf{r})\cdot\mathbf{E}(\mathbf{r})\,dv \tag{2.66}$$

Finally, we consider the volume V (enclosed by S) to be extremely large so that all charges are contained well within it. If we choose a large enough volume, say a sphere of radius R, the surface integral over S vanishes. The integrand of the surface integral becomes negligibly small $\phi \sim 1/R$ and $E \sim 1/R^2$, whereas $da \sim R^2$. Therefore, the energy associated with an electric field in a region of space is

$$U = \int\limits_V u_E(\mathbf{r})\,dv \tag{2.67}$$

where

$$u_E(\mathbf{r}) \equiv \frac{\epsilon_0 E^2}{2} \tag{2.68}$$

is interpreted as the energy density of the electric field.

Appendix 2.B Energy Density of Magnetic Fields

In a derivation similar to that in appendix 2.A, we consider the energy associated with magnetic fields. The magnetic vector potential $\mathbf{A}(\mathbf{r})$ (in units of energy per charge×velocity) describes the potential energy that a charge moving with velocity $\mathbf{v}$ would experience if placed in the field. The magnetic field and the vector potential are connected through

$$\mathbf{B}(\mathbf{r}) = \nabla \times \mathbf{A}(\mathbf{r}) \tag{2.69}$$

The energy U necessary to assemble a distribution of current can be written in terms of a summation over all of the currents (or current density $\mathbf{J}(\mathbf{r})$) located within the vector potential field:

$$U = \frac{1}{2}\int\limits_V \mathbf{J}(\mathbf{r}) \cdot \mathbf{A}(\mathbf{r})\, dv \tag{2.70}$$

As in (2.63), the factor 1/2 is necessary to avoid double counting the influence of the currents on each other.

Under the assumption of steady currents (no variations in time), we may substitute Ampere's law (1.14) into (2.70), which yields

$$U = \frac{1}{2\mu_0}\int\limits_V [\nabla \times \mathbf{B}(\mathbf{r})] \cdot \mathbf{A}(\mathbf{r})\, dv \tag{2.71}$$

Next we employ the vector identity P 0.18 from which the previous expression becomes

$$U = \frac{1}{2\mu_0}\int\limits_V \mathbf{B}(\mathbf{r}) \cdot [\nabla \times \mathbf{A}(\mathbf{r})]\, dv - \frac{1}{2\mu_0}\int\limits_V \nabla \cdot [\mathbf{A}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})]\, dv \tag{2.72}$$

Upon substituting (2.69) into the first equation and applying the Divergence theorem (0.26) on the second integral, this expression for total energy becomes

$$U = \frac{1}{2\mu_0}\int\limits_V \mathbf{B}(\mathbf{r}) \cdot \mathbf{B}(\mathbf{r})\, dv - \frac{1}{2\mu_0}\oint\limits_S [\mathbf{A}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})] \cdot \hat{\mathbf{n}}\, da \tag{2.73}$$

As was done in connection with (2.66), if we choose a large enough volume (a sphere with radius R), the surface integral vanishes because $A \sim 1/R$ and $B \sim 1/R^2$, whereas $da \sim R^2$. The total energy (2.73) then reduces to

$$U = \int\limits_V u_B(\mathbf{r})\, dv \tag{2.74}$$

where

$$u_B(\mathbf{r}) \equiv \frac{B^2}{2\mu_0} \tag{2.75}$$

is the energy density for a magnetic field.

Name	**Concept**	**Units**
Radiant Power (of a source)	Electromagnetic energy emitted per time from a source	$\mathrm{W} = \mathrm{J/s}$
Radiant Solid-Angle Intensity (of a source)	Radiant power per steradian emitted from a point-like source (4π steradians in a sphere)	W/Sr
Radiance or Brightness (of a source)	Radiant solid-angle intensity per unit projected area of an extended source. The *projected* area foreshortens by $\cos\theta$, where θ is the observation angle relative to the surface normal.	$\mathrm{W/(Sr \cdot cm^2)}$
Radiant Emittance or Exitance (from a source)	Radiant Power emitted per unit surface area of an extended source (the Poynting flux leaving).	$\mathrm{W/cm^2}$
Irradiance (to a receiver). Often called intensity	Electromagnetic power delivered per area to a receiver: Poynting flux arriving.	$\mathrm{W/cm^2}$

Table 2.2 Radiometric quantities and units.

Appendix 2.C Radiometry Versus Photometry

Photometry refers to the characterization of light sources in the context of the spectral response of the human eye. However, physicists most often deal with *radiometry*, which treats light of any wavelength on equal footing. Table 2.2 lists several concepts important in radiometry. The last two entries are associated with the average Poynting flux described in section 2.7.

The concepts used in photometry are similar, except that the radiometric quantities are multiplied by the spectral response of the human eye, a curve that peaks at $\lambda_{\text{vac}} = 555$ nm and drops to near zero for wavelengths longer than $\lambda_{\text{vac}} = 700$ nm or shorter than $\lambda_{\text{vac}} = 400$ nm. Photometric units, which may seem a little obscure, were first defined in terms of an actual candle with prescribed dimensions made from whale tallow. The basic unit of luminous power is called the lumen, defined to be (1/683) W of light with wavelength $\lambda_{\text{vac}} = 555$ nm, the peak of the eye's response. More radiant power is required to achieve the same number of lumens for wavelengths away from the center of the eye's spectral response. Photometric units are often used to characterize room lighting as well as photographic, projection, and display equipment. Table 2.3 gives the names of the various photometric quantities, which parallel the entries in table 2.2. We include a variety of units that are sometimes encountered.

Name	**Concept**	**Typical Units**
Luminous Power (of a source)	Visible light energy emitted per time from a source: **lumen** (lm).	lm=(1/683) W @ 555 nm
Luminous Solid-Angle Intensity (of a source)	Luminous power per steradian emitted from a point-like source: **candela** (cd).	cd = lm/Sr
Luminance (of a source)	Luminous solid-angle intensity per projected area of an extended source. (The *projected* area foreshortens by $\cos\theta$, where θ is the observation angle relative to the surface normal.)	cd/cm^2 = stilb cd/m^2 = nit nit = 3183 lamberts = 3.4 footlamberts
Luminous Emittance or Exitance (from a source)	Luminous Power emitted per unit surface area of an extended source	lm/cm^2
Illuminance (to a receiver)	Incident luminous power delivered per area to a receiver: **lux**.	lm/m^2 = lux lm/cm^2 = phot lm/ft^2 = footcandle

Table 2.3 Photometric quantities and units.

Exercises

2.3 Index of Refraction in Dielectrics

P2.1 Verify that (2.31) is a solution to (2.30).

P2.2 Derive the Sellmeier equation

$$n^2 = 1 + \frac{A\lambda_{\text{vac}}^2}{\lambda_{\text{vac}}^2 - \lambda_{0,\text{vac}}^2}$$

from (2.36) for a gas with negligible absorption (i.e. $\gamma \cong 0$, valid far from resonance ω_0), where $\lambda_{0,\text{vac}}$ corresponds to frequency ω_0 and A is a constant. Many materials (e.g. glass, air) have strong resonances in the ultraviolet. In such materials, do you expect the index of refraction for blue light to be greater than that for red light? Make a sketch of n as a function of wavelength for visible light down to the ultraviolet (where $\lambda_{0,\text{vac}}$ is located).

P2.3 In the Lorentz model, take $N = 10^{28}\ \text{m}^{-3}$ for the density of bound electrons in an insulator (note that N is number per volume, not just number), and a single transition at $\omega_0 = 6 \times 10^{15}$ rad/sec (in the UV), and damping $\gamma = \omega_0/5$ (quite broad). Assume E_0 is 10^4 V/m.

For three frequencies $\omega = \omega_0 - 2\gamma$, $\omega = \omega_0$, and $\omega = \omega_0 + 2\gamma$ find the *magnitude* and *phase* of the following (give the phase relative to the phase of E_0). Give correct SI units with each quantity. You don't need to worry about vector directions.

(a) The charge displacement amplitude r_{micro} (2.31)

(b) The polarization amplitude $P(\omega)$

(c) The susceptibility $\chi(\omega)$. What would the susceptibility be for twice the E-field strength as before?

For the following no phase is needed:

(d) Find n and κ at the three frequencies. You will have to solve for the real and imaginary parts of $(n + i\kappa)^2 = 1 + \chi(\omega)$.

(e) Find the three speeds of light in terms of c. Find the three wavelengths λ.

(f) Find how far light penetrates into the material before only $1/e$ of the amplitude of E remains. Find how far light penetrates into the material before only $1/e$ of the intensity I remains.

P2.4 (a) Use a computer graphing program and the Lorentz model to plot n and κ as a function of ω frequency for a dielectric (i.e. obtain graphs such as the ones in Fig. 2.4(a)). Use these parameters to keep things simple: $\omega_{\text{p}} = 1$, $\omega_0 = 10$, and $\gamma = 1$; plot your function from $\omega = 0$ to $\omega = 20$.

(b) Plot n and κ as a function of frequency for a material that has three resonant frequencies: $\omega_{01} = 10$, $\gamma_1 = 1$, $f_1 = 0.5$; $\omega_{02} = 15$, $\gamma_2 = 1$, $f_2 = 0.25$; and $\omega_{03} = 25$, $\gamma_3 = 3$, $f_3 = 0.25$. Use $\omega_{\text{p}} = 1$ for all three resonances, and plot the results from $\omega = 0$ to $\omega = 30$. Comment on your plots.

2.5 Conductor Model of Refractive Index and Absorption

P2.5 Verify that (2.43) is a solution to (2.42).

P2.6 For silver, the complex refractive index is characterized by $n = 0.2$ and $\kappa = 3.4$. Find the distance that light travels inside of silver before the field is reduced by a factor of 1/e. Assume a wavelength of $\lambda_{\text{vac}} = 633$ nm. What is the speed of the wave crests in the silver (written as a number times c)? Are you surprised?

P2.7 Show that the dielectric model and the conductor model give identical results for n in the case of a low-density plasma where there is no restoring force (i.e. $\omega_0 = 0$) and no dragging term (i.e., $\gamma = 0$). Write n in terms of the plasma frequency ω_p.

P2.8 Use the result from P 2.7.

(a) If the index of refraction of the ionosphere is $n = 0.9$ for an FM station at $\nu = \omega/2\pi = 100$ MHz, calculate the number of free electrons per cubic meter.

(b) What is the complex refractive index for KSL radio at 1160 kHz? Assume the same density of free electrons as in part (a). For your information, AM radio reflects better than FM radio from the ionosphere (like visible light from a metal mirror). At night, the lower layer of the ionosphere goes away so that AM radio waves reflect from a higher layer.

P2.9 Use a computer graphing program to plot n and κ as a function of frequency for a conductor (obtain plots such as the ones in Fig. 2.4(b)). Use these parameters to keep things simple: $\omega_\text{p} = 1$ and $\gamma = 0.02$. Plot your function from $\omega = 0.6$ to $\omega = 2$.

2.7 Irradiance of a Plane Wave

P2.10 In the case of a linearly-polarized plane wave, where the phase of each vector component of $\mathbf{E}_0$ is the same, re-derive (2.60) directly from the real field (2.23). For simplicity, you may ignore absorption (i.e. $\kappa \cong 0$).

HINT: The time-average of $\cos^2(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi)$ is 1/2.

P2.11 (a) Find the intensity (in W/cm^2) produced by a short laser pulse (linearly polarized) with duration $\Delta t = 2.5 \times 10^{-14}$ s and energy $E = 100$ mJ, focused in vacuum to a round spot with radius $r = 5$ μm.

(b) What is the peak electric field (in V/Å)?

HINT: The SI units of electric field are N/C = V/m.

(c) What is the peak magnetic field (in T = kg/(s · C)?

P2.12 What is the intensity (in W/cm^2) *on the retina* when looking directly at the sun? Assume that the eye's pupil has a radius $r_{\text{pupil}} = 1$ mm. Take the Sun's irradiance at the earth's surface to be 1.4 kW/m^2, and neglect refractive index (i.e. set $n = 1$). HINT: The Earth-Sun distance is $d_\text{o} = 1.5 \times 10^8$ km and the pupil-retina distance is $d_\text{i} = 22$ mm. The radius of the Sun $r_\text{Sun} = 7.0 \times 10^5$ km is de-magnified on the retina according to the ratio d_i/d_o.

P2.13 What is the intensity at the retina when looking directly into a 1 mW HeNe laser? Assume that the smallest radius of the laser beam is $r_\text{waist} = 0.5$ mm positioned $d_\text{o} = 2$ m in front of the eye, and that the entire beam enters the pupil. Compare with P 2.12 (see HINT).

P2.14 Show that the magnetic field of an intense laser with $\lambda = 1\ \mu$m becomes important for a free electron oscillating in the field at intensities above 10^{18} W/cm^2. This marks the transition to relativistic physics. Nevertheless, for convenience, use classical physics in making the estimate.

HINT: At lower intensities, the oscillating electric field dominates, so the electron motion can be thought of as arising solely from the electric field. Use this motion to calculate the magnetic force on the moving electron, and compare it to the electric force. The forces become comparable at 10^{18} W/cm^2.

Chapter 3

Reflection and Refraction

3.1 Introduction

In the previous chapter, we considered a plane wave propagating in a homogeneous isotropic medium. In this chapter, we examine what happens when such a wave propagates from one material (characterized by index n or even by complex index $\mathcal{N}$) to another material. As we know from everyday experience, when light arrives at an interface between materials it is partially reflected and partially transmitted. We will derive expressions for the amount of reflection and transmission. The results depend on the angle of incidence (i.e. the angle between $\mathbf{k}$ and the normal to the surface) as well as on the orientation of the electric field (called polarization—not to be confused with $\mathbf{P}$, also called polarization).

As we develop the connection between incident, reflected, and transmitted light waves, many familiar relationships will emerge naturally (e.g. Snell's law, Brewster's angle). The formalism also describes polarization-dependent phase shifts upon reflection (especially interesting in the case of total internal reflection or in the case of reflections from absorbing surfaces such as metals), described in sections 3.6 and 3.7.

For simplicity, we initially neglect the imaginary part of refractive index. Each plane wave is thus characterized by a real wave vector $\mathbf{k}$. We will write each plane wave in the form $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \exp\left[i\left(\mathbf{k} \cdot \mathbf{r} - \omega t\right)\right]$, where, as usual, only the real part of the field corresponds to the physical field. The restriction to real indices is not as serious as it might seem since the results can be extended to include complex indices, and we do this in section 3.7. The use of the letter n instead of $\mathcal{N}$ hardly matters. The math is all the same, which demonstrates the power of the complex notation.

In an isotropic medium, the electric field amplitude $\mathbf{E}_0$ is confined to a plane perpendicular to $\mathbf{k}$. Therefore, $\mathbf{E}_0$ can always be broken into two orthogonal polarization components within that plane. The two vector components of $\mathbf{E}_0$ contain the individual phase information for each dimension. If the phases of the two components of $\mathbf{E}_0$ are the same, then the polarization of the electric field is said to be linear. If the components of the vector $\mathbf{E}_0$ differ in phase, then the electric field polarization is said to be elliptical (or circular) as will be studied in chapter 4.

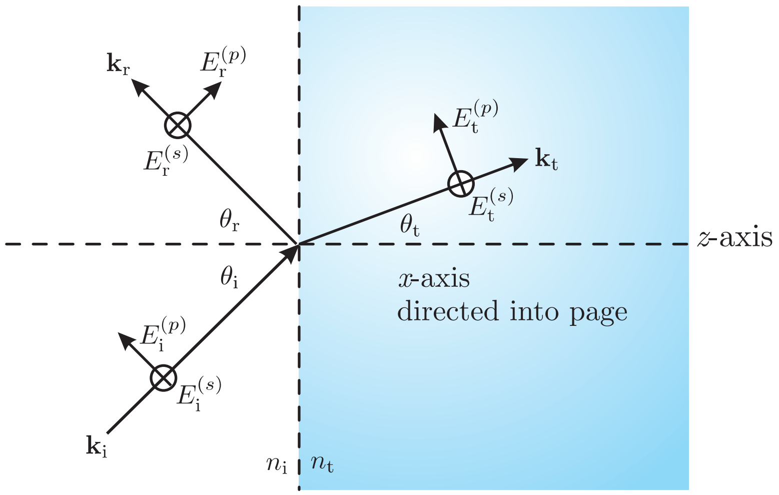


Figure 3.1 Incident, reflected, and transmitted plane wave fields at a material interface.

3.2 Refraction at an Interface

To study the reflection and transmission of light at a material interface, we will examine three distinct waves traveling in the directions $\mathbf{k}_\mathrm{i}$, $\mathbf{k}_\mathrm{r}$, and $\mathbf{k}_\mathrm{t}$ as depicted in the Fig. 3.1. In the upcoming development, we will refer to Fig. 3.1 often. We assume a planar boundary between the two materials. The index n_i characterizes the material on the left, and the index n_t characterizes the material on the right. $\mathbf{k}_\mathrm{i}$ specifies an incident plane wave making an angle θ_i with the normal to the interface. $\mathbf{k}_\mathrm{r}$ specifies a reflected plane wave making an angle θ_r with the interface normal. These two waves exist only to the left of the interface. $\mathbf{k}_\mathrm{t}$ specifies a transmitted plane wave making an angle θ_t with the interface normal. The transmitted wave exists only to the right of the material interface.

We choose the y–z plane to be the *plane of incidence*, containing $\mathbf{k}_\mathrm{i}$, $\mathbf{k}_\mathrm{r}$, and $\mathbf{k}_\mathrm{t}$ (i.e. the plane represented by the surface of this page). By symmetry, all three k-vectors must lie in a single plane, assuming an isotropic material. We are free to orient our coordinate system in many different ways (and every textbook seems to do it differently!). We choose the normal incidence on the interface to be along the z-direction. The x-axis points into the page.

For a given $\mathbf{k}_\mathrm{i}$, the electric field vector $\mathbf{E}_\mathrm{i}$ can be decomposed into arbitrary components as long as they are perpendicular to $\mathbf{k}_\mathrm{i}$. For convenience, we choose one of the electric field vector components to be that which lies within the plane of incidence as depicted in Fig. 3.1. $E_\mathrm{i}^{(p)}$ denotes this component, represented by an arrow in the plane of the page. The remaining electric field vector component, denoted by $E_\mathrm{i}^{(s)}$, is directed normal to the plane of incidence. The superscript s stands for *senkrecht*, a German word meaning perpendicular. In Fig. 3.1, $E_{\rm i}^{(s)}$ is represented by the tail of an arrow pointing into the page, or the x-direction, by our convention. The other fields $\mathbf{E}_{\rm r}$ and $\mathbf{E}_{\rm t}$ are similarly split into s and p components as indicated in Fig. 3.1. (Our choice of coordinate system orientation is motivated in part by the fact that it is easier to draw arrow tails rather than arrow tips to represent the electric field in the s-direction.) All field components are considered to be positive when they point in the direction of their respective arrows.[1]

By inspection of Fig. 3.1, we can write the various k-vectors in terms of the $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ unit vectors:

$$\begin{aligned}\mathbf{k}_{\rm i} &= k_{\rm i}\left(\hat{\mathbf{y}}\sin\theta_{\rm i} + \hat{\mathbf{z}}\cos\theta_{\rm i}\right)\\ \mathbf{k}_{\rm r} &= k_{\rm r}\left(\hat{\mathbf{y}}\sin\theta_{\rm r} - \hat{\mathbf{z}}\cos\theta_{\rm r}\right)\\ \mathbf{k}_{\rm t} &= k_{\rm t}\left(\hat{\mathbf{y}}\sin\theta_{\rm t} + \hat{\mathbf{z}}\cos\theta_{\rm t}\right)\end{aligned} \tag{3.1}$$

Also by inspection of Fig. 3.1 (following the conventions for the electric fields depicted by the arrows), we can write the incident, reflected, and transmitted fields in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$:

$$\begin{aligned}\mathbf{E}_{\rm i} &= \left[E_{\rm i}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_{\rm i} - \hat{\mathbf{z}}\sin\theta_{\rm i}\right) + \hat{\mathbf{x}}E_{\rm i}^{(s)}\right] e^{i[k_{\rm i}(y\sin\theta_{\rm i}+z\cos\theta_{\rm i})-\omega_{\rm i}t]}\\ \mathbf{E}_{\rm r} &= \left[E_{\rm r}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_{\rm r} + \hat{\mathbf{z}}\sin\theta_{\rm r}\right) + \hat{\mathbf{x}}E_{\rm r}^{(s)}\right] e^{i[k_{\rm r}(y\sin\theta_{\rm r}-z\cos\theta_{\rm r})-\omega_{\rm r}t]}\\ \mathbf{E}_{\rm t} &= \left[E_{\rm t}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_{\rm t} - \hat{\mathbf{z}}\sin\theta_{\rm t}\right) + \hat{\mathbf{x}}E_{\rm t}^{(s)}\right] e^{i[k_{\rm t}(y\sin\theta_{\rm t}+z\cos\theta_{\rm t})-\omega_{\rm t}t]}\end{aligned} \tag{3.2}$$

Each field has the form (2.7), and we have utilized the k-vectors (3.1) in the exponents of (3.2).

Now we are ready to apply a boundary condition on the fields. The tangential component of $\mathbf{E}$ (parallel to the surface) must be identical on either side of the plane $z = 0$, as explained in appendix 3.A (see (3.52)). This means that at $z = 0$ the parallel components (in the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ directions only) of the combined incident and reflected fields must match the parallel components of the transmitted field:

$$\begin{aligned}&\left[E_{\rm i}^{(p)}\hat{\mathbf{y}}\cos\theta_{\rm i} + \hat{\mathbf{x}}E_{\rm i}^{(s)}\right] e^{i(k_{\rm i}y\sin\theta_{\rm i}-\omega_{\rm i}t)} + \left[E_{\rm r}^{(p)}\hat{\mathbf{y}}\cos\theta_{\rm r} + \hat{\mathbf{x}}E_{\rm r}^{(s)}\right] e^{i(k_{\rm r}y\sin\theta_{\rm r}-\omega_{\rm r}t)}\\ &\qquad = \left[E_{\rm t}^{(p)}\hat{\mathbf{y}}\cos\theta_{\rm t} + \hat{\mathbf{x}}E_{\rm t}^{(s)}\right] e^{i(k_{\rm t}y\sin\theta_{\rm t}-\omega_{\rm t}t)}\end{aligned} \tag{3.3}$$

Since this equation must hold for all conceivable values of t and y, we are compelled to set all exponential factors equal to each other. This requires the frequency of all waves to be the same:

$$\omega_{\rm i} = \omega_{\rm r} = \omega_{\rm t} \equiv \omega \tag{3.4}$$

(We could have guessed that all frequencies would be the same; otherwise wave fronts would be annihilated or created at the interface.) Equating the terms in the exponents of (3.3) also requires

$$k_{\rm i}\sin\theta_{\rm i} = k_{\rm r}\sin\theta_{\rm r} = k_{\rm t}\sin\theta_{\rm t} \tag{3.5}$$

[1] Many textbooks draw the arrow for $E_{\rm r}^{(p)}$ in the direction opposite of ours. However, that choice leads to an awkward situation at normal incidence (i.e. $\theta_{\rm i} = \theta_{\rm r} = 0$) where the arrows for the incident and reflected fields are parallel for the s-component but anti parallel for the p-component.

Willebrord Snell

(1580–1626, Dutch)

Snell was an astronomer and mathematician. He is probably most famous for determining the law that connects refracted angles to incident angles when waves come to a boundary. He was an accomplished mathematician, and developed a new method for calculating π, and an improved method for measuring the circumference of the earth.

Now recall from (2.21) the relations $k_\text{i} = k_\text{r} = n_\text{i}\omega/c$ and $k_\text{t} = n_\text{t}\omega/c$. With these relations, (3.5) yields the law of reflection

$$\theta_\text{r} = \theta_\text{i} \tag{3.6}$$

and Snell's law

$$n_\text{i} \sin\theta_\text{i} = n_\text{t} \sin\theta_\text{t} \tag{3.7}$$

The three angles θ_i, θ_r, and θ_t are not independent. The reflected angle matches the incident angle, and the transmitted angle obeys Snell's law. The phenomenon of *refraction* refers to the fact that θ_i and θ_t are different.

Because the exponents are all identical, (3.3) reduces to two relatively simple equations (one for each dimension, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$):

$$E_\text{i}^{(s)} + E_\text{r}^{(s)} = E_\text{t}^{(s)} \tag{3.8}$$

and

$$\left(E_\text{i}^{(p)} + E_\text{r}^{(p)}\right) \cos\theta_\text{i} = E_\text{t}^{(p)} \cos\theta_\text{t} \tag{3.9}$$

We have derived these equations from the simple boundary condition (3.52) on the parallel component of the electric field. We have yet to use the boundary condition (3.56) on the parallel component of the magnetic field, from which we can derive two similar but distinct equations.

From Maxwell's equation (1.37), we have for a plane wave

$$\mathbf{B} = \frac{\mathbf{k} \times \mathbf{E}}{\omega} = \frac{n}{c}\hat{\mathbf{u}} \times \mathbf{E} \tag{3.10}$$

where $\hat{\mathbf{u}} \equiv \mathbf{k}/k$ is a unit vector in the direction of $\mathbf{k}$. We have also utilized (2.21). This expression is useful to obtain expressions for $\mathbf{B}_\text{i}$, $\mathbf{B}_\text{r}$, and $\mathbf{B}_\text{t}$ in terms of the electric field components that we have already introduced. By injecting (3.1) and (3.2) into (3.10), the incident, reflected, and transmitted magnetic fields are seen to be

$$\begin{aligned}
\mathbf{B}_{\mathrm{i}} &= \frac{n_{\mathrm{i}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{i}}^{(p)} + E_{\mathrm{i}}^{(s)}\left(-\hat{\mathbf{z}}\sin\theta_{\mathrm{i}} + \hat{\mathbf{y}}\cos\theta_{\mathrm{i}}\right)\right] e^{i[k_{\mathrm{i}}(y\sin\theta_{\mathrm{i}} + z\cos\theta_{\mathrm{i}}) - \omega_{\mathrm{i}}t]} \\
\mathbf{B}_{\mathrm{r}} &= \frac{n_{\mathrm{r}}}{c}\left[\hat{\mathbf{x}}E_{\mathrm{r}}^{(p)} + E_{\mathrm{r}}^{(s)}\left(-\hat{\mathbf{z}}\sin\theta_{\mathrm{r}} - \hat{\mathbf{y}}\cos\theta_{\mathrm{r}}\right)\right] e^{i[k_{\mathrm{r}}(y\sin\theta_{\mathrm{r}} - z\cos\theta_{\mathrm{r}}) - \omega_{\mathrm{r}}t]} \\
\mathbf{B}_{\mathrm{t}} &= \frac{n_{\mathrm{t}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{t}}^{(p)} + E_{\mathrm{t}}^{(s)}\left(-\hat{\mathbf{z}}\sin\theta_{\mathrm{t}} + \hat{\mathbf{y}}\cos\theta_{\mathrm{t}}\right)\right] e^{i[k_{\mathrm{t}}(y\sin\theta_{\mathrm{t}} + z\cos\theta_{\mathrm{t}}) - \omega_{\mathrm{t}}t]}
\end{aligned} \tag{3.11}$$

Next, we apply the boundary condition (3.56), which requires the components of $\mathbf{B}$ parallel to the surface (i.e. the components in the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ directions) to be the same on either side of the plane $z = 0$. Since we already know that the exponents are all equal and that $\theta_{\mathrm{r}} = \theta_{\mathrm{i}}$ and $n_{\mathrm{i}} = n_{\mathrm{r}}$, the boundary condition gives

$$\frac{n_{\mathrm{i}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{i}}^{(p)} + E_{\mathrm{i}}^{(s)}\hat{\mathbf{y}}\cos\theta_{\mathrm{i}}\right] + \frac{n_{\mathrm{i}}}{c}\left[\hat{\mathbf{x}}E_{\mathrm{r}}^{(p)} - E_{\mathrm{r}}^{(s)}\hat{\mathbf{y}}\cos\theta_{\mathrm{i}}\right] = \frac{n_{\mathrm{t}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{t}}^{(p)} + E_{\mathrm{t}}^{(s)}\hat{\mathbf{y}}\cos\theta_{\mathrm{t}}\right] \tag{3.12}$$

As before, (3.12) reduces to two relatively simple equations (one for the $\hat{\mathbf{x}}$ dimension and one for the $\hat{\mathbf{y}}$ dimension):

$$n_{\mathrm{i}}\left(E_{\mathrm{i}}^{(p)} - E_{\mathrm{r}}^{(p)}\right) = n_{\mathrm{t}}E_{\mathrm{t}}^{(p)} \tag{3.13}$$

and

$$n_{\mathrm{i}}\left(E_{\mathrm{i}}^{(s)} - E_{\mathrm{r}}^{(s)}\right)\cos\theta_{\mathrm{i}} = n_{\mathrm{t}}E_{\mathrm{t}}^{(s)}\cos\theta_{\mathrm{t}} \tag{3.14}$$

These two equations (wherein the permeability μ_0 was considered to be the same on both sides of the boundary) together with (3.8) and (3.9) give a complete description of how the fields on each side of the boundary relate to each other. If we choose an incident field $\mathbf{E}_{\mathrm{i}}$, these equations can be used to predict $\mathbf{E}_{\mathrm{r}}$ and $\mathbf{E}_{\mathrm{t}}$. To use these equations, we must break the fields into their respective s and p polarization components. However, (3.8), (3.9), (3.13), and (3.14) are not yet in their most convenient form.

3.3 The Fresnel Coefficients

Augustin Fresnel first developed the equations derived in the previous section. However, at the time he did not have the benefit of Maxwell's equations, since he lived well before Maxwell's time. Instead, Fresnel thought of light as transverse mechanical waves propagating within materials. (We can see why Fresnel was a great proponent of the later-discredited luminiferous ether.) Instead of relating the parallel components of the electric and magnetic fields across the boundary between the materials, Fresnel used the principle that, as a transverse mechanical wave propagates from one material to the other, the two materials should not slip past each other at the interface. This "gluing" of the materials at the interface also forbids the possibility of the materials detaching from one another (creating gaps) or passing through one another as they experience the wave vibration. This mechanical approach to light worked splendidly and explained polarization effects along with the variations in reflectance and transmittance as a function of the incident angle of the light.

Fresnel wrote the relationships between the various plane waves depicted in Fig. 3.1 in terms of coefficients that compare the reflected and transmitted field amplitudes to those of the incident field. He then calculated the ratio of the reflected and transmitted

Augustin Fresnel

(1788–1829, French)

Fresnel was a major proponent of the wave theory of light. He studied polarization, and invented the Fresnel romb for generating circularly polarized light. He also invented the fresnel lens, originally for use in light houses. Today fresnel lenses are used in many applications such as overhead projectors.

field components to the incident field components for each polarization. In the following example, we illustrate this procedure for s-polarized light. It is left as a homework exercise to solve the equations for p-polarized light (see P 3.1).

Example 3.1

Calculate the ratio of transmitted field to the incident field and the ratio of the reflected field to incident field for s-polarized light.

Solution: We use (3.8)

$$E_{\mathrm{i}}^{(s)} + E_{\mathrm{r}}^{(s)} = E_{\mathrm{t}}^{(s)} \qquad [3.8]$$

and (3.14), which with the help of Snell's law is written

$$E_{\mathrm{i}}^{(s)} - E_{\mathrm{r}}^{(s)} = \frac{\sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}} E_{\mathrm{t}}^{(s)} \qquad (3.15)$$

If we add these two equations, we get

$$2E_{\mathrm{i}}^{(s)} = \left[1 + \frac{\sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}\right] E_{\mathrm{t}}^{(s)} \qquad (3.16)$$

and after dividing by $E_{\mathrm{i}}^{(s)}$ and doing a little algebra, we obtain

$$\frac{E_{\mathrm{t}}^{(s)}}{E_{\mathrm{i}}^{(s)}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}.$$

To get the ratio of reflected to incident, we subtract (3.16) from (3.8) to obtain

$$2E_{\mathrm{r}}^{(s)} = \left[1 - \frac{\sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}\right] E_{\mathrm{t}}^{(s)} \qquad (3.17)$$

and then divide (3.17) by (3.16). After a little algebra, we arrive at

$$\frac{E_{\mathrm{r}}^{(s)}}{E_{\mathrm{i}}^{(s)}} = \frac{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} - \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}$$

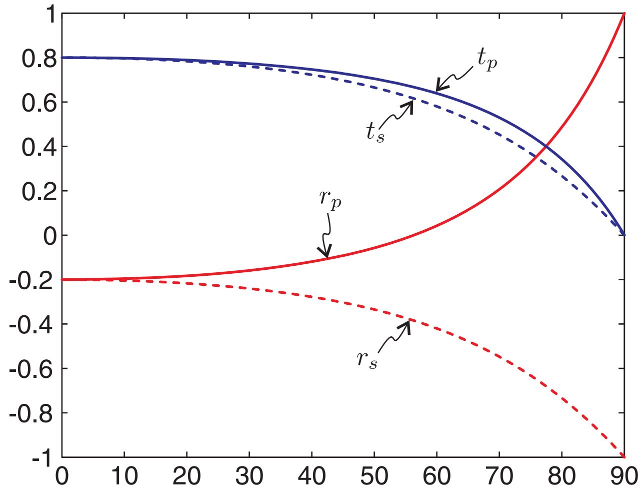


Figure 3.2 The Fresnel coefficients plotted versus θ_{i} for the case of a air-glass interface ($n_{\mathrm{i}} = 1$ and $n_{\mathrm{t}} = 1.5$).

The ratio of the reflected and transmitted field components to the incident field components are specified by the following coefficients, called Fresnel coefficients:

$$r_s \equiv \frac{E_{\mathrm{r}}^{(s)}}{E_{\mathrm{i}}^{(s)}} = \frac{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} - \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}} = -\frac{\sin(\theta_{\mathrm{i}} - \theta_{\mathrm{t}})}{\sin(\theta_{\mathrm{i}} + \theta_{\mathrm{t}})} = \frac{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} - n_{\mathrm{t}}\cos\theta_{\mathrm{t}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + n_{\mathrm{t}}\cos\theta_{\mathrm{t}}} \tag{3.18}$$

$$t_s \equiv \frac{E_{\mathrm{t}}^{(s)}}{E_{\mathrm{i}}^{(s)}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin(\theta_{\mathrm{i}} + \theta_{\mathrm{t}})} = \frac{2n_{\mathrm{i}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + n_{\mathrm{t}}\cos\theta_{\mathrm{t}}} \tag{3.19}$$

$$r_p \equiv \frac{E_{\mathrm{r}}^{(p)}}{E_{\mathrm{i}}^{(p)}} = \frac{\cos\theta_{\mathrm{t}}\sin\theta_{\mathrm{t}} - \cos\theta_{\mathrm{i}}\sin\theta_{\mathrm{i}}}{\cos\theta_{\mathrm{t}}\sin\theta_{\mathrm{t}} + \cos\theta_{\mathrm{i}}\sin\theta_{\mathrm{i}}} = -\frac{\tan(\theta_{\mathrm{i}} - \theta_{\mathrm{t}})}{\tan(\theta_{\mathrm{i}} + \theta_{\mathrm{t}})} = \frac{n_{\mathrm{i}}\cos\theta_{\mathrm{t}} - n_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{t}} + n_{\mathrm{t}}\cos\theta_{\mathrm{i}}} \tag{3.20}$$

$$t_p \equiv \frac{E_{\mathrm{t}}^{(p)}}{E_{\mathrm{i}}^{(p)}} = \frac{2\cos\theta_{\mathrm{i}}\sin\theta_{\mathrm{t}}}{\cos\theta_{\mathrm{t}}\sin\theta_{\mathrm{t}} + \cos\theta_{\mathrm{i}}\sin\theta_{\mathrm{i}}} = \frac{2\cos\theta_{\mathrm{i}}\sin\theta_{\mathrm{t}}}{\sin(\theta_{\mathrm{i}} + \theta_{\mathrm{t}})\cos(\theta_{\mathrm{i}} - \theta_{\mathrm{t}})} = \frac{2n_{\mathrm{i}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{t}} + n_{\mathrm{t}}\cos\theta_{\mathrm{i}}} \tag{3.21}$$

All of the above forms of the Fresnel coefficients are commonly used. Remember that the angles in the coefficient cannot be independently chosen, but are subject to Snell's law (3.7). (The right-most form of each coefficient is obtained from the other forms using Snell's law).

The Fresnel coefficients allow us to easily connect the electric field amplitudes on the two sides of the boundary. They also keep track of phase shifts at a boundary. In Fig. 3.2 we have plotted the Fresnel coefficients for the case of a air-glass interface. Notice that the reflection coefficients are sometimes negative in this plot, which corresponds to a phase shift of π upon reflection (remember $e^{i\pi} = -1$). Later we will see that when absorbing materials are encountered, more complicated phase shifts can arise due to the complex index of refraction.

3.4 Reflectance and Transmittance

We are often interested in knowing the fraction of intensity that transmits through or reflects from a boundary. Since intensity is proportional to the square of the amplitude of

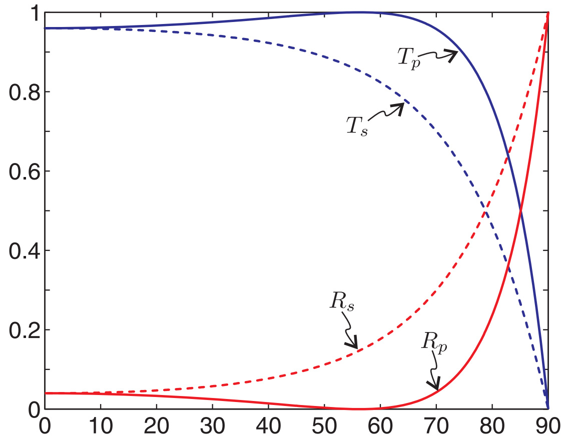


Figure 3.3 The reflectance and transmittance plotted versus θ_i for the case of an air-glass interface ($n_\mathrm{i} = 1$ and $n_\mathrm{t} = 1.5$).

the electric field, we can write the fraction of the light reflected from the surface (called *reflectance*) in terms of the Fresnel coefficients as

$$R_s \equiv |r_s|^2 \qquad \text{and} \qquad R_p \equiv |r_p|^2 \tag{3.22}$$

These expressions are applied individually to each polarization component (s or p). The intensity reflected for each of these orthogonal polarizations is additive because the two electric fields are orthogonal and do not interfere with each other. The total reflected intensity is therefore

$$I_\mathrm{r}^{(\mathrm{total})} = I_\mathrm{r}^{(s)} + I_\mathrm{r}^{(p)} = R_s I_\mathrm{i}^{(s)} + R_p I_\mathrm{i}^{(p)} \tag{3.23}$$

where the incident intensity is given by (2.61):

$$I_\mathrm{i}^{(\mathrm{total})} = I_\mathrm{i}^{(s)} + I_\mathrm{i}^{(p)} = \frac{1}{2} n_\mathrm{i} \epsilon_0 c \left[\left| E_\mathrm{i}^{(s)} \right|^2 + \left| E_\mathrm{i}^{(p)} \right|^2 \right] \tag{3.24}$$

Since intensity is power per area, we can rewrite (3.23) as incident and reflected power:

$$P_\mathrm{r}^{(\mathrm{total})} = P_\mathrm{r}^{(s)} + P_\mathrm{r}^{(p)} = R_s P_\mathrm{i}^{(s)} + R_p P_\mathrm{i}^{(p)} \tag{3.25}$$

Using this expression and requiring that energy be conserved (i.e. $P_\mathrm{i}^{(\mathrm{total})} = P_\mathrm{r}^{(\mathrm{total})} + P_\mathrm{t}^{(\mathrm{total})}$), we find the fraction of the power that transmits:

$$\begin{aligned} P_\mathrm{t}^{(\mathrm{total})} &= \left(P_\mathrm{i}^{(\mathrm{s})} + P_\mathrm{i}^{(\mathrm{p})} \right) - \left(P_\mathrm{r}^{(\mathrm{s})} + P_\mathrm{r}^{(\mathrm{p})} \right) \\ &= (1 - R_s) P_\mathrm{i}^{(\mathrm{s})} + (1 - R_p) P_\mathrm{i}^{(\mathrm{p})} \end{aligned} \tag{3.26}$$

From this expression we see that the *transmittance* (i.e. the fraction of the light that transmits) for either polarization is

$$T_s \equiv 1 - R_s \qquad \text{and} \qquad T_p \equiv 1 - R_p \tag{3.27}$$

Figure 3.3 shows typical reflectance and transmittance values for an air-glass interface.

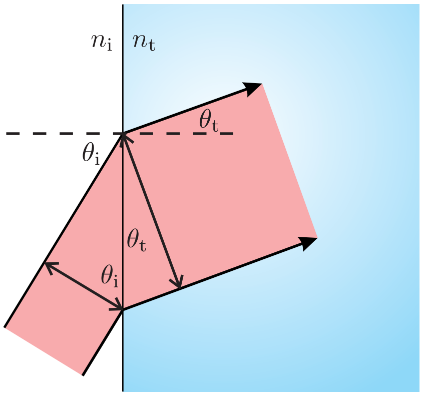


Figure 3.4 Light refracting into a surface.

You might be surprised at first to learn that

$$T_s \neq |t_s|^2 \qquad \text{and} \qquad T_p \neq |t_p|^2 \tag{3.28}$$

However, recall that the transmitted intensity (in terms of the transmitted fields) depends also on the refractive index. The Fresnel coefficients t_s and t_p relate the bare electric fields to each other, whereas the transmitted intensity (similar to (3.24)) is

$$I_\mathrm{t}^{(\text{total})} = I_\mathrm{t}^{(s)} + I_\mathrm{t}^{(p)} = \frac{1}{2} n_\mathrm{t} \epsilon_0 c \left[\left| E_\mathrm{t}^{(s)} \right|^2 + \left| E_\mathrm{t}^{(p)} \right|^2 \right] \tag{3.29}$$

Therefore, we expect T_s and T_p to depend on the ratio of the refractive indices n_t and n_i as well as on the squares of t_s and t_p.

There is another more subtle reason for the inequalities in (3.28). Consider a lateral strip of the power associated with a plane wave incident upon the material interface in Fig. 3.4. Upon refraction into the second medium, the strip is seen to change its width by the factor $\cos\theta_\mathrm{t}/\cos\theta_\mathrm{i}$. This is a geometrical artifact, owing to the change in propagation direction at the interface. The change in direction alters the intensity (power per area) but not the power. In computing the transmittance, we must remove this geometrical effect from the ratio of the intensities, which leads to the following transmittance coefficients:

$$\begin{aligned} T_s &= \frac{n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}} |t_s|^2 \\ T_p &= \frac{n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}} |t_p|^2 \end{aligned} \qquad \text{(valid when no total internal reflection)} \tag{3.30}$$

Note that (3.30) is valid only if a real angle θ_t exists; it does not hold when the incident angle exceeds the critical angle for total internal reflection, discussed in section 3.6. In that situation, we must stick with (3.27).

Example 3.2

Show *analytically* for p-polarized light that $R_p + T_\mathrm{p} = 1$, where R_p is given by (3.22) and T_p is given by (3.30).

Solution: From (3.20) we have

$$\begin{aligned} R_p &= \left|\frac{\cos\theta_{\rm t}\sin\theta_{\rm t} - \cos\theta_{\rm i}\sin\theta_{\rm i}}{\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}}\right|^2 \\ &= \frac{\cos^2\theta_{\rm t}\sin^2\theta_{\rm t} - 2\cos\theta_{\rm i}\sin\theta_{\rm i}\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos^2\theta_{\rm i}\sin^2\theta_{\rm i}}{\left(\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}\right)^2} \end{aligned}$$

From (3.21) and (3.30) we have

$$\begin{aligned} T_p &= \frac{n_{\rm t}\cos\theta_{\rm t}}{n_{\rm i}\cos\theta_{\rm i}}\left[\frac{2\cos\theta_{\rm i}\sin\theta_{\rm t}}{\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}}\right]^2 \\ &= \frac{\sin\theta_{\rm i}\cos\theta_{\rm t}}{\sin\theta_{\rm t}\cos\theta_{\rm i}}\frac{4\cos^2\theta_{\rm i}\sin^2\theta_{\rm t}}{\left(\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}\right)^2} \\ &= \frac{4\cos\theta_{\rm i}\sin\theta_{\rm t}\sin\theta_{\rm i}\cos\theta_{\rm t}}{\left(\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}\right)^2} \end{aligned}$$

Then

$$\begin{aligned} R_p + T_p &= \frac{\cos^2\theta_{\rm t}\sin^2\theta_{\rm t} + 2\cos\theta_{\rm i}\sin\theta_{\rm i}\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos^2\theta_{\rm i}\sin^2\theta_{\rm i}}{\left(\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}\right)^2} \\ &= \frac{\left(\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}\right)^2}{\left(\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}\right)^2} \\ &= 1 \end{aligned}$$

3.5 Brewster's Angle

Notice r_p and R_p go to zero at a certain angle in Figs. 3.2 and 3.3, indicating that no p-polarized light is reflected at this angle. This behavior is quite general, as we can see from the second form of the Fresnel coefficient formula for r_p in (3.20), which has $\tan(\theta_{\rm i} + \theta_{\rm t})$ in the denominator. Since the tangent "blows up" at $\pi/2$, the reflection coefficient goes to zero when

$$\theta_{\rm i} + \theta_{\rm t} = \frac{\pi}{2} \qquad \text{(requirement for zero } p\text{-polarized reflection)} \qquad (3.31)$$

By inspecting Fig. 3.1, we see that this condition occurs when the reflected and transmitted k-vectors, $\mathbf{k}_{\rm r}$ and $\mathbf{k}_{\rm t}$, are perpendicular to each other. If we insert (3.31) into Snell's law (3.7), we can solve for the incident angle $\theta_{\rm i}$ that gives rise to this special circumstance:

$$n_{\rm i}\sin\theta_{\rm i} = n_{\rm t}\sin\left(\frac{\pi}{2} - \theta_{\rm i}\right) = n_{\rm t}\cos\theta_{\rm i} \qquad (3.32)$$

The special incident angle that satisfies this equation, in terms of the refractive indices, is found to be

$$\theta_{\rm B} = \tan^{-1}\frac{n_{\rm t}}{n_{\rm i}} \qquad (3.33)$$

We have replaced the specific $\theta_{\rm i}$ with $\theta_{\rm B}$ in honor of Sir David Brewster (1781-1868) who first discovered the phenomenon. The angle $\theta_{\rm B}$ is called Brewster's angle. At Brewster's angle, no p-polarized light reflects (see L 3.6). Physically, the p-polarized light cannot reflect because $\mathbf{k}_\mathrm{r}$ and $\mathbf{k}_\mathrm{t}$ are perpendicular. A reflection would require the microscopic dipoles at the surface of the second material to radiate along their axes, which they cannot do. Maxwell's equations "know" about this, and so everything is nicely consistent.

3.6 Total Internal Reflection

From Snell's law (3.7), we can compute the transmitted angle in terms of the incident angle:

$$\theta_\mathrm{t} = \sin^{-1}\left(\frac{n_\mathrm{i}}{n_\mathrm{t}}\sin\theta_\mathrm{i}\right) \tag{3.34}$$

The angle θ_t is real only if the argument of the inverse sine is less than or equal to one. If $n_\mathrm{i} > n_\mathrm{t}$, we can find a critical angle at which the argument begins to exceed one:

$$\theta_\mathrm{c} \equiv \sin^{-1}\frac{n_\mathrm{t}}{n_\mathrm{i}} \tag{3.35}$$

When $\theta_\mathrm{i} > \theta_\mathrm{c}$, then there is total internal reflection and we can directly show that $R_s = 1$ and $R_p = 1$ (see P 3.8). To demonstrate this, one computes the Fresnel coefficients (3.18) and (3.20) while employing the following substitutions:

$$\sin\theta_\mathrm{t} = \frac{n_\mathrm{i}}{n_\mathrm{t}}\sin\theta_\mathrm{i} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \qquad \text{(Snell's law)} \tag{3.36}$$

and

$$\cos\theta_\mathrm{t} = i\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i} - 1} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \tag{3.37}$$

(see P 0.7).

In this case, θ_t is a complex number. However, we do not assign geometrical significance to it in terms of any direction. Actually, we don't even need to know the value for θ_t; we need only the values for $\sin\theta_\mathrm{t}$ and $\cos\theta_\mathrm{t}$, as specified in (3.36) and (3.37). Even though $\sin\theta_\mathrm{t}$ is greater than one and $\cos\theta_\mathrm{t}$ is imaginary, we can use their values to compute r_s, r_p, t_s, and t_p. (Complex notation is wonderful!)

Upon substitution of (3.36) and (3.37) into the Fresnel reflection coefficients (3.18) and (3.20) we obtain

$$r_s = \frac{\frac{n_\mathrm{i}}{n_\mathrm{t}}\cos\theta_\mathrm{i} - i\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i} - 1}}{\frac{n_\mathrm{i}}{n_\mathrm{t}}\cos\theta_\mathrm{i} + i\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i} - 1}} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \tag{3.38}$$

and

$$r_p = -\frac{\cos\theta_\mathrm{i} - i\frac{n_\mathrm{i}}{n_\mathrm{t}}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i} - 1}}{\cos\theta_\mathrm{i} + i\frac{n_\mathrm{i}}{n_\mathrm{t}}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i} - 1}} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \tag{3.39}$$

These Fresnel coefficients can be manipulated (see P 3.8) into the forms

$$r_s = \exp\left\{-2i\tan^{-1}\left[\frac{n_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i} - 1}\right]\right\} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \tag{3.40}$$

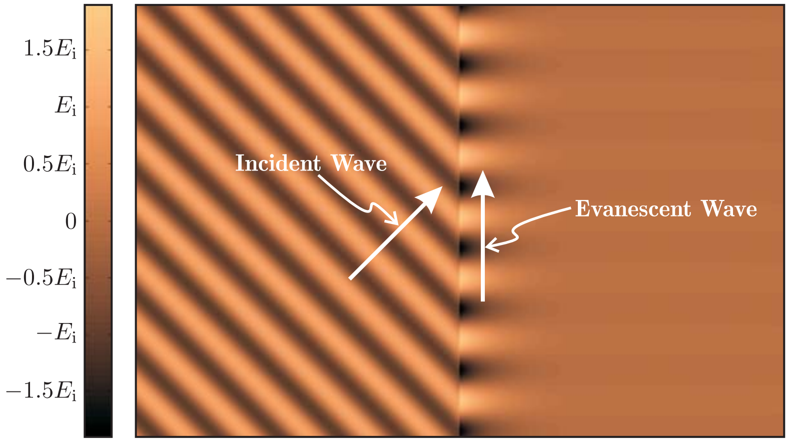


Figure 3.5 An incident wave experiences total internal reflection and creates an evanescent wave which propagates parallel to the interface ($\theta_\mathrm{i} = 45°$, $n_\mathrm{i} = 1.5$, $n_\mathrm{t} = 1$). (The reflected wave is not shown in this figure.)

and

$$r_p = -\exp\left\{-2i\tan^{-1}\left[\frac{n_\mathrm{i}}{n_\mathrm{t}\cos\theta_\mathrm{i}}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}\right]\right\} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \qquad (3.41)$$

Each coefficient has a different phase (note $n_\mathrm{i}/n_\mathrm{t}$ vs. $n_\mathrm{t}/n_\mathrm{i}$ in the expressions), which means that the s- and p-polarized fields experience different phase shifts upon reflection. Nevertheless, we definitely have $|r_s| = 1$ and $|r_p| = 1$. We rightly conclude that 100% of the light reflects. Even so, the boundary conditions from Maxwell's equations (see appendix 3.A) require that the fields be non-zero on the transmitted side of the boundary, meaning $t_s \neq 0$ and $t_\mathrm{p} \neq 0$. This may seem puzzling, but it does not contradict our assertion that 100% of the light reflects. The transmitted power is still zero as dictated by (3.25). For total internal reflection, one should not employ (3.29).

The coefficients t_s and t_p characterize *evanescent waves* that exist on the transmitted side of the interface. The evanescent wave travels *parallel* to the interface so that no energy is conveyed away from the interface deeper into the medium on the transmission side. In the direction perpendicular to the boundary, the strength of the evanescent wave decays exponentially. To compute the explicit form of the evanescent wave, we plug (3.36) and (3.37) into the transmitted field (3.2):

$$\begin{aligned}\mathbf{E}_\mathrm{t} &= \left[E_\mathrm{t}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_\mathrm{t} - \hat{\mathbf{z}}\sin\theta_\mathrm{t}\right) + \hat{\mathbf{x}}E_\mathrm{t}^{(s)}\right]e^{i[k_\mathrm{t}(y\sin\theta_\mathrm{t}+z\cos\theta_\mathrm{t})-\omega t]}\\ &= \left[t_pE_\mathrm{i}^{(p)}\left(\hat{\mathbf{y}}i\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1} - \hat{\mathbf{z}}\frac{n_\mathrm{i}}{n_\mathrm{t}}\sin\theta_\mathrm{i}\right) + \hat{\mathbf{x}}t_sE_\mathrm{i}^{(s)}\right]e^{-k_\mathrm{t}z\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}}\,e^{i\left[k_\mathrm{t}y\frac{n_\mathrm{i}}{n_\mathrm{t}}\sin\theta_\mathrm{i}-\omega t\right]}\end{aligned} \qquad (3.42)$$

Figure 3.5 plots the evanescent wave described by (3.42) along with the associated incident wave. Note that the evanescent wave propagates parallel to the boundary (in the y-dimension) and its strength diminishes away from the boundary (in the z-dimension) as dictated by the exponential terms at the end of (3.42). We leave the calculation of t_s and t_p as an exercise (P 3.9).

3.7 Reflection from Metallic or other Absorptive Surfaces

In this section we extend our analysis to materials with complex refractive index $\mathcal{N} \equiv n+i\kappa$ as studied in chapter 2. As a reminder, the imaginary part of the index controls attenuation of a wave as it propagates within a material. The real part of the index governs the oscillatory nature of the wave. It turns out that both the imaginary and real parts of the index strongly influence the reflection of light from a surface. The reader may be grateful that there is no need to re-derive the Fresnel coefficients (3.18)–(3.21) for the case of complex indices. The coefficients remain valid whether the index is real or complex. We just need to be a bit careful when applying them.

We restrict our discussion to *reflections* from a metallic or other absorbing material surface. To employ Fresnel reflection coefficients (3.18) and (3.20), we actually do not need to know the transmitted angle θ_t. We need only acquire expressions for $\cos\theta_\mathrm{t}$ and $\sin\theta_\mathrm{t}$, and we can obtain these from Snell's law (3.7). To minimize complications, we let the incident refractive index be $n_\mathrm{i} = 1$ (which is often the case). Let the index on the transmitted side be written simply as $\mathcal{N}_\mathrm{t} = \mathcal{N}$. Then by Snell's law the sine of the transmitted angle is

$$\sin\theta_\mathrm{t} = \frac{\sin\theta_\mathrm{i}}{\mathcal{N}} \tag{3.43}$$

This expression is of course complex since $\mathcal{N}$ is complex, but that is just fine. The cosine of the same angle is

$$\cos\theta_\mathrm{t} = \sqrt{1-\sin^2\theta_\mathrm{t}} = \frac{1}{\mathcal{N}}\sqrt{\mathcal{N}^2-\sin^2\theta_\mathrm{i}} \tag{3.44}$$

The positive sign in front of the square root is appropriate since it is clearly the right choice if the imaginary part of the index approaches zero.

Upon substitution of these expressions, the Fresnel reflection coefficients (3.18) and (3.20) become

$$r_s = \frac{\cos\theta_\mathrm{i} - \sqrt{\mathcal{N}^2-\sin^2\theta_\mathrm{i}}}{\cos\theta_\mathrm{i} + \sqrt{\mathcal{N}^2-\sin^2\theta_\mathrm{i}}} \tag{3.45}$$

and

$$r_p = \frac{\sqrt{\mathcal{N}^2-\sin^2\theta_\mathrm{i}} - \mathcal{N}^2\cos\theta_\mathrm{i}}{\sqrt{\mathcal{N}^2-\sin^2\theta_\mathrm{i}} + \mathcal{N}^2\cos\theta_\mathrm{i}} \tag{3.46}$$

These expressions are tedious to evaluate. When evaluating the expressions, it is usually desirable to put them into the form

$$r_s = |r_s|\,e^{i\phi_s} \tag{3.47}$$

and

$$r_p = |r_p|\,e^{i\phi_p} \tag{3.48}$$

However, we refrain from putting (3.45) and (3.46) into this form using the general expressions; we would get a big mess. It is a good idea to let your calculator or a computer do

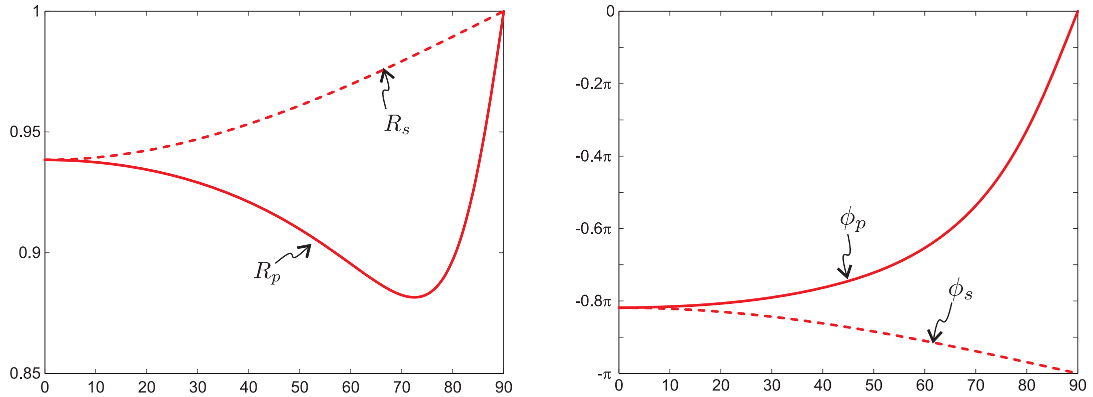


Figure 3.6 The transmittance and reflectance (left) and the phase upon reflection (right) for a metal with $n = 0.2$ $\kappa = 3.4$. Note the minimum of R_p where Brewster's angle occurs.

it after a specific value for $\mathcal{N} \equiv n + i\kappa$ is chosen. An important point to notice is that the phases upon reflection can be very different for s and p-polarization components (i.e. ϕ_p and ϕ_s can be very different). This in general is true even when the reflectivity is high (i.e. $|r_s|$ and $|r_\mathrm{p}|$ on the order of unity).

Brewster's angle exists also for surfaces with complex refractive index. However, in general the expressions (3.46) and (3.48) do not go to zero at any angle θ_i. Rather, the reflection of p-polarized light can go through a minimum at some angle θ_i, which we refer to as Brewster's angle (see Fig. 3.6). This minimum is best found numerically since the general expression for $|r_p|$ in terms of n and κ and as a function of θ_i can be unwieldy.

Appendix 3.A Boundary Conditions For Fields at an Interface

We are interested in the continuity of fields across a boundary from one medium with index n_1 to another medium with index n_2. We will show that the components of electric field parallel to the interface surface must be the same on the two sides of the surface (adjacent to the interface). This result is independent of the refractive index of the materials. We will also show that the component of magnetic field parallel to the interface surface is the same on the two sides (assuming the permeability μ_0 is the same on both sides).

Consider a surface S (a rectangle) that is *perpendicular* to the interface between the two media and which extends into both media, as depicted in Fig. 3.7.

First we examine the implications of Faraday's law (1.23):

$$\oint_C \mathbf{E} \cdot d\ell = -\frac{\partial}{\partial t} \int_S \mathbf{B} \cdot \hat{\mathbf{n}}\, da \tag{3.49}$$

We apply Faraday's law to the rectangular contour depicted in Fig. 3.7. We can perform the path integration on the left-hand side of (3.49). The integration around the loop gives

$$\oint \mathbf{E} \cdot d\ell = E_{1||} d - E_{1\perp}\ell_1 - E_{2\perp}\ell_2 - E_{2||} d + E_{2\perp}\ell_2 + E_{1\perp}\ell_1 = \left(E_{1||} - E_{2||}\right) d \tag{3.50}$$

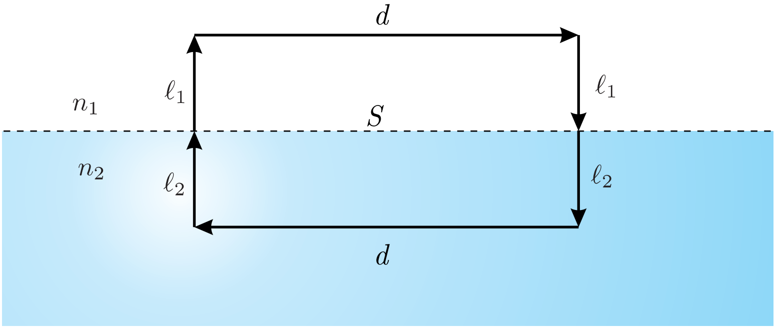


Figure 3.7 Interface of two materials.

Here, $E_{1||}$ refers to the component of the electric field in the material with index n_1 that is parallel to the interface. $E_{1\perp}$ refers to the component of the electric field in the material with index n_1 which is perpendicular to the interface. Similarly, $E_{2||}$ and $E_{2\perp}$ are the parallel and perpendicular components of the electric field in the material with index n_2. We have assumed that the rectangle is small enough that the fields are uniform within the half rectangle on either side of the boundary.

We can continue to shrink the loop down until it has zero surface area by letting the lengths ℓ_1 and ℓ_2 go to zero. In this situation, the right-hand side of Faraday's law goes to zero

$$\int_S \mathbf{B} \cdot \hat{\mathbf{n}}\, da \to 0 \tag{3.51}$$

and we are left with

$$E_{1||} = E_{2||} \tag{3.52}$$

This simple relation is a general boundary condition, which is met at any material interface. The component of the electric field that lies in the plane of the interface must be the same on both sides of the interface.

We now derive a similar boundary condition for the magnetic field. Maxwell's equation (1.38), upon integration over the surface S in Fig. 3.7 and after applying Stokes' theorem (0.27) to the magnetic field term, can be written as

$$\oint_C \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 \int_S \left(\mathbf{J}_{\text{free}} + \frac{\partial \mathbf{P}}{\partial t} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \cdot \hat{\mathbf{n}}\, da \tag{3.53}$$

As before, we are able to perform the path integration on the left-hand side for the geometry depicted in the figure. When we integrate around the loop we get

$$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = B_{1||} d - B_{1\perp}\ell_1 - B_{2\perp}\ell_2 - B_{2||} d + B_{2\perp}\ell_2 + B_{1\perp}\ell_1 = \left(B_{1||} - B_{2||}\right) d \tag{3.54}$$

The notation for parallel and perpendicular components on either side of the interface is similar to that used in (3.50).

Again, we can continue to shrink the loop down until it has zero surface area by letting the lengths ℓ_1 and ℓ_2 go to zero. In this situation, the right-hand side of (3.53) goes to zero (not considering the possibility of surface currents):

$$\int\limits_S \left(\mathbf{J}_{\text{free}} + \frac{\partial \mathbf{P}}{\partial t} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}\right) \cdot \hat{\mathbf{n}}\, da \rightarrow 0 \tag{3.55}$$

and we are left with

$$B_{1||} = B_{2||} \tag{3.56}$$

This is a general boundary condition that must be satisfied at the material interface.

Exercises

3.3 The Fresnel Coefficients

P3.1 Derive the Fresnel coefficients (3.20) and (3.21) for *p*-polarized light.

P3.2 Verify the first alternative form given in each of (3.18)–(3.21).

P3.3 Verify the alternative forms given in each of (3.18)–(3.21). Show that at normal incidence (i.e. $\theta_\mathrm{i} = \theta_\mathrm{t} = 0$) the Fresnel coefficients reduce to

$$\lim_{\theta_\mathrm{i}\to 0} r_s = \lim_{\theta_\mathrm{i}\to 0} r_p = -\frac{n_\mathrm{t} - n_\mathrm{i}}{n_\mathrm{t} + n_\mathrm{i}}$$

and

$$\lim_{\theta_\mathrm{i}\to 0} t_s = \lim_{\theta_\mathrm{i}\to 0} t_p = \frac{2n_\mathrm{i}}{n_\mathrm{t} + n_\mathrm{i}}$$

P3.4 Undoubtedly the most important interface in optics is when air meets glass. Use a computer graphing program to make the following plots for this interface as a function of the incident angle. Use $n_\mathrm{i} = 1$ for air and $n_\mathrm{t} = 1.54$ for glass. Explicitly label Brewster's angle on all of the applicable graphs.

(a) r_p and t_p (plot together on same graph)

(b) R_p and T_p (plot together on same graph)

(c) r_s and t_s (plot together on same graph)

(d) R_s and T_s (plot together on same graph)

3.4 Reflectance and Transmittance

P3.5 Show *analytically* for *s*-polarized light that $R_s + T_s = 1$, where R_s is given by (3.22) and T_s is given by (3.30).

L3.6 Use a computer to calculate the theoretical air-to-glass reflectance as a function of incident angle (i.e. plot R_s and R_p as a function of θ_i). Take the index of refraction for glass to be $n_\mathrm{t} = 1.54$ and the index for air to be one. Plot this theoretical calculation as a smooth line on a graph.

In the laboratory, measure the reflectance for both *s* and *p* polarized light at about ten points, and plot the points on your graph (not points connected by lines). You can normalize the detector by placing it in the incident beam of light before the glass surface. Especially watch for Brewster's angle (described in section 3.5). Figure 3.8 illustrates the experimental setup.

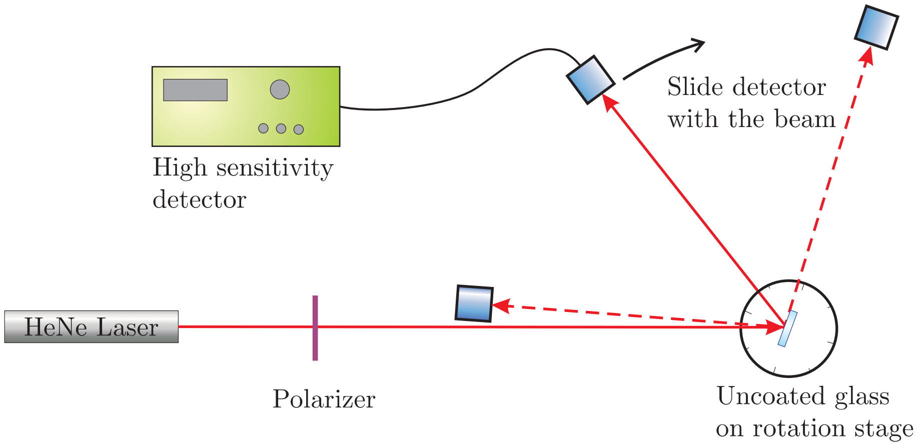


Figure 3.8 Experimental setup for lab 3.6.

3.5 Brewster's Angle

P3.7 Find Brewster's angle for glass $n = 1.5$.

3.6 Total Internal Reflection

P3.8 Derive (3.40) and (3.41) and show that $R_s = 1$ and $R_p = 1$.

HINT:

$$\frac{a - ib}{a + ib} = \frac{\sqrt{a^2 + b^2} e^{-i \tan^{-1} \frac{b}{a}}}{\sqrt{a^2 + b^2} e^{i \tan^{-1} \frac{b}{a}}} = \frac{e^{-i \tan^{-1} \frac{b}{a}}}{e^{i \tan^{-1} \frac{b}{a}}} = e^{-2i \tan^{-1} \frac{b}{a}}$$

where a is positive and real and b is real.

P3.9 Compute t_s and t_p in the case of total internal reflection.

P3.10 Use a computer to plot the air-to-water transmittance as a function of incident angle (i.e. plot (3.27) as a function of θ_i). Also plot the water-to-air transmittance on a separate graph. Plot both T_s and T_p on each graph. The index of refraction for water is $n = 1.33$. Take the index of air to be one.

P3.11 Light ($\lambda_\text{vac} = 500$ nm) reflects internally from a glass surface ($n = 1.5$) surrounded by air. The incident angle is $\theta_\text{i} = 45°$. An evanescent wave travels parallel to the surface on the air side. At what distance from the surface is the amplitude of the evanescent wave $1/e$ of its value at the surface?

3.7 Reflection from Metallic or other Absorptive Surfaces

P3.12 The complex index for silver is given by $n = 0.2$ and $\kappa = 3.4$. Find r_s and r_p when $\theta_\text{i} = 80°$ and put them into the forms (3.47) and (3.48). Find the result using the rules of complex arithmetic and real-valued function on your calculator. (You can use the complex number abilities of your calculator to check your answer.)

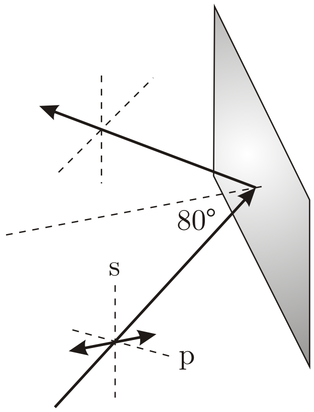


Figure 3.9 Geometry for P 3.12

P3.13 Using a computer graphing program that understands complex numbers (e.g. Matlab), plot $|r_s|$, $|r_p|$ versus θ_i for silver ($n = 0.2$ and $\kappa = 3.4$). Make a separate plot of the phases ϕ_s and ϕ_p from (3.47) and (3.48). Clearly label each plot, and comment on how the phase shifts are different from those experienced when reflecting from glass.

P3.14 Find Brewster's angle for silver ($n = 0.2$ and $\kappa = 3.4$) by calculating R_p and finding its minimum. You will want to use a computer program to do this (Matlab, Maple, Mathematica, etc.).

Chapter 4

Polarization

4.1 Linear, Circular, and Elliptical Polarization

Consider the plane-wave solution to Maxwell's equations given by

$$\mathbf{E}\left(\mathbf{r},t\right) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{4.1}$$

The wave vector $\mathbf{k}$ specifies the direction of propagation. We neglect absorption so that the refractive index is real and $k = n\omega/c = 2\pi n/\lambda_{\text{vac}}$ (see (2.21)–(2.26)). In an isotropic medium, $\mathbf{k}$ and $\mathbf{E}_0$ are perpendicular. Thus, once the direction of $\mathbf{k}$ is specified, $\mathbf{E}_0$ is still only confined to two dimensions. If we orient our coordinate system with the z-axis in the direction of $\mathbf{k}$, we can write (4.1) as

$$\mathbf{E}\left(z,t\right) = \left(E_{0x}\hat{\mathbf{x}} + E_{0y}\hat{\mathbf{y}}\right) e^{i(kz-\omega t)} \tag{4.2}$$

Only the real part of (4.2) is physically relevant. The complex amplitudes of E_{0x} and E_{0y} keep track of the phase of the oscillating field components. In general the complex phases of E_{0x} and E_{0y} can differ, so that the wave in one of the dimensions lags or leads the wave in the other dimension.

The relationship between E_{0x} and E_{0y} describes the *polarization* of the light. For example, if the y-component of the field E_{0y} is zero, the plane wave is said to be *linearly polarized* along the x-dimension. Linearly polarized light can have any orientation in the x–y plane, and it occurs whenever E_{0x} and E_{0y} have the same complex phase (or differ by an integer times π). We often take the x-dimension to be horizontal and the y-dimension to be vertical.

As an example, suppose $E_{0y} = iE_{0x}$, where E_{0x} is real. The y-component of the field is then out of phase with the x-component by the factor $i = e^{i\pi/2}$. Taking the real part of the field (4.2) we get

$$\begin{aligned}\mathbf{E}\left(z,t\right) &= \text{Re}\left[E_{0x}e^{i(kz-\omega t)}\right]\hat{\mathbf{x}} + \text{Re}\left[e^{i\pi/2}E_{0x}e^{i(kz-\omega t)}\right]\hat{\mathbf{y}} \\ &= E_{0x}\cos\left(kz-\omega t\right)\hat{\mathbf{x}} + E_{0x}\cos\left(kz-\omega t+\pi/2\right)\hat{\mathbf{y}} \qquad \text{(left circular)} \\ &= E_{0x}\left[\cos\left(kz-\omega t\right)\hat{\mathbf{x}} - \sin\left(kz-\omega t\right)\hat{\mathbf{y}}\right]\end{aligned} \tag{4.3}$$

In this example, the field in the y-dimension lags the field in the x-dimension by a quarter cycle. That is, the behavior seen in the x-dimension happens in the y-dimension a quarter

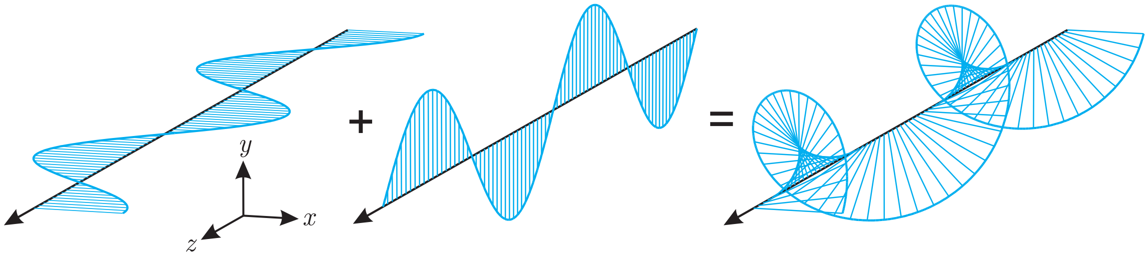


Figure 4.1 The combination of two orthogonally polarized plane waves that are out of phase results in elliptically polarized light. Here we have left circularly polarized light created as specified by (4.3).

cycle later. The field never goes to zero simultaneously in both dimensions. In fact, in this example the strength of the electric field is constant, and it rotates in a circular pattern in the $x-y$ dimensions. For this reason, this type of field is called *circularly polarized*. Figure 4.1 graphically shows the two linear polarized pieces in (4.3) adding to make circularly polarized light.

If we view the field in (4.3) throughout space at a frozen instant in time, the electric field vector spirals as we move along the z-dimension. If the sense of the spiral (with time frozen) matches that of a common wood screw oriented along the z-axis, the polarization is called *right handed*. (It makes no difference whether the screw is flipped end for end.) If instead the field spirals in the opposite sense, then the polarization is called *left handed*. The field in (4.3) is an example of left-handed circularly polarized light.

An equivalent way to view the handedness convention is to imagine the light impinging on a screen as a function of time. The field of a right-handed circularly polarized wave rotates counter clockwise at the screen, when looking along the **k** direction (towards the front side of the screen). The field rotates clockwise for a left-handed circularly polarized wave.

In the next section, we develop a convenient way for keeping track of polarization in terms of a two-dimensional vector, called the Jones vector. In section 4.3, we introduce polarizing filters and describe how their effect on a light field can be represented as a 2×2 matrix operating on the polarization vector. In subsequent sections we show how to deal with polarizers oriented at arbitrary angles with respect to the coordinate system. The analysis applies also to wave plates, devices that retard one field component with respect to the other. A wave plate is used to convert, for example, linearly polarized light into circularly polarized light.

Beginning in section 4.6, we investigate how reflection and transmission at a material interface influences field polarization. The Fresnel coefficients studied in the previous chapter can be conveniently incorporated into the 2×2 matrix formulation for handling polarization. As we saw, the amount of light reflected from a surface depends on the type of polarization, s or p. In addition, upon reflection, s-polarized light can acquire a phase lag or phase advance relative to p-polarized light. This is especially true at metal surfaces, which have complex indices of refraction (i.e. highly absorptive).

Linear polarized light can become circularly or, in general, *elliptically* polarized after reflection from a metal surface if the incident light has both s- and p-polarized components.

R. Clark Jones

(1916–2004, United States)

Jones was educated at Harvard and spent his professional career working for Polaroid corporation. He is well-known for his work in polarization, but also studied many other fields. He was an avid train enthusiast, and even wrote papers on railway engineering.

Every good experimentalist working with light needs to know this. For reflections involving materials with real indices such as glass (for visible light), the situation is less complicated and linearly polarized light remains linear. However, even if the index is real, there are interesting phase shifts (different for s and p components) for total internal reflection. In section 4.7 we briefly discuss *ellipsometry*, which is the science of characterizing optical properties of materials by observing the polarization of light reflected from surfaces.

Throughout this chapter, we consider light to have well characterized polarization. However, most natural sources of light have rapidly varying, random polarization (e.g. sunlight or the light from an incandescent lamp). Such sources are commonly referred to as *unpolarized.* It is possible to have a mixture of unpolarized and polarized light, called *partially polarized* light. In appendix 4.A, we describe a formalism for dealing with light having an arbitrary degree of polarization of an arbitrary kind.

4.2 Jones Vectors for Representing Polarization

In 1941, R. Clark Jones introduced a two-dimensional matrix algebra that is useful for keeping track of light polarization and the effects of optical elements that influence polarization. The algebra deals with light having a definite polarization, such as plane waves. It does not apply to un-polarized or partially polarized light (e.g. sunlight). For partially polarized light, a four-dimensional algebra known as Stokes calculus is used (see Appendix 4.A).

In preparation for introducing Jones vectors, we explicitly write the complex phases of the field components in (4.2) as

$$\mathbf{E}\left(z,t\right)=\left(|E_{0x}|e^{i\delta_x}\hat{\mathbf{x}}+|E_{0y}|e^{i\delta_y}\hat{\mathbf{y}}\right)e^{i(kz-\omega t)} \tag{4.4}$$

and then factor (4.4) as follows:

$$\mathbf{E}\left(z,t\right)=E_{\text{eff}}\left(A\hat{\mathbf{x}}+Be^{i\delta}\hat{\mathbf{y}}\right)e^{i(kz-\omega t)} \tag{4.5}$$

where

$$E_{\text{eff}} \equiv \sqrt{|E_{0x}|^2 + |E_{0y}|^2}\, e^{i\delta_x} \tag{4.6}$$

$$A \equiv \frac{|E_{0x}|}{\sqrt{|E_{0x}|^2 + |E_{0y}|^2}} \tag{4.7}$$

$$B \equiv \frac{|E_{0y}|}{\sqrt{|E_{0x}|^2 + |E_{0y}|^2}} \tag{4.8}$$

$$\delta \equiv \delta_y - \delta_x \tag{4.9}$$

Please notice that A and B are real non-negative dimensionless numbers that satisfy $A^2 + B^2 = 1$. If the x-component of the field E_{0x} happens to be zero, then its phase $e^{i\delta_x}$ is indeterminant. In this case we let $E_{\text{eff}} = |E_{0y}|e^{i\delta_y}$, $B = 1$, and $\delta = 0$. (If E_{0y} is zero, then $e^{i\delta_y}$ is indeterminant. However, this is not a problem since $B = 0$ in this case, so that(4.5) is still well-defined.)

The overall field strength E_{eff} is often unimportant in a discussion of polarization. It represents the strength of an effective linearly polarized field that would give the same intensity that (4.4) would yield. Specifically, from (4.5) and (2.61) we have

$$I = \langle S \rangle_t = \frac{1}{2} n c \epsilon_0 \mathbf{E}_0 \cdot \mathbf{E}_0^* = \frac{1}{2} n c \epsilon_0 \left| E_{\text{eff}} \right|^2 \tag{4.10}$$

The phase of E_{eff} represents an overall phase shift that one can trivially adjust by physically moving the light source (a laser, say) forward or backward by a fraction of a wavelength.

The portion of (4.5) that is interesting in the current discussion is the vector $A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}$, referred to as the *Jones vector*. This vector contains the essential information regarding field polarization. Notice that the Jones vector is a kind of unit vector, in that $(A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}) \cdot (A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}})^* = 1$ (the asterisk represents the complex conjugate). When writing a Jones vector we dispense with the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ notation and organize the components into a column vector (for later use in matrix algebra) as follows:

$$\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} \tag{4.11}$$

This vector can describe the polarization state of any plane wave field. Table 4.1 lists a number of Jones vectors representing various polarization states. The last Jones vector in the table corresponds to the example given in (4.3). All of the vectors in Table 4.1 are special cases of the general Jones vector (4.11).

In general, (4.11) represents a polarization state in between linear and circular. This "in-between" state is known as *elliptically polarized* light. As the wave travels, the field vector undergoes a spiral motion. If we observe the field vector at a point as the field goes by, the field vector traces out an ellipse oriented perpendicular to the direction of travel (i.e. in the x–y plane). One of the axes of the ellipse occurs at the angle (see P 4.8)

$$\alpha = \frac{1}{2} \tan^{-1} \left(\frac{2AB\cos\delta}{A^2 - B^2} \right) \tag{4.12}$$

Vector	Description
$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	linearly polarized along x-dimension
$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	linearly polarized along y-dimension
$\begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix}$	linearly polarized at an angle α from the x-axis
$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$	right circularly polarized
$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$	left circularly polarized

Table 4.1 Jones Vectors for various polarization states

with respect to the x-axis. This angle sometimes corresponds to the minor axis and sometimes to the major axis of the ellipse, depending on the exact values of A, B, and δ. The other axis of the ellipse (major or minor) then occurs at $\alpha \pm \pi/2$ (see Fig. 4.2). We can deduce whether (4.12) corresponds to the major or minor axis of the ellipse by comparing the strength of the electric field when it spirals through the direction specified by α and when it spirals through $\alpha \pm \pi/2$. The strength of the electric field at α is given by (see P 4.8)

$$E_\alpha = |E_{\text{eff}}| \sqrt{A^2\cos^2\alpha + B^2\sin^2\alpha + AB\cos\delta\sin 2\alpha} \qquad (E_{\text{max}} \text{ or } E_{\text{min}}) \tag{4.13}$$

and the strength of the field when it spirals through the orthogonal direction ($\alpha \pm \pi/2$) is given by

$$E_{\alpha\pm\pi/2} = |E_{\text{eff}}| \sqrt{A^2\sin^2\alpha + B^2\cos^2\alpha - AB\cos\delta\sin 2\alpha} \qquad (E_{\text{max}} \text{ or } E_{\text{min}}) \tag{4.14}$$

After computing (4.13) and (4.14), we decide which represents E_{min} and which E_{max} according to

$$E_{\text{max}} \geq E_{\text{min}} \tag{4.15}$$

(We could predict in advance which of (4.13) and (4.14) corresponds to the major axis and which corresponds to the minor axis. However, making this prediction is as complicated as simply evaluating (4.13) and (4.14) and determining which is greater.)

Elliptically polarized light is often characterized by the ratio of the minor axis to the major axis. This ratio is called the *ellipticity*, which is a dimensionless number:

$$e \equiv \frac{E_{\text{min}}}{E_{\text{max}}} \tag{4.16}$$

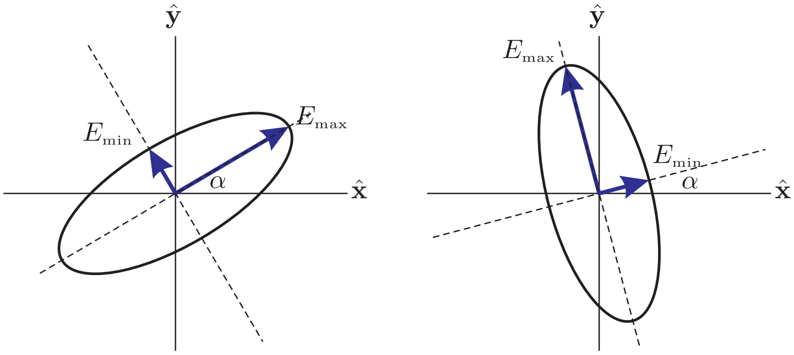


Figure 4.2 The electric field of elliptically polarized light traces an ellipse in the plane perpendicular to its propagation direction. Depending on the values of A, B, and δ, the angle α can describe the major axis (left figure) or the minor axis (right figure).

The ellipticity e ranges between zero (corresponding to linearly polarized light) and one (corresponding to circularly polarized light). Finally, the *helicity* or handedness of elliptically polarized light is as follows (see P 4.2):

$$0 < \delta < \pi \qquad \text{(left-handed helicity)} \tag{4.17}$$

$$\pi < \delta < 2\pi \qquad \text{(right-handed helicity)} \tag{4.18}$$

4.3 Jones Matrices

In 1928, Edwin Land invented Polaroid at the age of nineteen. He did it by stretching a polymer sheet and infusing it with iodine. The stretching causes the polymer chains to align along a common direction, whereupon the sheet is cemented to a substrate. The infusion of iodine causes the individual chains to become conductive. When light impinges upon the Polaroid sheet, the component of electric field that is *parallel* to the polymer chains causes a current $\mathbf{J}_{\rm free}$ to oscillate in that dimension. The resistance to the current quickly dissipates the energy (i.e. the refractive index is complex) and the light is absorbed. The thickness of the Polaroid sheet is chosen sufficiently large to ensure that virtually none of the light with electric field component oscillating along the chains makes it through the device.

The component of electric field that is orthogonal to the polymer chains encounters electrons that are essentially bound, unable to leave their polymer chains. For this polarization component, the wave passes through the material like it does through typical dielectrics such as glass (i.e. the refractive index is real). Today, there are a wide variety of technologies for making polarizers, many very different from Polaroid.

A polarizer can be represented as a 2 × 2 matrix that operates on Jones vectors. The function of a polarizer is to pass only the component of electric field that is oriented along the polarizer transmission axis (perpendicular to the polymer chains). Thus, if a polarizer is oriented with its transmission axis along the x-dimension, then only the x-component of polarization transmits; the y-component is killed. If the polarizer is oriented with its

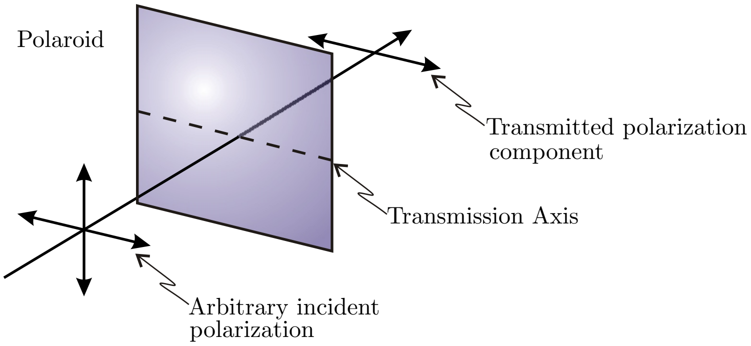

Polaroid
Transmitted polarization component
Transmission Axis
Arbitrary incident polarization

Figure 4.3 Light transmitting through a Polaroid sheet.

transmission axis along the y-dimension, then only the y-component of the field transmits, and the x-component is killed. These two scenarios can be represented with the following Jones matrices:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{(polarizer with transmission along x-axis)} \tag{4.19}$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{(polarizer with transmission along y-axis)} \tag{4.20}$$

These matrices operate on any Jones vector representing the polarization of incident light. The result gives the Jones vector for the light exiting the polarizer. As an example, consider a horizontally polarized plane wave traversing a polarizer with its transmission axis oriented also horizontally (x-dimension):

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{(horizontal polarizer on horizontally polarized field)} \tag{4.21}$$

As expected, the polarization state is unaffected by the polarizer (ignoring small surface reflections).

Now consider vertically polarized light traversing the same horizontal polarizer. In this case, we have :

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{(horizontal polarizer on vertical linear polarization)} \tag{4.22}$$

As expected, the polarizer extinguishes the light. When a horizontally oriented polarizer operates on light with an arbitrary Jones vector (4.11), we have

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} = \begin{bmatrix} A \\ 0 \end{bmatrix} \qquad \text{(horizontal polarizer on arbitrary polarization)} \tag{4.23}$$

Only the horizontal component of polarization is transmitted through the polarizer.

4.4 Jones Matrix for Polarizers at Arbitrary Angles

While students will readily agree that the matrices given in (4.19) and (4.20) can be used to get the right result for light traversing a horizontal or a vertical polarizer, the real advantage of the matrix formulation has yet to be demonstrated. The usefulness of the formalism becomes clear as we consider the problem of passing a plane wave with arbitrary polarization through a polarizer with its transmission axis aligned at an arbitrary angle θ with the x-axis.

We will analyze this problem in a general context so that we can take advantage of present work when we discuss wave plates in the next section. To help keep things on a more conceptual level, let us revert back to (4.4). We will make the connection with Jones calculus at a later point. The electric field of our plane wave is

$$\mathbf{E}\left(z,t\right) = E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}} \tag{4.24}$$

where

$$\begin{aligned} E_x &\equiv E_{0x}e^{i(kz-\omega t)} \\ E_y &\equiv E_{0y}e^{i(kz-\omega t)} \end{aligned} \tag{4.25}$$

In the upcoming discussion, let the transmission axis of the polarizer be called axis 1 and the absorption axis of the polarizer be called axis 2 (orthogonal to axis 1) as depicted in Fig. 4.4. Axis 1 is oriented at an angle θ from the x-axis. We need to write the electric field components in terms of the new basis specified by the unit vectors $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ as shown in Fig. 4.5. These new unit vectors are connected to the original ones via

$$\hat{\mathbf{x}} = \cos\theta\hat{\mathbf{e}}_1 - \sin\theta\hat{\mathbf{e}}_2 \tag{4.26}$$

and

$$\hat{\mathbf{y}} = \sin\theta\hat{\mathbf{e}}_1 + \cos\theta\hat{\mathbf{e}}_2 \tag{4.27}$$

By direct substitution of (4.26) and (4.27) into (4.24), the electric field can be written as

$$\mathbf{E}\left(z,t\right) = E_1\hat{\mathbf{e}}_1 + E_2\hat{\mathbf{e}}_2 \tag{4.28}$$

where

$$\begin{aligned} E_1 &\equiv E_x\cos\theta + E_y\sin\theta \\ E_2 &\equiv -E_x\sin\theta + E_y\cos\theta \end{aligned} \tag{4.29}$$

At this point, we can introduce the effect of the polarizer on the field: E_1 is transmitted unaffected, and E_2 is killed. Let us multiply E_2 by a parameter ξ to signify the effect of the device. In the case of the polarizer, ξ is zero, but in the next section we will consider other values for ξ. After traversing the polarizer, the field becomes

$$\mathbf{E}_{\text{after}}\left(z,t\right) = E_1\hat{\mathbf{e}}_1 + \xi E_2\hat{\mathbf{e}}_2 \tag{4.30}$$

This completes the job since we now have the field after the polarizer. However, it would be nice to rewrite it in terms of the original x–y basis. By inverting (4.26) and (4.27), or by inspection of Fig. 4.5, if preferred, we see that

$$\hat{\mathbf{e}}_1 = \cos\theta\hat{\mathbf{x}} + \sin\theta\hat{\mathbf{y}} \tag{4.31}$$

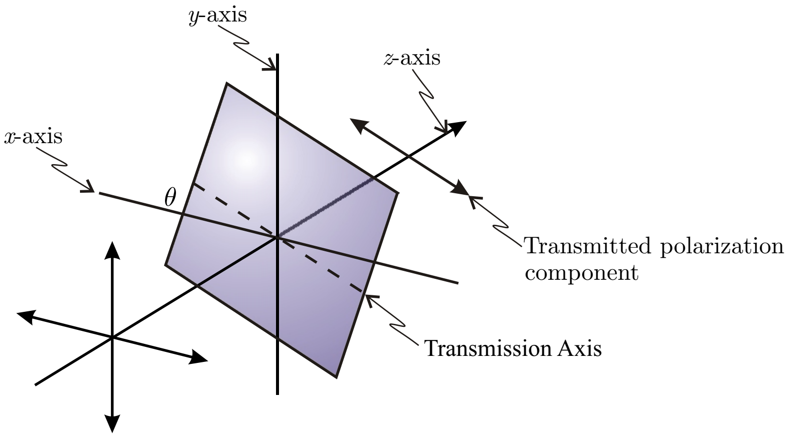


Figure 4.4 Polarizer oriented with transmission axis at angle θ from x-axis.

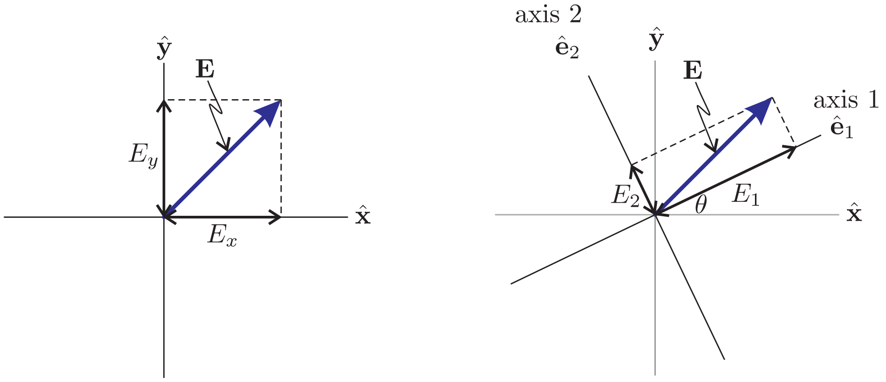


Figure 4.5 Electric field components written in either the $\hat{\mathbf{x}}$–$\hat{\mathbf{y}}$ basis or the $\hat{\mathbf{e}}_1$–$\hat{\mathbf{e}}_2$ basis.

and

$$\hat{\mathbf{e}}_2 = -\sin\theta\hat{\mathbf{x}} + \cos\theta\hat{\mathbf{y}} \tag{4.32}$$

Substitution of these relationships into (4.30) together with the definitions (4.29) for E_1 and E_2 yields

$$\begin{aligned}\mathbf{E}_{\text{after}}(z,t) &= (E_x\cos\theta + E_y\sin\theta)(\cos\theta\hat{\mathbf{x}} + \sin\theta\hat{\mathbf{y}}) \\ &\quad + \xi(-E_x\sin\theta + E_y\cos\theta)(-\sin\theta\hat{\mathbf{x}} + \cos\theta\hat{\mathbf{y}}) \\ &= \left[E_x\left(\cos^2\theta + \xi\sin^2\theta\right) + E_y(\sin\theta\cos\theta - \xi\sin\theta\cos\theta)\right]\hat{\mathbf{x}} \\ &\quad + \left[E_x(\sin\theta\cos\theta - \xi\sin\theta\cos\theta) + E_y\left(\sin^2\theta + \xi\cos^2\theta\right)\right]\hat{\mathbf{y}}\end{aligned} \tag{4.33}$$

Notice that if $\xi = 1$ (i.e. no polarizer), then we get back exactly what we started with (i.e. (4.33) reduces to (4.24)). There remains only to recognize that (4.33) is a linear mixture of E_x and E_y, used to express $\mathbf{E}_{\text{after}}(z,t)$. This type of linear mixture can be represented with matrix algebra. If we represent $\mathbf{E}_{\text{after}}(z,t)$ as a two dimensional column vector with its x-component in the top and its y-component in the bottom (like a Jones vector), then we can rewrite (4.33) as

$$\mathbf{E}_{\text{after}}(z,t) = \begin{bmatrix}\cos^2\theta + \xi\sin^2\theta & \sin\theta\cos\theta - \xi\sin\theta\cos\theta \\ \sin\theta\cos\theta - \xi\sin\theta\cos\theta & \sin^2\theta + \xi\cos^2\theta\end{bmatrix}\begin{bmatrix}E_x \\ E_y\end{bmatrix} \tag{4.34}$$

The matrix here is a Jones matrix, appropriate for operating on Jones vectors (although the vector here is not a properly normalized Jones vector). We used the full representation of the electric field to make things easier to visualize, but we could have done the derivation using matrix and vector notation. We are now ready to write down the Jones matrix for a polarizer (with $\xi = 0$):

$$\begin{bmatrix}\cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta\end{bmatrix} \quad \text{(polarizer with transmission axis at angle } \theta) \tag{4.35}$$

Notice that when $\theta = 0$ this matrix reduces to that of a horizontal polarizer (4.19), and when$\theta = \pi/2$, it reduces to that of a vertical polarizer (4.20).

To the extent that part of the light is absorbed by the polarizer, the Jones vector of the exiting wave is no longer normalized to magnitude one. The Jones vector dotted with its complex conjugate gives the factor by which the *intensity* of the light decreases. In accordance with (4.10), the intensity of the exiting light is

$$\begin{aligned}I &= \frac{1}{2}nc\epsilon_0\left|E_{\text{eff}}\right|^2\begin{bmatrix}A' & B'e^{i\delta'}\end{bmatrix}^*\begin{bmatrix}A' \\ B'e^{i\delta'}\end{bmatrix} \\ &= \frac{1}{2}nc\epsilon_0\left|E_{\text{eff}}\right|^2\left(\left|A'\right|^2 + \left|B'\right|^2\right)\end{aligned} \tag{4.36}$$

where $\begin{bmatrix}A' \\ B'e^{i\delta'}\end{bmatrix}$ represents the Jones vector that emerges from the polarizer (or some other devices), and $\begin{bmatrix}A' & B'e^{i\delta'}\end{bmatrix}^*$ is the complex conjugate, or rather the Hermitian conjugate, written in a format conducive to vector multiplication resulting in a scalar.

The intensity is attenuated by the factor $|A'|^2 + |B'|^2$. Recall that E_{eff} represents the effective strength of the field before it enters the polarizer (or other device), so that the initial Jones vector is normalized to one (see (4.10)). By convention we normally remove an overall phase factor from the Jones vector so that A' is real and non-negative, and we choose δ' so that B' is real and non-negative. However, if we don't bother doing this, the absolute value signs on A' and B' in (4.36) ensure that we get the correct value for intensity.

A product of Jones matrices can represent a sequence of polarizers (with varying orientations). The matrices operate on the Jones vector in the order that the light encounters the devices. Therefore, the matrix for the first device is written on the *right*, and so on until the last device encountered, which is written on the *left*, farthest from the Jones vector.

4.5 Jones Matrices for Wave Plates

The other device for influencing polarization that we will consider is called a wave plate (or retarder). A wave plate is made from a non-isotropic material such as a crystal with low symmetry. Such materials have different indices of refraction, depending on the orientation of the electric field polarization. A wave plate has the appearance of a thin window through which the light passes. However, it has a fast and a slow axis, which are $\pi/2$ (90°) apart in the plane of the window. If the light is polarized along the fast axis, it experiences an index of refraction n_{fast}. This index is less than an index n_{slow} that light experiences when polarized along the orthogonal (slow) axis.

When a plane wave passes through a wave plate, the component of the electric field oriented along the fast axis travels faster than its orthogonal counterpart. The speed of the fast wave component is $v_{\text{fast}} = c/n_{\text{fast}}$ while the speed of the other component is $v_{\text{slow}} = c/n_{\text{slow}}$. The fast component gets ahead, and this introduces a relative phase between the two polarization components.

By adjusting the thickness of the wave plate, we can introduce any desired phase difference between the two components. From (2.24) and (2.26), we have for the **k**-vectors within the wave plate (associated with the two electric field components)

$$k_{\text{slow}} = \frac{2\pi n_{\text{slow}}}{\lambda_{\text{vac}}} \tag{4.37}$$

and

$$k_{\text{fast}} = \frac{2\pi n_{\text{fast}}}{\lambda_{\text{vac}}} \tag{4.38}$$

As light passes through a wave plate of thickness d, the phase difference in (4.2) that accumulates between the fast and the slow polarization components is

$$k_{\text{slow}}d - k_{\text{fast}}d = \frac{2\pi d}{\lambda_{\text{vac}}}\left(n_{\text{slow}} - n_{\text{fast}}\right) \tag{4.39}$$

The most common types of wave plates are the quarter-wave plate and the half-wave plate. The quarter-wave plate introduces a phase difference between the two polarization components equal to

$$k_{\text{slow}}d - k_{\text{fast}}d = \pi/2 + 2\pi m \qquad \text{(quarter-wave plate)} \tag{4.40}$$

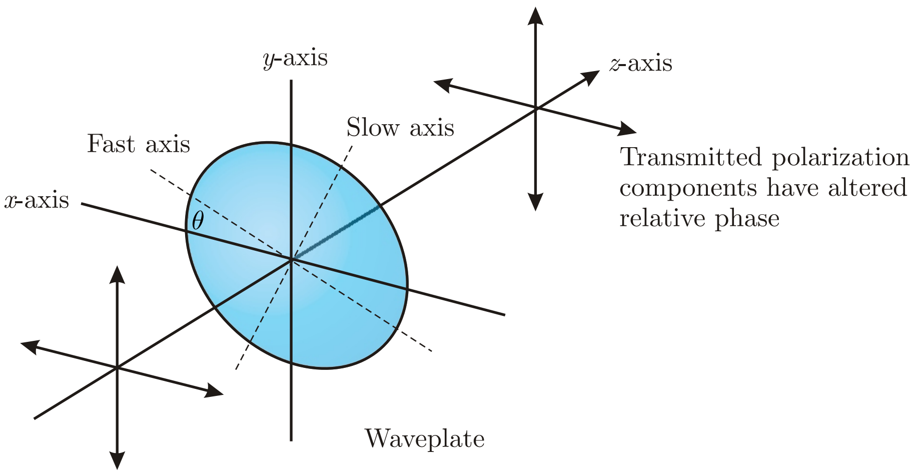

y-axis
z-axis
Slow axis
Fast axis
x-axis
θ
Transmitted polarization components have altered relative phase
Waveplate

Figure 4.6 Wave plate interacting with a plane wave.

where m is an integer. This means that the polarization component along the slow axis is delayed spatially by one quarter of a wavelength (or five quarters, etc.). The half-wave plate introduces a phase delay between the two polarization components equal to

$$k_{\text{slow}}d - k_{\text{fast}}d = \pi + 2\pi m \qquad \text{(half-wave plate)} \tag{4.41}$$

where m is an integer. This means that the polarization component along the slow axis is delayed spatially by half a wavelength (or three halves, etc.).

The derivation of the Jones matrix for the two wave plates is essentially the same as the derivation for the polarizer in the previous section. Let axis 1 correspond to the fast axis, and let axis 2 correspond to the slow axis. We proceed as before. However, instead of setting ξ equal to zero in (4.34), we must choose values for ξ appropriate for each wave plate. Since nothing is absorbed, ξ should have a magnitude equal to one. The important feature is the phase of ξ. As seen in (4.39), the field component along the slow axis accumulates excess phase relative to the component along the fast axis, and we let ξ account for this. In the case of the quarter-wave plate, the appropriate factor from (4.40) is

$$\xi = e^{i\pi/2} = i \qquad \text{(quarter-wave plate)} \tag{4.42}$$

For the polarization component along the slow axis, the term $-i\omega t$ in (4.2) is able to counteract this added phase only at a later time t. Thus, there is a *relative delay* for the light emerging with polarization along the slow axis. (We are not concerned with the overall delay of both polarization components relative to travel through vacuum. What concerns us is the difference between the two components.) For the half-wave plate, the appropriate factor is

$$\xi = e^{i\pi} = -1 \qquad \text{(half-wave plate)} \tag{4.43}$$

We can now write the Jones matrices (4.34) for the quarter-wave and half-wave plates:

$$\begin{bmatrix} \cos^2\theta + i\sin^2\theta & \sin\theta\cos\theta - i\sin\theta\cos\theta \\ \sin\theta\cos\theta - i\sin\theta\cos\theta & \sin^2\theta + i\cos^2\theta \end{bmatrix} \quad \text{(quarter-wave plate)} \qquad (4.44)$$

$$\begin{bmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \sin^2\theta - \cos^2\theta \end{bmatrix} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \quad \text{(half-wave plate)} \qquad (4.45)$$

Again, θ refers to the angle that the fast axis makes with respect to the x-axis.

These two matrices are especially interesting at $\theta = 45°$, where the Jones matrix for the quarter-wave plate reduces to

$$\frac{e^{i\pi/4}}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \quad \text{(quarter-wave plate, fast axis at } \theta = 45°\text{)} \qquad (4.46)$$

The factor $e^{i\pi/4}$ in front is not important since it merely accompanies the overall phase of the beam, which can be adjusted arbitrarily by moving the light source forwards or backwards through a fraction of a wavelength. The Jones matrix for the half-wave plate reduces to

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{(half-wave plate, fast axis at } \theta = 45°\text{)} \qquad (4.47)$$

As an example, consider the effect of the two wave plates (oriented at $\theta = 45°$) operating on horizontally polarized light. For the quarter-wave plate, we get

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix} \qquad (4.48)$$

Notice that the quarter-wave plate (properly oriented) turns linearly polarized light into right-circularly polarized light (see tabel 4.1). The half-wave plate operating on horizontally polarized light gives

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad (4.49)$$

The half-wave plate (when properly oriented) transforms horizontally polarized light into vertically polarized light.

4.6 Polarization Effects of Reflection and Transmission

When light encounters a material interface, the amount of reflected and transmitted light depends on the polarization. The Fresnel coefficients (3.18)–(3.21) dictate how much of each polarization is reflected and how much is transmitted. In addition, the Fresnel coefficients keep track of phases intrinsic in the reflection phenomenon. To the extent that the s and p components of the field behave differently, the overall polarization state is altered. For example, a linearly-polarized field upon reflection can become elliptically polarized (see L 4.9). Even when a wave reflects at normal incidence so that the s and p components are indistinguishable, right-circular polarized light becomes left-circular polarized. This is the same effect that causes a right-handed person to appear left-handed when viewed in a mirror.

We can use Jones calculus to keep track of how reflection and transmission influences polarization. However, before proceeding, we emphasize that in this context we do not strictly adhere to the coordinate system depicted in Fig. 3.1. (Please refer to Fig. 3.1 right now.) For purposes of examining polarization, we consider each plane wave as though traveling in its own z-direction, regardless of the incident angle in the figure. This loose manner of defining coordinate systems has a great advantage. The individual x and y dimensions for each of the three separate plane waves are each aligned parallel to their respective s and p field component. Let us adopt the convention that p-polarized light in all cases is associated with the x-dimension (horizontal). The s-polarized component then lies along the y-dimension (vertical).

We are now in a position to see why there is a handedness inversion upon reflection from a mirror. While referring to Fig. 3.1, notice that for the incident light, the s-component of the field crossed (vector cross product) into the p-component yields that beam's propagation direction. However, for the reflected light, the s-component crossed into the p-component points opposite to that beam's propagation direction.

The Jones matrix corresponding to reflection from a surface is simply

$$\begin{bmatrix} -r_p & 0 \\ 0 & r_s \end{bmatrix} \qquad \text{(Jones matrix for reflection)} \tag{4.50}$$

By convention, we place the minus sign on the coefficient r_p to take care of handedness inversion (the effect that 'moves' your watch from your left wrist to the right wrist when looking in a mirror). We could alternately have put the minus sign on r_s; the important point is that the two polarizations acquire a relative phase differential of π when the propagation direction flips. This effect changes right-hand polarized light into left-hand polarized light. The Fresnel coefficients specify the ratios of the exiting fields to the incident ones. When (4.50) operates on an arbitrary Jones vector such as (4.11), $-r_p$ multiplies the horizontal component of the field, and r_s multiplies the vertical component of the field. In the case of reflection from an absorbing surface such as a metal, the phases of the two polarization components can be very different (see P 4.11). Thus, linearly polarized light containing both s- and p-components in general becomes elliptically polarized when reflected from a metal surface. When light undergoes total internal reflection, again the phases of the s- and p-components can be very different, thus enabling the conversion of linearly polarized light into elliptically polarized light (see P 4.12).

Transmission through a material interface can also influence the polarization of the field. However, there is no handedness inversion, since the light continues on in a forward sense. Nevertheless, the relative amplitudes (and phases if materials are absorbing) of the field components are modified by the Fresnel transmission coefficients. The Jones matrix for this effect is

$$\begin{bmatrix} t_p & 0 \\ 0 & t_s \end{bmatrix} \qquad \text{(Jones matrix for transmission)} \tag{4.51}$$

If a beam of light encounters a series of mirrors, the final polarization is determined by multiplying the sequence of appropriate Jones matrices (4.50) onto the initial polarization. This procedure is straightforward if the normals to all of the mirrors lie in a single plane (say parallel to the surface of an optical bench). However, if the beam path deviates from this plane (due to vertical tilt on the mirrors), then we must reorient our coordinate

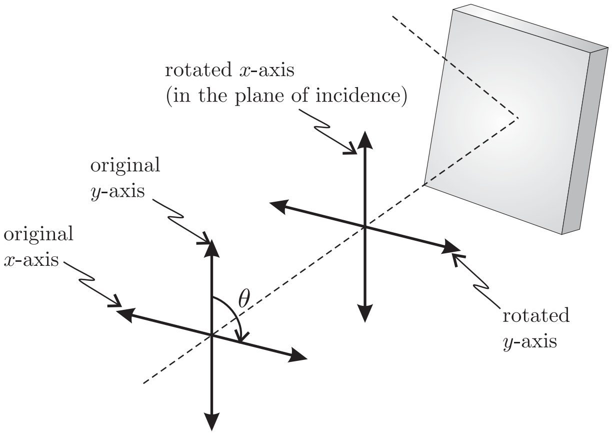


Figure 4.7 When light is reflected out of an optical system's plane of incidence a rotation matrix must be applied so that the rotated x-axis is in the new plane of incidence (i.e. so that p-polarized light remains associated with the x-component of a Jones vector).

system before each mirror to have a new "horizontal" (p-polarized dimension) and the new "vertical" (s-polarized dimension). We have already examined the rotation of a coordinate system through an angle θ in (4.29). This rotation can be accomplished by multiplying the following matrix onto the incident Jones vector:

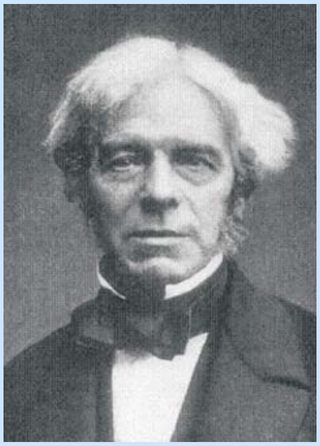
$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \qquad \text{(rotation of coordinates through an angle } \theta\text{)} \tag{4.52}$$

This is a rotation about the z-axis, and the angle of rotation θ is chosen such that the rotated x-axis lies in the plane of incidence for the mirror. When such a reorientation of coordinates is necessary, the two orthogonal field components in the initial coordinate system are stirred together to form the field components in the new system. This does not change the fundamental characteristics of the polarization, just their representation.

4.7 Ellipsometry

In this final section we mention that measuring the polarization of light reflected from a surface can yield information regarding the optical constants of that surface (i.e. n and κ). As done in L 4.9, it is possible to characterize the polarization of a beam of light using a quarter-wave plate and a polarizer. However, we often want to know n and κ at a range of frequencies, and this would require a different quarter-wave plate thickness d for each wavelength used (see (4.40)). Therefore, many commercial ellipsometers do not try to extract the helicity of the light, but only the ellipticity. In this case only polarizers are used, which can be made to work over a wide range of wavelengths.

Inasmuch as most commercial ellipsometers do not determine directly the helicity of the reflected light, the measurement is usually made for a variety of different incident angles on the sample. This adds enough redundancy that n and κ can be pinned down (allowing a



Michael Faraday
(1791–1867, English)

Faraday was one of the greatest experimental physicist in history. He is perhaps best known for his work that established the law of induction (i.e. changing magnetic fields produce electric fields). He also discovered that magnetic fields can interact with light. When a magnetic field is oriented along the direction of travel for light in a dielectric, the polarization of the light will rotate. This effect is used to build optical isolators, which prevent light from reflecting back into an optical system.

computer to take care of the busy work). If many different incident angles are measured at many different wavelengths, it is possible to extract detailed information about the optical constants and the thicknesses of possibly many layers of materials influencing the reflection. (We will learn to deal with multilayer coatings in chapter 6.)

Commercial ellipsometers typically employ two polarizers, one before and one after the sample, where s and p-polarized reflections take place. The first polarizer ensures that linearly polarized light arrives at the test surface (polarized at angle α to give both s and p-components). The Jones matrix for the test surface reflection is given by (4.50), and the Jones matrix for the analyzing polarizer oriented at angle θ is given by (4.35). The Jones vector for the light arriving at the detector is then

$$\begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{bmatrix} \begin{bmatrix} -r_p & 0 \\ 0 & r_s \end{bmatrix} \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} = \begin{bmatrix} -r_p\cos\alpha\cos^2\theta + r_s\sin\alpha\sin\theta\cos\theta \\ -r_p\cos\alpha\sin\theta\cos\theta + r_s\sin\alpha\sin^2\theta \end{bmatrix} \tag{4.53}$$

In an ellipsometer, the angle θ of the analyzing polarizer often rotates at a high speed, and the time dependence of the light reaching a detector is analyzed and correlated with the polarizer orientation. From the measurement of the intensity where θ and α are continuously varied, it is possible to extract the values of n and κ (with the aid of a computer!).

Appendix 4.A Partially Polarized Light

In this appendix, we outline an approach for dealing with partially polarized light, which is a mixture of polarized and unpolarized light. Most natural light such as sunshine is unpolarized. The transverse electric field direction in natural light varies rapidly (and quasi randomly). Such variations imply the superposition of multiple frequencies rather as opposed to the single frequency assumed in the formulation of Jones calculus earlier in this chapter. Unpolarized light can become partially polarized when it, for example, reflects from a surface at oblique incidence, since s and p components of the polarization might reflect with differing strength.

Stokes vectors are used to keep track of the partial polarization (and attenuation) of a light beam as the light progresses through an optical system. In contrast, Jones vectors only deal with pure polarization states. Partially polarized light is a mixture or polarized and unpolarized light. In fact, a beam of light can always be considered as an intensity sum of completely unpolarized light and perfectly polarized light:

$$I = I_{Pol} + I_{Un} \tag{4.54}$$

It is assumed that both types of light propagate in the same direction.

The main characteristic of unpolarized light is that it cannot be extinguished by a single polarizer (or combination of a wave plate and polarizer). Moreover, the transmission of unpolarized light through an ideal polarizer is always 50%. On the other hand, polarized light (be it linearly, circularly, or elliptically polarized) can always be represented by a Jones vector, and it is always possible to extinguish polarized light with a combination of a wave plate and a single polarizer.

We may introduce the *degree of polarization* as the fraction of the intensity that is in a definite polarization state:

$$P \equiv \frac{I_{Pol}}{I_{Pol} + I_{Un}} \tag{4.55}$$

The degree of polarization takes on values between zero and one. Thus, if the light is completely unpolarized (such that $I_{Pol} = 0$), then the degree of polarization is zero. On the other hand, if the beam is fully polarized (such that $I_{Un} = 0$), then the degree of polarization is one.

A Stokes vector, which characterizes a partially polarized beam, is a column vector written as

$$\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

The parameter

$$S_0 \equiv \frac{I}{I_{in}} \tag{4.56}$$

is a comparison of the beam's intensity (or power) with a benchmark intensity, I_{In}, measured before the beam enters an optical system under consideration. I represents the intensity at the point of investigation, where one wishes to characterize the beam. Thus, S_0 is normalized such that a value of one represents the input intensity. After the light goes through a polarizing system, S_0 can drop to values less than one, to account for attenuation of light by polarizers in the system. (Alternatively, S_0 could grow in the atypical case of amplification.)

The next parameter, S_1, describes how much the light looks either horizontally or vertically polarized, and it is defined as

$$S_1 \equiv \frac{2I_{Hor}}{I_{in}} - S_0 \tag{4.57}$$

Here, I_{Hor} represents the amount of light detected if an ideal linear polarizer is placed with its axis aligned horizontally directly in front of the detector (inserted where the light is characterized). S_1 ranges between negative one and one, taking on its extremes when the light is linearly polarized either horizontally or vertically, respectively. If the light has been attenuated, it may still be perfectly horizontally polarized even if S_1 has a magnitude less than one. (For convenience, one may wish to renormalize the beam, taking I_in to be the intensity at the point of investigation, or one can simply examine S_1/S_0, which is guaranteed to a number ranging between negative one and one.)

The parameter S_2 describes how much the light looks linearly polarized along the diagonals. It is given by

$$S_2 \equiv \frac{2I_{45^\circ}}{I_{in}} - S_0 \tag{4.58}$$

Similar to the previous case, I_{45° represents the amount of light detected if an ideal linear polarizer is placed with its axis at 45° directly in front of the detector (inserted where the light is characterized). As before, S_2 ranges between negative and one, taking on extremes when the light is linearly polarized either at 45° or 135°.

Finally, S_3 characterizes the extent to which the beam is either right or left circularly polarized:

$$S_3 \equiv \frac{2I_{R-cir}}{I_{in}} - S_0 \tag{4.59}$$

Here, I_{R-cir} represents the amount of light detected if an ideal right-circular polarizer is placed directly in front of the detector. A right-circular polarizer is one that passes right-handed polarized light, but blocks left handed polarized light. One way to construct such a polarizer is a half wave plate followed by a linear polarizer with the transmission axis aligned 45° from the wave-plate fast axis (see P 4.13). Again, this parameter ranges between negative one and one, taking on the extremes for right and left circular polarization, respectively.

Importantly, if any of the parameters S_1, S_2, or S_3 take on their extreme values (i.e., a magnitude equal to S_0), the other two parameters necessarily equal zero. As an example, if a beam is linearly horizontally polarized with $I = I_{in}$, then we have $I_{Hor} = I_{in}$, $I_{45^\circ} = I_{in}/2$, and $I_{R-cir} = I_{in}/2$. This yields $S_0 = 1$, $S_1 = 1$, $S_2 = 0$, and $S_3 = 0$. As a second example, suppose that the light has been attenuated to $I = I_{in}/3$ but is purely left circularly polarized. Then we have $I_{Hor} = I_{in}/6$, $I_{45^\circ} = I_{in}/6$, and $I_{R-cir} = 0$. Whereas the Stokes parameters are $S_0 = 1/3$, $S_1 = 0$, $S_2 = 0$, and $S_3 = -1/3$.

Another interesting case is completely unpolarized light, which transmits 50% through any of the polarizers discussed above. In this case, $I_{Hor} = I_{45^\circ} = I_{R-cir} = I/2$ and $S_1 = S_2 = S_3 = 0$.

Example 4.1

Find the Stokes parameters for perfectly polarized light, represented by an arbitrary Jones vector

$$\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix}$$

where A, B, and δ are all real. (Note that depending on the values A, B, and δ, the polarization can follow any ellipse.)

Solution: The intensity of this polarized beam is $I_{Pol} = A^2 + B^2$, according to Eq. (4.36), where we absorb the factor $\frac{1}{2}\epsilon_0 c \left|E_{eff}\right|^2$ into A and B for convenience. The Jones vector for the light that passes through a horizontal polarizer is

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} = \begin{bmatrix} A \\ 0 \end{bmatrix}$$

which gives a measured intensity of $I_{Hor} = A^2$. Similarly, the Jones vector when the beam is passed through a polarizer oriented at 45° is

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} A + Be^{i\delta} \\ A + Be^{i\delta} \end{bmatrix}$$

leading to an intensity of

$$I_{45^\circ} = \frac{A^2 + B^2 + 2AB\cos\delta}{2}$$

Finally, the Jones vector for light passing through a right-circular polarizer (see P 4.13) is

$$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix} \begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} A + iBe^{i\delta} \\ -iA + Be^{i\delta} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} (A - B\sin\delta) + iB\cos\delta \\ B\cos\delta + i\,(B\sin\delta - A) \end{bmatrix}$$

giving an intensity of

$$I_{R-cir} = \frac{A^2 - 2AB\sin\delta + B^2\sin^2\delta + B^2\cos^2\delta}{2} = \frac{A^2 + B^2 - 2AB\sin\delta}{2}$$

Thus, the Stokes parameters become

$$S_0 = \frac{A^2 + B^2}{I_{in}}$$

$$S_1 = \frac{2A^2}{I_{in}} - \frac{A^2 + B^2}{I_{in}} = \frac{A^2 - B^2}{I_{in}}$$

$$S_2 = \frac{A^2 + B^2 + 2AB\cos\delta}{I_{in}} - \frac{A^2 + B^2}{I_{in}} = \frac{2AB\cos\delta}{I_{in}}$$

$$S_3 = \frac{A^2 + B^2 - 2AB\sin\delta}{I_{in}} - \frac{A^2 + B^2}{I_{in}} = -\frac{2AB\sin\delta}{I_{in}}$$

It is clear from the linear dependence of S_0, S_1, S_2, and S_3 on intensity (see Eqs. (4.56)–(4.59)) that the overall Stokes vector may be regarded as the sum of the individual Stokes vectors for polarized and unpolarized light. That is, we may write $S_i = S_i^{(Pol)} + S_i^{(Un)}$, $i = 0, 1, 2, 3$.

This is certainly true for

$$S_0 = \frac{I}{I_{in}} = \frac{I_{Pol} + I_{Un}}{I_{in}} \tag{4.60}$$

and in the other cases the unpolarized portion does not contribute to the Stokes parameters, since an equal contribution from the unpolarized light appears in both terms in each of Eqs. (4.57)-(4.59) and therefore cancels out.

A completely general form of the Stokes vector may then be written as (see Example 4.1)

$$\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \frac{1}{I_{in}} \begin{bmatrix} I_{Pol} + I_{Un} \\ A^2 - B^2 \\ 2AB\cos\delta \\ 2AB\sin\delta \end{bmatrix} \tag{4.61}$$

where the Jones vector

$$\begin{pmatrix} A \\ Be^{i\delta} \end{pmatrix}$$

describes the polarized portion of the light, which has intensity

$$I_{Pol} = A^2 + B^2 \tag{4.62}$$

We would like to express the degree of polarization in terms of the Stokes parameters. We first note that the quantity $\sqrt{S_1^2 + S_2^2 + S_3^2}$ can be expressed as

$$\begin{aligned} \sqrt{S_1^2 + S_2^2 + S_3^2} &= \sqrt{\left(\frac{A^2 - B^2}{I_{in}}\right)^2 + \left(\frac{2AB\cos\delta}{I_{in}}\right)^2 + \left(\frac{2AB\sin\delta}{I_{in}}\right)^2} \\ &= \frac{1}{I_{in}}\sqrt{\left(A^2 - B^2\right)^2 + 4A^2B^2\left(\cos^2\delta + \sin^2\delta\right)} \\ &= \frac{A^2 + B^2}{I_{in}} \\ &= \frac{I_{Pol}}{I_{in}} \end{aligned} \tag{4.63}$$

Substituting (4.60) and (4.63) into the expression for the degree of polarization (4.55) yields

$$P \equiv \frac{1}{S_0}\sqrt{S_1^2 + S_2^2 + S_3^2} \tag{4.64}$$

If the light is polarized such that it perfectly transmits through or is perfectly extinguished by one of the three test polarizers associated with S_1, S_2, or S_3, then the degree of polarization will be unity. Obviously, it is possible to have pure polarization states that are not aligned with the axes of any one of these test polarizers. In this situation, the degree of polarization is still one, although the values S_1, S_2, and S_3 may all three contribute to (4.62).

Finally, it is possible to represent polarizing devices as matrices that operate on the Stokes vectors in much the same way that Jones operate on Jones vectors. Since Stokes vectors are four-dimensional, the matrices used are four-by-four. These are known as Mueller matrices.

Example 4.2

Determine the Mueller matrix that represents a linear polarizer with transmission axis at arbitrary angle θ.

Solution: We know that the 50% of the unpolarized light transmits through the polarizer, ending up with Jones vector

$$\begin{bmatrix} A_1' \\ B_1' \end{bmatrix} = \frac{I_{Un}}{2}\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

(see table 4.1). We also know that the Jones matrix (4.36) acts on the polarized portion of the light, represented by arbitrary Jones vector

$$\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix}$$

This gives a transmitted Jones vector of

$$\begin{aligned}\begin{bmatrix} A_2' \\ B_2' e^{i\delta_2'} \end{bmatrix} &= \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix} \begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} \\ &= \left[A\cos\theta + B\sin\theta e^{i\delta}\right] \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}\end{aligned}$$

One might be tempted to add the two Jones vectors, but this would be wrong, since the two beams are not coherent. As mentioned previously, unpolarized light necessarily contains multiple frequencies, and so the fields from the polarized and unpolarized beam destructively interfere as often as they constructively interfere. In this case, we add intensities rather than fields. That is, we have

$$\begin{aligned}\left|A'\right|^2 = \left|A_1'\right|^2 + \left|A_2'\right|^2 &= \left[\frac{I_{Un}}{2} + A^2\cos^2\theta + B^2\sin^2\theta + 2AB\cos\theta\sin\theta\cos\delta\right]\cos^2\theta \\ &= \left[\frac{I_{Un} + A^2 + B^2}{2} + \left(A^2 - B^2\right)\frac{\cos 2\theta}{2} + 2AB\cos\delta\frac{\sin 2\theta}{2}\right]\cos^2\theta \\ &= \left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\cos^2\theta\end{aligned}$$

Similarly,

$$\left|B'\right|^2 = \left|B_{11}'\right|^2 + \left|B_{22}'\right|^2 = \left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\sin^2\theta$$

This gives

$$\begin{aligned}S_0' &= \left|A'\right|^2 + \left|B'\right|^2 \\ &= \frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2 \\ S_1' &= \left|A'\right|^2 - \left|B'\right|^2 \\ &= \left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\left(\cos^2\theta - \sin^2\theta\right) \\ &= \frac{S_0\cos 2\theta}{2} + \frac{\cos^2 2\theta}{2}S_1 + \frac{\sin 4\theta}{4}S_2\end{aligned}$$

and since $\delta' = 0$ we have

$$\begin{aligned}S_2' &= 2\left|A'\right|\left|B'\right|\cos\delta' \\ &= 2\left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\cos\theta\sin\theta \\ &= \frac{S_0\sin 2\theta}{2} + \frac{\sin 4\theta}{4}S_1 + \frac{\sin^2 2\theta}{2}S_2 \\ S_3' &= 2\left|A'\right|\left|B'\right|\sin\delta' \\ &= 0\end{aligned}$$

These transformations expressed in matrix format become

$$\begin{bmatrix} S_0' \\ S_1' \\ S_2' \\ S_3' \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & \cos 2\theta & \sin 2\theta & 0 \\ \cos 2\theta & \cos^2 2\theta & \frac{1}{2}\sin 4\theta & 0 \\ \sin 2\theta & \frac{1}{2}\sin 4\theta & \sin^2 2\theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

which reveals the Mueller matrix for a linear polarizer.

Exercises

4.2 Jones Vectors for Representing Polarization

P4.1 Show that $\left(A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}\right)\cdot\left(A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}\right)^* = 1$, as defined in connection with (4.5).

P4.2 Prove that if $0 < \delta < \pi$, the helicity is left-handed, and if $\pi < \delta < 2\pi$ the helicity is right-handed.

HINT: Write the relevant real field associated with (4.5)

$$\mathbf{E}(z,t) = |E_{\text{eff}}|\left[\hat{\mathbf{x}}A\cos(kz-\omega t+\phi) + \hat{\mathbf{y}}B\cos(kz-\omega t+\phi+\delta)\right]$$

where ϕ is the phase of E_{eff}. Freeze time at, say, $t = \phi/\omega$. Determine the field at $z = 0$ and at $z = \lambda/4$ (a quarter cycle), say. If $\mathbf{E}(0,t)\times\mathbf{E}(\lambda/4,t)$ points in the direction of $\mathbf{k}$, then the helicity matches that of a wood screw.

P4.3 For the following cases, what is the orientation of the major axis, and what is the ellipticity of the light? Case I: $A = B = 1/\sqrt{2}$; $\delta = 0$ Case II: $A = B = 1/\sqrt{2}$; $\delta = \pi/2$; Case III: $A = B = 1/\sqrt{2}$; $\delta = \pi/4$.

L4.4 Determine how much right-handed circularly polarized light ($\lambda_{\text{vac}} = 633$ nm) is delayed (or advanced) with respect to left-handed circularly polarized light as it goes through approximately 3 cm of Karo syrup (the neck of the bottle). This phenomenon is called *optical activity.* Because of a definite-handedness to the molecules in the syrup, right- and left-handed polarized light experience slightly different refractive indices.

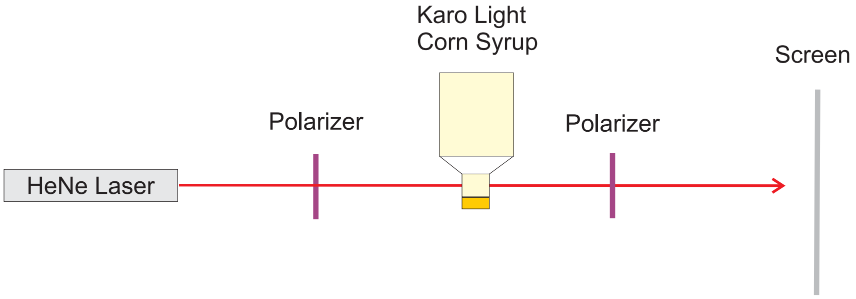


Figure 4.8 Lab schematic for L 4.4

HINT: Linearly polarized light contains equal amounts of right and left circularly polarized light. Consider

$$\frac{1}{2}\begin{bmatrix}1\\ i\end{bmatrix} + \frac{e^{i\phi}}{2}\begin{bmatrix}1\\ -i\end{bmatrix}$$

where ϕ is the phase delay of the right circular polarization. Show that this can be written as

$$e^{i\delta}\begin{bmatrix}\cos\phi/2\\ \sin\phi/2\end{bmatrix}$$

Compare this with

$$\begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix}$$

where α is the angle through which the polarization is rotated, beginning with horizontally polarized light. The overall phase δ is unimportant.

4.3 Jones Matrices

P4.5 (a) Suppose that linearly polarized light is oriented at an angle α with respect to the horizontal axis (x-axis) (see table 4.1). What fraction of the original *intensity* gets through a vertically oriented polarizer?

(b) If the original light is right-circularly polarized, what fraction of the original *intensity* gets through the same polarizer?

4.4 Jones Matrix for Polarizers at Arbitrary Angles

P4.6 Horizontally polarized light ($\alpha = 0$) is sent through two polarizers, the first oriented at $\theta_1 = 45°$ and the second at $\theta_2 = 90°$. What fraction of the original intensity emerges? What is the fraction if the ordering of the polarizers is reversed?

P4.7 (a) Suppose that linearly polarized light is oriented at an angle α with respect to the horizontal or x-axis. What fraction of the original *intensity* emerges from a polarizer oriented with its transmission at angle θ from the x-axis?

Answer: $\cos^2(\theta - \alpha)$; compare with P 4.5.

(b) If the original light is right circularly polarized, what fraction of the original *intensity* emerges from the same polarizer?

P4.8 Derive (4.12), (4.13), and (4.14).

HINT: Analyze the Jones vector just as you would analyze light in the laboratory. Put a polarizer in the beam and observe the intensity of the light as a function of polarizer angle. Compute the intensity via (4.36). Then find the polarizer angle (call it α) that gives a maximum (or a minimum) of intensity. The angle then corresponds to an axis of the ellipse followed by the E-field as it spirals. When taking the arctangent, remember that it is defined only over a range of π. You can add π for another valid result (which corresponds to the second ellipse axis).

4.5 Jones Matrices for Wave Plates

L4.9 Create a source of unknown elliptical polarization by reflecting a linearly polarized laser beam (with both s and p-components) from a metal mirror with a large incident angle (i.e. $\theta_\mathrm{i} \geq 80°$). Use a quarter-wave plate and a polarizer to determine the Jones vector of the reflected beam. Find the ellipticity, the helicity (right or left handed), and the orientation of the major axis.

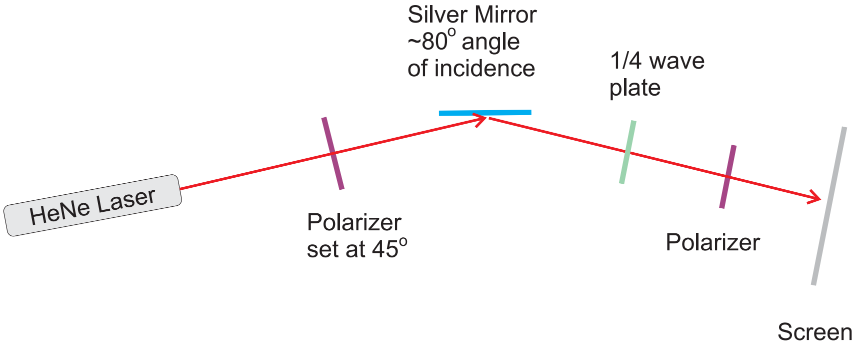


Figure 4.9 Lab schematic for L 4.9

HINT: A polarizer alone can reveal the direction of the major and minor axes and the ellipticity, but it does not reveal the helicity. Use a quarter-wave plate (oriented at a special angle θ) to convert the unknown elliptically polarized light into linearly polarized light. A subsequent polarizer can then extinguish the light, from which you can determine the Jones vector of the light coming through the wave plate. This must equal the original (unknown) Jones vector (4.11) operated on by the wave plate (4.44). As you solve the matrix equation, it is helpful to note that the inverse of (4.44) is its own complex conjugate.

P4.10 What is the minimum thickness (called zero-order thickness) of a quartz plate made to operate as a quarter-wave plate for $\lambda_{\text{vac}} = 500$ nm? The indices of refraction are $n_{\text{fast}} = 1.54424$ and $n_{\text{slow}} = 1.55335$.

4.6 Polarization Effects of Reflection and Transmission

P4.11 Light is linearly polarized at $\alpha = 45°$ with a Jones vector according to table 4.1. The light is reflected from a vertical silver mirror with angle of incidence $\theta_{\text{i}} = 80°$, as described in (P 3.12). Find the Jones vector representation for the polarization of the reflected light. NOTE: The answer may be somewhat different than the result measured in L 4.9. For one thing, we have not considered that a silver mirror inevitably has a thin oxide layer.

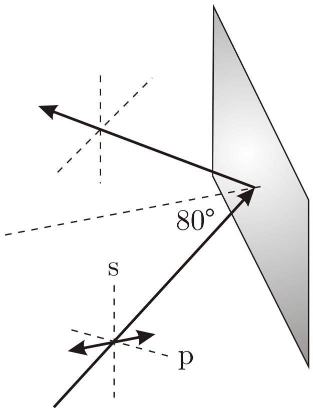


Figure 4.10 Geometry for P 4.11

Answer: $\begin{pmatrix} 0.668 \\ 0.702e^{1.13i} \end{pmatrix}$.

P4.12 Calculate the angle θ to cut the glass in a Fresnel rhomb such that after the two internal reflections there is a phase difference of $\pi/2$ between the two polarization states. The rhomb then acts as a quarter wave plate.

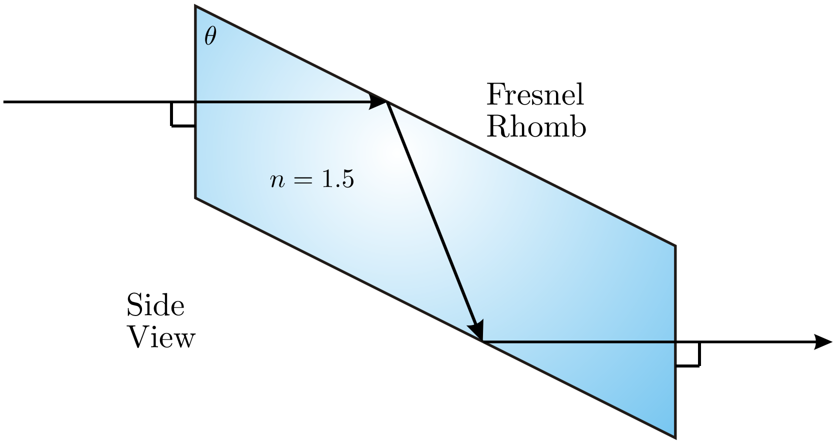


Figure 4.11 Fresnel Rhomb geometry for P 4.12

HINT: You need to find the phase difference between (3.40) and (3.41). Set the difference equal to $\pi/4$ for each bounce. The equation you get does not have a clean analytic solution, but you can plot it to find a numerical solution.

Answer: There are two angles that work: $\theta \cong 50°$ and $\theta \cong 53°$.

4.A Partially Polarized Light

P4.13 (a) Construct the Jones matrix for a right-circular polarizer using a quarter wave plate with fast axis at , followed by a linear polarizer oriented vertically, and finally a quarter wave plate with fast axis at . Answer:

(b) Check that the device leaves right-circularly polarized light unaltered while killing left-circularly polarized light.

P4.14 Derive the Mueller matrix for a half wave plate.

P4.15 Derive the Mueller matrix for a quarter wave plate.

Chapter 5

Light Propagation in Crystals

5.1 Introduction

In a crystal, the connection between $\mathbf{P}$ and $\mathbf{E}$ is more complicated than in an isotropic medium. Fig. 5.1 depicts an electron bound in a crystal lattice. The electron is bound as though by tiny springs, which have different strengths in different dimensions. The lattice can cause directional asymmetries so that the polarization $\mathbf{P}$ of the material in the crystal does not respond necessarily in the same direction as the applied electric field $\mathbf{E}$ (i.e. $\mathbf{P} \neq \epsilon_0 \chi \mathbf{E}$). However, except in the case of extremely intense light, the response of the material is still linear (or proportionate). The linear constitutive relation which connects $\mathbf{P}$ to $\mathbf{E}$ in a crystal can be expressed in its most general form as

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \tag{5.1}$$

The matrix in (5.1) is called a tensor. This relation really contains three different equations:

$$\begin{aligned} P_x &= \epsilon_0 \chi_{xx} E_x + \epsilon_0 \chi_{xy} E_y + \epsilon_0 \chi_{xz} E_z, \\ P_y &= \epsilon_0 \chi_{yx} E_x + \epsilon_0 \chi_{yy} E_y + \epsilon_0 \chi_{yz} E_z, \\ P_z &= \epsilon_0 \chi_{zx} E_x \quad + \epsilon_0 \chi_{zy} E_y + \epsilon_0 \chi_{zz} E_z \end{aligned}$$

As we consider the propagation of light within a crystal, we ignore the possibility of absorption. Hence, we take all χ_{ij} to be real.

The geometrical interpretation of the many coefficients χ_{ij} is clear. In a crystal, if we apply an electric field in the x-direction, the induced polarization can acquire y and z-components in addition to an x-component. This is similar to sliding an object on an inclined plane. As we apply a force in the horizontal direction, there is a vertical component to the acceleration if the object is constrained to slide along the inclined plane.

The tensor in (5.1) must be symmetric (i.e. $\chi_{ij} = \chi_{ji}$). This means that an electric field applied in the x-dimension results in P_y equal to P_x that results when the same electric field is applied instead in the y-dimension. Returning to the object sliding on an inclined plane, if we apply the same force instead to the vertical direction, the resulting horizontal component of acceleration is the same as the vertical component previously.

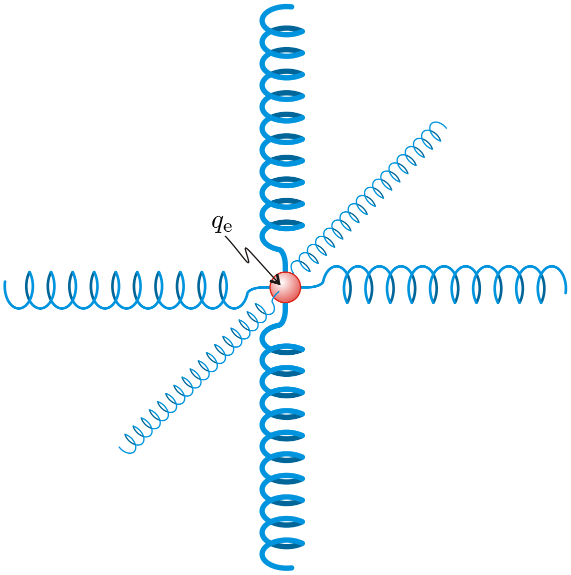


Figure 5.1 Electron bound in a crystal lattice.

Fortunately, (5.1) can be considerably simplified by a judicious choice of coordinate system. In every crystal there exists a coordinate system for which off-diagonal elements of the matrix in (5.1) vanish (see Appendix 5.A). This is true even if the lattice planes in the crystal are not mutually orthogonal (e.g. rhombus, hexagonal, etc.).

We allow the crystal to dictate the orientation of the coordinate system, aligned to the *principal axes* of the crystal for which the off-diagonal elements of (5.1) are zero. Then the constitutive relation simplifies to

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \tag{5.2}$$

or

$$\mathbf{P} = \hat{\mathbf{x}}\epsilon_0\chi_x E_x + \hat{\mathbf{y}}\epsilon_0\chi_y E_y + \hat{\mathbf{z}}\epsilon_0\chi_z E_z \tag{5.3}$$

In section 5.2 we apply Maxwell's equations to a plane wave traveling in a crystal. The analysis leads to *Fresnel's equation*, which connects the components of the **k**-vector with χ_x, χ_y, and χ_z. In section 5.4 we apply Fresnel's equation to a *uniaxial* crystal (e.g. quartz, sapphire) where $\chi_x = \chi_y \neq \chi_z$. In section 5.5 we examine the flow of energy in a uniaxial crystal and show that the Poynting vector and the **k**-vector in general are not parallel. In Appendix 5.B we describe light propagation in a crystal using the method of Christian Huygens (1629-1695) who lived more than a century before Fresnel. Huygens successfully described *birefringence* in crystals using the idea of elliptical wavelets. His method gives the direction of the Poynting vector associated with the *extraordinary* ray in a crystal. It was Huygens who coined the term "extraordinary" since one of the rays in a *birefringent* material appeared not to obey Snell's law. Actually, the **k**-vector always obeys Snell's law, but in a crystal the **k**-vector points in a different direction than the Poynting vector, and

Christiaan Huygens

(1629–1695, Dutch)

Huygens championed the wave theory of light. He was able to explain birefringence in terms of different an indexe of refraction that varied with direction (Newton was also able to explain birefringence with particles by assuming that the crystal sorted light-particles according to their geometric properties.) Huygens made many advancements in clock-making technology and statistical theory.

it is the Poynting vector that delivers the energy seen by an observer.

5.2 Wave Propagation in Non-Isotropic Media

We will search for plane-wave solutions to the wave equation (1.42) in a crystal with $\mathbf{J}_{\text{free}} = 0$. As a trial solution, we consider a plane wave with frequency ω, similar to the plane-wave solution we have studied in isotropic materials. In this case, the fields $\mathbf{E}$, $\mathbf{B}$, and $\mathbf{P}$ are all associated with the same plane wave according to

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ \mathbf{B} &= \mathbf{B}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ \mathbf{P} &= \mathbf{P}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \end{aligned} \tag{5.4}$$

The phase of each wave is included in the amplitudes $\mathbf{E}_0$, $\mathbf{B}_0$, and $\mathbf{P}_0$.

The fields must satisfy Maxwell's equations, two of which (1.35)–(1.36) are

$$\nabla \cdot (\epsilon_0 \mathbf{E} + \mathbf{P}) = 0 \tag{5.5}$$

and

$$\nabla \cdot \mathbf{B} = 0 \tag{5.6}$$

When our trial solutions (5.4) are inserted into these, we find

$$\mathbf{k} \cdot (\epsilon_0 \mathbf{E} + \mathbf{P}) = 0 \tag{5.7}$$

and

$$\mathbf{k} \cdot \mathbf{B} = 0 \tag{5.8}$$

Notice that we have the following peculiarity: From its definition, the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$ (2.51) is perpendicular to both $\mathbf{E}$ and $\mathbf{B}$, and by (5.8) the $\mathbf{k}$-vector is perpendicular to $\mathbf{B}$. However, by (5.7) the $\mathbf{k}$-vector is not perpendicular to $\mathbf{E}$ (since in general $\mathbf{k}\cdot\mathbf{E} \neq 0$ if $\mathbf{P}$ points in a direction other than $\mathbf{E}$). Therefore, $\mathbf{k}$ and $\mathbf{S}$ are not necessarily parallel in a crystal. In other words, the flow of energy and the direction of the wave propagation are not the same.

We start with the wave equation in the form (1.41),

$$\nabla^2\mathbf{E} - \mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} = \mu_0\frac{\partial^2\mathbf{P}}{\partial t^2} + \nabla\left(\nabla\cdot\mathbf{E}\right) \tag{5.9}$$

under the assumption $\mathbf{J}_{\text{free}} = 0$. Upon substitution of our trial solutions (5.4) into this equation, we obtain

$$k^2\mathbf{E} - \omega^2\mu_0\left(\epsilon_0\mathbf{E} + \mathbf{P}\right) = \mathbf{k}\left(\mathbf{k}\cdot\mathbf{E}\right) \tag{5.10}$$

At this point we make use of the constitutive relation (5.3) for crystals. The requirement (5.10) imposed by Maxwell's equations then becomes

$$k^2\mathbf{E} - \omega^2\mu_0\epsilon_0\left[\left(1+\chi_x\right)E_x\hat{\mathbf{x}} + \left(1+\chi_y\right)E_y\hat{\mathbf{y}} + \left(1+\chi_z\right)E_z\hat{\mathbf{z}}\right] = \mathbf{k}\left(\mathbf{k}\cdot\mathbf{E}\right) \tag{5.11}$$

This relation actually contains three equations, one for each dimension. Explicitly, these equations are

$$\left[k^2 - \frac{\omega^2}{c^2}\left(1+\chi_x\right)\right]E_x = k_x\left(\mathbf{k}\cdot\mathbf{E}\right) \tag{5.12}$$

$$\left[k^2 - \frac{\omega^2}{c^2}\left(1+\chi_y\right)\right]E_y = k_y\left(\mathbf{k}\cdot\mathbf{E}\right) \tag{5.13}$$

and

$$\left[k^2 - \frac{\omega^2}{c^2}\left(1+\chi_z\right)\right]E_z = k_z\left(\mathbf{k}\cdot\mathbf{E}\right) \tag{5.14}$$

We have replaced the constants $\mu_0\epsilon_0$ with $1/c^2$ according to (1.44).

5.3 Fresnel's Equation

Equations (5.12)–(5.14) are unwieldy since the electric field components appear in the expressions. This did not cause a problem when we investigated isotropic materials for which the $\mathbf{k}$-vector is perpendicular to $\mathbf{E}$, making the right-hand side of the equations zero. Nevertheless, through a direct procedure, we can eliminate the electric field components from the expressions.

We multiply (5.12)–(5.14) respectively by k_x, k_y, and k_z. We also move the factor in square brackets in each equation to the denominator on the right-hand side. Then if we add the three equations together we get

$$\frac{k_x^2\left(\mathbf{k}\cdot\mathbf{E}\right)}{\left[k^2 - \frac{\omega^2(1+\chi_x)}{c^2}\right]} + \frac{k_y^2\left(\mathbf{k}\cdot\mathbf{E}\right)}{\left[k^2 - \frac{\omega^2(1+\chi_y)}{c^2}\right]} + \frac{k_z^2\left(\mathbf{k}\cdot\mathbf{E}\right)}{\left[k^2 - \frac{\omega^2(1+\chi_z)}{c^2}\right]} = k_xE_x + k_yE_y + k_zE_z = \left(\mathbf{k}\cdot\mathbf{E}\right) \tag{5.15}$$

This nice trick allows us to get rid of the electric field by dividing the equation by $\mathbf{k}\cdot\mathbf{E}$. If we also multiply the equation by ω^2/c^2 we have

$$\frac{k_x^2}{[k^2c^2/\omega^2-(1+\chi_x)]}+\frac{k_y^2}{[k^2c^2/\omega^2-(1+\chi_y)]}+\frac{k_z^2}{[k^2c^2/\omega^2-(1+\chi_z)]}=\frac{\omega^2}{c^2} \tag{5.16}$$

This equation allows us to find a suitable k-vector, given values for ω, χ_x, χ_y, and χ_z. Nevertheless, with only this information the solution to this equation is far from unique. In particular, we must decide on a direction for the wave to travel. We must choose the ratios between k_x, k_y, and k_z. To remind ourselves of this fact, let us introduce a unit vector that points in the direction of the k-vector we wish to find:

$$\mathbf{k}=k_x\hat{\mathbf{x}}+k_y\hat{\mathbf{y}}+k_z\hat{\mathbf{z}}=k\left(u_x\hat{\mathbf{x}}+u_y\hat{\mathbf{y}}+u_z\hat{\mathbf{z}}\right)=k\hat{\mathbf{u}} \tag{5.17}$$

With this unit vector inserted, (5.16) becomes

$$\frac{u_x^2}{[k^2c^2/\omega^2-(1+\chi_x)]}+\frac{u_y^2}{[k^2c^2/\omega^2-(1+\chi_y)]}+\frac{u_z^2}{[k^2c^2/\omega^2-(1+\chi_z)]}=\frac{\omega^2}{k^2c^2} \tag{5.18}$$

We are now ready to introduce the refractive index for non-isotropic materials. As we have seen before, the speed of a wave having the form (5.4) is $v=\omega/k$ (see P 1.10). By definition, the refractive index of a material is the ratio of c to the speed v (see (2.18)). Therefore, the refractive index for the wave is

$$n=\frac{kc}{\omega} \tag{5.19}$$

as we saw in (2.23). Although the above equation looks innocent enough—and we have seen it before—the relationship between k and ω depends on the direction of propagation in the crystal according to

$$\frac{u_x^2}{(n^2-n_x^2)}+\frac{u_y^2}{\left(n^2-n_y^2\right)}+\frac{u_z^2}{(n^2-n_z^2)}=\frac{1}{n^2} \tag{5.20}$$

Motivated by (2.18), we have replaced the susceptibility parameters in (5.18) with three new constants:

$$\begin{aligned} n_x &\equiv \sqrt{1+\chi_x}\\ n_y &\equiv \sqrt{1+\chi_y}\\ n_z &\equiv \sqrt{1+\chi_z} \end{aligned} \tag{5.21}$$

Equation (5.20) is called *Fresnel's equation* (not to be confused with the Fresnel coefficients studied in chapter 3). The relationship contains the yet unknown index n that varies with the direction of the **k**-vector (i.e. the direction of the unit vector $\hat{\mathbf{u}}$).

Actually, for a given **k**-vector there are two possible values for n, one associated with each of two orthogonal field polarizations. As we shall see, in the special cases of the electric field being oriented along the x, y, or z-directions, the refractive index n takes on the values n_x, n_y, or n_z, respectively. When the electric field points in other directions, n takes on values which are mixtures of n_x, n_y, and n_z.

Our next task is to solve (5.20) for n. To this end, (5.20) can be manipulated into the form

$$\begin{aligned}0 = &\left[\left(u_x^2 + u_y^2 + u_z^2\right) - 1\right] n^6 \\ &+ \left[\left(n_x^2 + n_y^2 + n_z^2\right) - u_x^2\left(n_y^2 + n_z^2\right) - u_y^2\left(n_x^2 + n_z^2\right) - u_z^2\left(n_x^2 + n_y^2\right)\right] n^4 \\ &- \left[\left(n_x^2 n_y^2 + n_x^2 n_z^2 + n_y^2 n_z^2\right) - u_x^2 n_y^2 n_z^2 - u_y^2 n_x^2 n_z^2 - u_z^2 n_x^2 n_y^2\right] n^2 + n_x^2 n_y^2 n_z^2\end{aligned} \tag{5.22}$$

The coefficient of n^6 is identically zero since by definition we have $u_x^2 + u_y^2 + u_z^2 = 1$. This leaves a quadratic equation in n^2:

$$An^4 - Bn^2 + C = 0 \tag{5.23}$$

where

$$A \equiv u_x^2 n_x^2 + u_y^2 n_y^2 + u_z^2 n_z^2 \tag{5.24}$$

$$B \equiv u_x^2 n_x^2\left(n_y^2 + n_z^2\right) + u_y^2 n_y^2\left(n_x^2 + n_z^2\right) + u_z^2 n_z^2\left(n_x^2 + n_y^2\right) \tag{5.25}$$

$$C \equiv n_x^2 n_y^2 n_z^2 \tag{5.26}$$

The solutions to (5.23) are

$$n^2 = \frac{B \pm \sqrt{B^2 - 4AC}}{2A} \tag{5.27}$$

Let us review what has been accomplished here. Given values for χ_x, χ_y, and χ_z associated with a frequency ω, one defines the indices n_x, n_y, and n_z, according to (5.21). Next, a direction for the **k**-vector is chosen (i.e. u_x, u_y, and u_z are chosen). Finally, the index for that direction of propagation is found from (5.27). However, (5.27) has two (positive) solutions for n! This should not be alarming since, after all, we are interested in birefringence! The upper and lower signs in (5.27) correspond to two orthogonal electric field polarizations in connection with the direction of $\hat{\mathbf{u}}$. The two polarization components of the wave travel at different speeds, according to the values for n. Therefore, even though the frequency ω is the same for both polarization components, the wavelength for each is different (within the crystal).

5.4 Uniaxial Crystal

A crystal is said to be *uniaxial* when the index of refraction for two of the three dimensions in the crystal is the same. We will study the behavior of uniaxial crystals (as opposed to biaxial) as an example of how to apply Fresnel's equation. In this case we have

$$n_x = n_y = n_\text{o} \tag{5.28}$$

and

$$n_z = n_\text{e} \tag{5.29}$$

where we have chosen the unique axis (called the *optic axis*) to be in the z-direction. The subscripts "o" and "e" stand for *ordinary* and *extraordinary*, so named by Huygens (see appendix 5.B).

For simplicity, consider a wave that propagates in the y–z plane, making an angle ϕ with the z-axis as depicted in Fig. 5.2. This poses no restriction since the x and y dimensions are indistinguishable. We could as easily consider propagation in the x–z plane or any other plane containing the z-axis. For a uniaxial crystal, these all yield the same result. Under our convention that propagation takes place in the y–z plane, the vector components of $\mathbf{k}$ are

$$\begin{aligned} k_x &= 0 \\ k_y &= k \sin\phi \\ k_z &= k\cos\phi \end{aligned} \tag{5.30}$$

or $u_x = 0$, $u_y = \sin\phi$, and $u_z = \cos\phi$.

When these parameters are used to evaluate (5.25)–(5.27) the two indices of refraction become

$$n = n_\mathrm{o} \quad \text{(uniaxial crystal)} \tag{5.31}$$

and

$$n = \frac{n_\mathrm{o} n_\mathrm{e}}{\sqrt{n_\mathrm{o}^2 \sin^2\phi + n_\mathrm{e}^2 \cos^2\phi}} \quad \text{(uniaxial crystal)} \tag{5.32}$$

Example 5.1

Derive (5.31) and (5.32).

Solution:

$$\begin{aligned} A &= n_\mathrm{o}^2 \sin^2\phi + n_\mathrm{e}^2\cos^2\phi \\ B &= n_\mathrm{o}^2\left(n_\mathrm{o}^2 + n_\mathrm{e}^2\right)\sin^2\phi + n_\mathrm{e}^2\left(n_\mathrm{o}^2 + n_\mathrm{o}^2\right)\cos^2\phi = n_\mathrm{o}^2 n_\mathrm{e}^2 + n_\mathrm{o}^4 \sin^2\phi + n_\mathrm{e}^2 n_\mathrm{o}^2\cos^2\phi \\ C &= n_\mathrm{o}^4 n_\mathrm{e}^2 \end{aligned}$$

$$\begin{aligned} B^2 - 4AC &= \left(n_\mathrm{o}^2 n_\mathrm{e}^2 + n_\mathrm{o}^4\sin^2\phi + n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right)^2 - 4\left(n_\mathrm{o}^2\sin^2\phi + n_\mathrm{e}^2\cos^2\phi\right) n_\mathrm{o}^4 n_\mathrm{e}^2 \\ &= \left(n_\mathrm{o}^2 n_\mathrm{e}^2 + n_\mathrm{o}^4\sin^2\phi + n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right)^2 - 4\left(n_\mathrm{o}^2 n_\mathrm{e}^2\right)\left(n_\mathrm{o}^4\sin^2\phi\right) \\ &\quad - \left(n_\mathrm{o}^2 n_\mathrm{e}^2\right) 4\left(n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right) \\ &= \left(-n_\mathrm{o}^2 n_\mathrm{e}^2 + n_\mathrm{o}^4\sin^2\phi + n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right)^2 \end{aligned}$$

$$\begin{aligned} n^2 &= \frac{B \pm \sqrt{B^2 - 4AC}}{2A} \\ &= \frac{\left(n_\mathrm{o}^2 n_\mathrm{e}^2 + n_\mathrm{o}^4\sin^2\phi + n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right) \pm \left(-n_\mathrm{o}^2 n_\mathrm{e}^2 + n_\mathrm{o}^4\sin^2\phi + n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right)}{2\left(n_\mathrm{o}^2\sin^2\phi + n_\mathrm{e}^2\cos^2\phi\right)} \\ &= \frac{2\left(n_\mathrm{o}^4\sin^2\phi + n_\mathrm{o}^2 n_\mathrm{e}^2\cos^2\phi\right)}{2\left(n_\mathrm{o}^2\sin^2\phi + n_\mathrm{e}^2\cos^2\phi\right)}, \frac{2 n_\mathrm{o}^2 n_\mathrm{e}^2}{2\left(n_\mathrm{o}^2\sin^2\phi + n_\mathrm{e}^2\cos^2\phi\right)} \\ &= n_\mathrm{o}^2, \frac{n_\mathrm{o}^2 n_\mathrm{e}^2}{\left(n_\mathrm{o}^2\sin^2\phi + n_\mathrm{e}^2\cos^2\phi\right)} \end{aligned}$$

The first index (5.31) corresponds to the electric field component which points in the x-direction (i.e. the arrow tail in Fig. 5.2). We shall avoid the analysis to prove this.

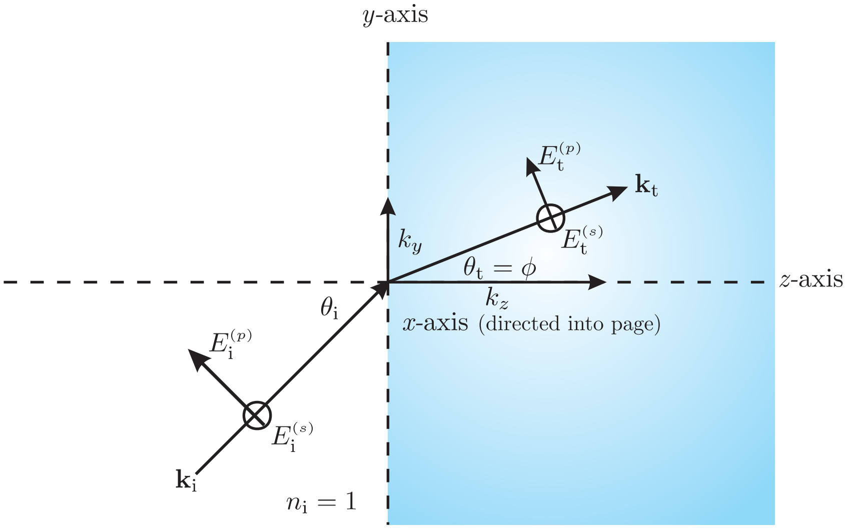


Figure 5.2 Propagation of light in a uniaxial crystal.

However, this assignment makes sense because, regardless of ϕ, the field component in the x-direction never notices the index n_e associated with excitation of the crystal lattice in the z-dimension. The x field component is associated with the *ordinary* wave because just as in an isotropic medium such as glass, the index does not vary with ϕ. The other index (5.32), which does vary with angle ϕ, is associated with the electric field component contained in the y–z plane. This component of the electric field is directed partly along the optic axis, and it is called the extraordinary wave.

If $\phi = 0$, then the **k**-vector is directed exactly along the optic axis, and neither electric field component experiences the unusual dimension (i.e. the z-direction). Notice that when $\phi = 0$, (5.32) reduces to $n = n_\mathrm{o}$ so that both indices are the same. On the other hand, if $\phi = \pi/2$ then (5.32) reduces to $n = n_\mathrm{e}$. (A wave plate is cut with the optic axis parallel to the surface (as opposed to Fig. 5.2). Then when the light goes through at normal incidence, the angle is $\phi = \pi/2$, and there is a slow and a fast axis with indices n_o and n_e.)

Finally, let us consider refraction as light enters a uniaxial crystal. Suppose that the crystal is cut such that the *optic axis* lies perpendicular to the surface (not the way a wave plate is cut). For this geometry, the surface of the crystal lies parallel to the x–y plane as depicted in Fig. 5.2. If the light hits the crystal surface straight on, the index of refraction is n_o, regardless of the orientation of polarization since $\phi = 0$ for normal incidence. However, if the light strikes the surface at an angle, then p-polarized light experiences an index that varies with angle since the electric field has a component of polarization along the optic axis. In contrast, s-polarized always experiences an index n_o since the electric field never points in the z-direction regardless of the incident angle. As was previously mentioned, in a uniaxial crystal we are free to choose the x-axis to be any direction perpendicular to the optic axis, and so we choose it to be along the s-polarized direction.

Snell's law (3.7) describes the connection between the **k**-vectors incident upon and transmitted through the surface. If we assume that the index outside of the crystal is $n_i = 1$, Snell's law may be written as

$$\sin\theta_{\rm i} = n\sin\theta_{\rm t} \tag{5.33}$$

where n is the index inside the crystal and the transmitted angle is $\theta_{\rm t} = \phi$. Since s-polarized light sees only the index $n_{\rm o}$, regardless of the incident angle $\theta_{\rm i}$, (5.33) can be solved for $\theta_{\rm t}$ in terms of the incident angle $\theta_{\rm i}$:

$$\sin\theta_{\rm t} = \frac{\sin\theta_{\rm i}}{n_{\rm o}} \quad (s\text{-polarized, optic axis} \perp \text{surface}) \tag{5.34}$$

This corresponds to an "ordinary" wave since it behaves the same way as light entering an isotropic medium.

The situation is more complicated for p-polarized light where n is a function of $\theta_{\rm t}$ via (5.32). It is left as an exercise (see P 5.2) to invert (5.33) for the transmitted angle $\theta_{\rm t}$ in terms of $\theta_{\rm i}$:

$$\tan\theta_{\rm t} = \frac{n_{\rm e}\sin\theta_{\rm i}}{n_{\rm o}\sqrt{n_{\rm e}^2 - \sin^2\theta_{\rm i}}} \quad (p\text{-polarized, optic axis} \perp \text{surface}) \tag{5.35}$$

It is no wonder that Huygens called this behavior "extraordinary"; as strange as this formula looks, it is Snell's law, but with an angularly dependent index. The s- and p-polarized light refract into the crystal at two different angles; they travel at two different velocities in the crystal; and they have two different wavelengths in the crystal.

5.5 Poynting Vector in a Uniaxial Crystal

We continue our investigation of a uniaxial crystal cut with its optic axis perpendicular to the surface as in Fig. 5.2. Note that this is a special case of crystal orientation, and the formulas that we derive are specific to that orientation. Ordinary polarized light (s-polarized in our crystal orientation) refracts at the surface the same way that it does for an isotropic material such as glass (i.e. the material index does not vary with incident angle). In a uniaxial crystal, the Poynting vector **S** for purely ordinary polarized light points in the same direction as the **k**-vector.

The refraction of extraordinary polarized light (p-polarized in our crystal orientation) is another story since the index varies with angle. For extraordinary polarized light, Snell's law generates the connection described by (5.35), relating the directions of the incident and transmitted **k**-vectors. The Poynting vector **S**, however, connects to the incident angle $\theta_{\rm i}$ through yet a different expression. (Recall from the discussion in connection with (5.7) and (5.8) that **S** and **k** are not parallel to each other.) In this section, we derive an expression similar to (5.35), but which applies to **S** rather than to **k**.

To find the direction of energy flow, we must calculate $\mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0$. To do this we need to know **E**. From the constitutive relation (5.3) and the definitions (5.21) we have

$$\begin{aligned}\epsilon_0\mathbf{E} + \mathbf{P} &= \epsilon_0\left[(1+\chi_x)E_x\hat{\mathbf{x}} + (1+\chi_y)E_y\hat{\mathbf{y}} + (1+\chi_z)E_z\hat{\mathbf{z}}\right]\\ &= \epsilon_0\left(n_{\rm o}^2E_x\hat{\mathbf{x}} + n_{\rm o}^2E_y\hat{\mathbf{y}} + n_{\rm e}^2E_z\hat{\mathbf{z}}\right)\end{aligned} \tag{5.36}$$

Again, let us consider the **k**-vector to lie in the y-z plane so that its components are given as before by (5.30). Upon substitution of these expressions into (5.7) we have

$$\begin{aligned}\mathbf{k}\cdot(\epsilon_0\mathbf{E}+\mathbf{P}) &= k\left(\hat{\mathbf{y}}\sin\phi+\hat{\mathbf{z}}\cos\phi\right)\cdot\epsilon_0\left(n_\mathrm{o}^2E_x\hat{\mathbf{x}}+n_\mathrm{o}^2E_y\hat{\mathbf{y}}+n_\mathrm{e}^2E_z\hat{\mathbf{z}}\right)\\ &= \epsilon_0 k\left(n_\mathrm{o}^2E_y\sin\phi+n_\mathrm{e}^2E_z\cos\phi\right)\\ &= 0\end{aligned}\tag{5.37}$$

Therefore, the y and z components of the field are related through

$$E_z = -\frac{n_\mathrm{o}^2E_y}{n_\mathrm{e}^2}\tan\phi\tag{5.38}$$

Note that E_y and E_z are exactly the components of the electric field that comprise extraordinary polarized light; the ordinary component of the field points in the x-direction. We may write the extraordinary polarized electric field as

$$\mathbf{E} = E_y\left(\hat{\mathbf{y}}-\hat{\mathbf{z}}\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\tan\phi\right)\qquad\text{(ordinary polarized)}\tag{5.39}$$

Before computing the Poynting vector, we also need to express the magnetic field in terms of the electric field. To do this we take advantage of (5.30):

$$\begin{aligned}\mathbf{B} &= \frac{\mathbf{k}\times\mathbf{E}}{\omega}\\ &= \frac{k\left(\hat{\mathbf{y}}\sin\phi+\hat{\mathbf{z}}\cos\phi\right)\times E_y\left(\hat{\mathbf{y}}-\hat{\mathbf{z}}\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\tan\phi\right)}{\omega}\\ &= -\hat{\mathbf{x}}\frac{kE_y}{\omega}\left(\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\sin\phi\tan\phi+\cos\phi\right)\end{aligned}\tag{5.40}$$

We proceed with the computation of the Poynting vector. Using (5.39) and (5.40) we get

$$\begin{aligned}\mathbf{S} &= \mathbf{E}\times\frac{\mathbf{B}}{\mu_0}\\ &= -E_y\left(\hat{\mathbf{y}}-\hat{\mathbf{z}}\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\tan\phi\right)\times\frac{kE_y}{\mu_0\omega}\left(\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\sin\phi\tan\phi+\cos\phi\right)\hat{\mathbf{x}}\\ &= \frac{kE_y^2}{\mu_0\omega}\left(\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\sin\phi\tan\phi+\cos\phi\right)\left(\hat{\mathbf{z}}+\hat{\mathbf{y}}\frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\tan\phi\right)\end{aligned}\tag{5.41}$$

Keep in mind that ϕ refers to the direction of the **k**-vector. The above equation demonstrates that the Poynting vector **S** lies along another direction. Let us label the direction of the Poynting vector with the angle ϕ'. This angle can be obtained from the ratio of the two vector components of **S** as follows:

$$\tan\phi' \equiv \frac{S_y}{S_z} = \frac{n_\mathrm{o}^2}{n_\mathrm{e}^2}\tan\phi\qquad\text{(extraordinary polarized)}\tag{5.42}$$

While the **k**-vector is characterized by the angle ϕ, the Poynting vector is characterized by the angle ϕ'. We can find the connection between the incident angle θ_i and ϕ' by taking advantage of the already known connection between θ_i and $\theta_\mathrm{t} = \phi$. Combining (5.35) and (5.42), we obtain

$$\tan\phi' = \frac{n_\mathrm{o}\sin\theta_\mathrm{i}}{n_\mathrm{e}\sqrt{n_\mathrm{e}^2 - \sin^2\theta_\mathrm{i}}} \quad \text{(extraordinary polarized)} \tag{5.43}$$

This describes the direction that *extraordinary rays* take through the crystal. Since this corresponds to the direction of energy flow, it also corresponds to what is seen by an observer. When an object is observed through a crystal (acting as a window), the energy associated with ordinary and extraordinary polarized light follow different paths, giving rise to two different images. This phenomenon is called *birefringence.* The energy flow associated with ordinary polarized light obeys Snell's law, while energy flow associated with extraordinary polarized light does not. When Huygens saw this, he said "how extraordinary!" (See Appendix 5.B.)

Due to our specific choice of orientation for the optic axis in this section, we had the case where ordinary polarized light is s-polarized light, and extraordinary polarized light is p-polarized light. This is the generally the case for arbitrary orientations of the optic axis. In general, the s- and p-polarized portions of the incident light can each give rise to both extraordinary and ordinary rays.

Appendix 5.A Rotation of Coordinates

In this appendix, we go through the tedious labor of showing that (5.1) can always be written as (5.3), given that the susceptibility tensor is symmetric (i.e. $\chi_{ij} = \chi_{ji}$). This amounts to an eigenvalue problem, which we accomplish here via rotations of the coordinate system. We have

$$\mathbf{P} = \epsilon_0\chi\mathbf{E} \tag{5.44}$$

where

$$\mathbf{E} \equiv \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \qquad \mathbf{P} \equiv \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} \qquad \chi \equiv \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{xy} & \chi_{yy} & \chi_{yz} \\ \chi_{xz} & \chi_{yz} & \chi_{zz} \end{bmatrix} \tag{5.45}$$

Our task now is to find a new coordinate system x', y', and z' for which the susceptibility tensor is diagonal. That is, we want to choose x', y', and z' such that

$$\mathbf{P}' = \epsilon_0\chi'\mathbf{E}', \tag{5.46}$$

where

$$\mathbf{E}' \equiv \begin{bmatrix} E'_{x'} \\ E'_{y'} \\ E'_{z'} \end{bmatrix} \qquad \mathbf{P}' \equiv \begin{bmatrix} P'_{x'} \\ P'_{y'} \\ P'_{z'} \end{bmatrix} \qquad \chi' \equiv \begin{bmatrix} \chi'_{x'x'} & 0 & 0 \\ 0 & \chi'_{y'y'} & 0 \\ 0 & 0 & \chi'_{z'z'} \end{bmatrix} \tag{5.47}$$

To arrive at the new coordinate system, we are free to make pure rotation transformations. From (4.31) and (4.32), a rotation through an angle γ about the z-axis, followed by a rotation through an angle β about the resulting y-axis, and finally a rotation through an angle α about the new x-axis, can be written as

$$\begin{aligned}\mathsf{R} &\equiv \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \cos\beta\cos\gamma & \cos\beta\sin\gamma & \sin\beta \\ -\cos\alpha\sin\gamma - \sin\alpha\sin\beta\cos\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\beta \\ \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma & -\sin\alpha\cos\gamma - \cos\alpha\sin\beta\sin\gamma & \cos\alpha\cos\beta \end{bmatrix}\end{aligned} \tag{5.48}$$

The matrix R produces an arbitrary rotation of coordinates in three dimensions. We can use this matrix to transform from the original coordinate system to the new one:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5.49}$$

More important to our present purpose are the connections

$$\begin{aligned}\mathbf{E}' &= \mathsf{R}\mathbf{E} \\ \mathbf{P}' &= \mathsf{R}\mathbf{P}\end{aligned} \tag{5.50}$$

These transformations can be inverted to give

$$\begin{aligned}\mathbf{E} &= \mathsf{R}^{-1}\mathbf{E}' \\ \mathbf{P} &= \mathsf{R}^{-1}\mathbf{P}'\end{aligned} \tag{5.51}$$

where

$$\begin{aligned}\mathsf{R}^{-1} &= \begin{bmatrix} \cos\beta\cos\gamma & -\cos\alpha\sin\gamma - \sin\alpha\sin\beta\cos\gamma & \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma \\ \cos\beta\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & -\sin\alpha\cos\gamma - \cos\alpha\sin\beta\sin\gamma \\ \sin\beta & \sin\alpha\cos\beta & \cos\alpha\cos\beta \end{bmatrix} \\ &= \begin{bmatrix} R_{11} & R_{21} & R_{31} \\ R_{12} & R_{22} & R_{32} \\ R_{13} & R_{23} & R_{33} \end{bmatrix} = \mathsf{R}^T\end{aligned} \tag{5.52}$$

Note that the inverse of the rotation matrix is the same as its transpose, an important feature that we exploit in what follows.

Upon inserting (5.51) into (5.44) we have

$$\mathsf{R}^{-1}\mathbf{P}' = \epsilon_0\chi\mathsf{R}^{-1}\mathbf{E}' \tag{5.53}$$

or

$$\mathbf{P}' = \epsilon_0\mathsf{R}\chi\mathsf{R}^{-1}\mathbf{E}' \tag{5.54}$$

From this equation we see that the new susceptibility tensor we seek for (5.46) is

$$\begin{aligned}\chi' &\equiv \mathbf{R}\chi\mathbf{R}^{-1}\\ &= \begin{bmatrix} R_{11} & R_{12} & R_{13}\\ R_{21} & R_{22} & R_{23}\\ R_{31} & R_{32} & R_{33}\end{bmatrix}\begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz}\\ \chi_{xy} & \chi_{yy} & \chi_{yz}\\ \chi_{xz} & \chi_{yz} & \chi_{zz}\end{bmatrix}\begin{bmatrix} R_{11} & R_{21} & R_{31}\\ R_{12} & R_{22} & R_{32}\\ R_{13} & R_{23} & R_{33}\end{bmatrix}\\ &= \begin{bmatrix} \chi'_{x'x'} & \chi'_{x'y'} & \chi'_{x'z'}\\ \chi'_{x'y'} & \chi'_{y'y'} & \chi'_{y'z'}\\ \chi'_{x'z'} & \chi'_{y'z'} & \chi'_{z'z'}\end{bmatrix}\end{aligned} \tag{5.55}$$

We have expressly indicated that the off-diagonal terms of χ' are symmetric (i.e. $\chi'_{ij} = \chi'_{ji}$). This can be verified by performing the multiplication in (5.55). It is a consequence of χ being symmetric and $\mathbf{R}^{-1}$ being equal to $\mathbf{R}^T$

The three off-diagonal elements (appearing both above and below the diagonal) are

$$\begin{aligned}\chi'_{x'y'} &= R_{11}\left(R_{21}\chi_{xx} + R_{22}\chi_{xy} + R_{23}\chi_{xz}\right) + R_{12}\left(R_{21}\chi_{xy} + R_{22}\chi_{yy} + R_{23}\chi_{yz}\right)\\ &\quad + R_{13}\left(R_{21}\chi_{xz} + R_{22}\chi_{yz} + R_{23}\chi_{zz}\right)\\ \chi'_{x'z'} &= R_{11}\left(R_{31}\chi_{xx} + R_{32}\chi_{xy} + R_{33}\chi_{xz}\right) + R_{12}\left(R_{31}\chi_{xy} + R_{32}\chi_{yy} + R_{33}\chi_{yz}\right)\\ &\quad + R_{13}\left(R_{31}\chi_{xz} + R_{32}\chi_{yz} + R_{33}\chi_{zz}\right)\\ \chi'_{y'z'} &= R_{21}\left(R_{31}\chi_{xx} + R_{32}\chi_{xy} + R_{33}\chi_{xz}\right) + R_{22}\left(R_{31}\chi_{xy} + R_{32}\chi_{yy} + R_{33}\chi_{yz}\right)\\ &\quad + R_{23}\left(R_{31}\chi_{xz} + R_{32}\chi_{yz} + R_{33}\chi_{zz}\right)\end{aligned} \tag{5.56}$$

These expressions are cumbersome. However, we can make all three of them equal to zero since we have three degrees of freedom in the angles α, β, and γ. Although, we do not expressly solve for the angles, we have demonstrated that it is always possible to set

$$\begin{aligned}\chi'_{x'y'} &= 0\\ \chi'_{x'z'} &= 0\\ \chi'_{y'z'} &= 0\end{aligned} \tag{5.57}$$

This justifies (5.3).

Appendix 5.B Huygens' Elliptical Construct for a Uniaxial Crystal

In 1690 Christian Huygens developed a way to predict the direction of extraordinary rays in a crystal by examining an elliptical wavelet. The point on the elliptical wavelet that propagates along the optic axis is assumed to experience the index n_e. The point on the elliptical wavlet that propagates perpendicular to the optic axis is assumed to experience the index n_o. It turns out that Huygens' approach agreed with the direction energy propagation (5.43) (as opposed to the direction of the **k**-vector). This was quite satisfactory in Huygens' day (except that he was largely ignored for a century, owing to Newton's corpuscular theory) since the direction of energy propagation is what an observer sees.

Consider a plane wave entering a uniaxial crystal. In Huygens' point of view, each point on a wave front acts as a wavelet source which combines with neighboring wavelets

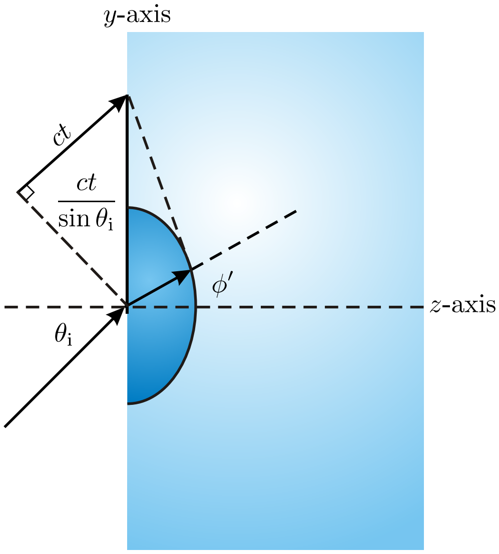


Figure 5.3 Elliptical wavelet.

to preserve the overall plane wave pattern. Inside the crystal, the wavelets propagate in the shape of an ellipse. The equation for an elliptical wave front after propagating during a time t is

$$\frac{y^2}{(ct/n_e)^2}+\frac{z^2}{(ct/n_o)^2}=1 \tag{5.58}$$

After rearranging, the equation of the ellipse inside the crystal can also be written as

$$z=\frac{ct}{n_o}\sqrt{1-\frac{y^2}{(ctn_e)^2}} \tag{5.59}$$

In order to have the wavelet joint neatly with other wavelets to build a plane wave, the wave front of the ellipse must be parallel to a new wave front entering the surface at a distance $ct/\sin\theta_i$ above the original point. This distance is represented by the hypotenuse of the right triangle seen in Fig. 5.3. Let the point where the wave front touches the ellipse be denoted by $(y,z)=(z\tan\phi',z)$. The slope (rise over run) of the line that connects these two points is then

$$\frac{dz}{dy}=-\frac{z}{ct/\sin\theta_i - z\tan\phi'} \tag{5.60}$$

At the point where the wave front touches the ellipse (i.e., $(y,z)=(z\tan\phi',z)$), the slope of the curve for the ellipse is

$$\frac{dz}{dy}=\frac{-yn_e^2}{n_oct\sqrt{1-\frac{y^2}{(ct/n_e)^2}}}=-\frac{n_e^2y}{n_o^2z}=-\frac{n_e^2}{n_o^2}\tan\phi' \tag{5.61}$$

We would like these two slopes to be the same. We therefore set them equal to each other:

$$-\frac{n_e^2}{n_o^2}\tan\phi'=-\frac{z}{ct/\sin\theta_i - z\tan\phi'} \Rightarrow \frac{ct}{z}\frac{n_e^2}{n_o^2}\frac{\tan\phi'}{\sin\theta_i}=\frac{n_e^2}{n_o^2}\tan^2\phi'+1 \tag{5.62}$$

If we evaluate (5.58) for the point $(y, z) = (z \tan \phi', z)$, we obtain

$$\frac{ct}{z} = n_\mathrm{o}\sqrt{\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\phi' + 1} \tag{5.63}$$

Upon substitution of this into (5.62) we arrive at

$$\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\frac{\tan\phi'}{\sin\theta_\mathrm{i}} = \sqrt{\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\phi' + 1} \Rightarrow \frac{n_\mathrm{e}^4}{n_\mathrm{o}^2}\frac{\tan^2\phi'}{\sin^2\theta_\mathrm{i}} = \frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\phi' + 1 \tag{5.64}$$

$$\Rightarrow \left[\frac{n_\mathrm{e}^2}{\sin^2\theta_\mathrm{i}} - 1\right]\tan^2\phi' = \frac{n_\mathrm{o}^2}{n_\mathrm{e}^2} \Rightarrow \tan\phi' = \frac{n_\mathrm{o}\sin\theta_\mathrm{i}}{n_\mathrm{e}\sqrt{n_\mathrm{e}^2 - \sin^2\theta_\mathrm{i}}} \tag{5.65}$$

This agrees with (5.43) as anticipated. Again, Huygens' approach obtained the correct direction of the Poynting vector associated with the extraordinary wave.

Exercises

5.3 Fresnel's Equation

P5.1 Suppose you have a crystal with $n_x = 1.5$, $n_y = 1.6$, and $n_z = 2.0$. Use Fresnel's equation to determine what the two indices of refraction are for a k-vector in the crystal along the $\hat{\mathbf{u}} = (\hat{\mathbf{x}} + 2\hat{\mathbf{y}} + 3\hat{\mathbf{z}})/\sqrt{14}$ direction.

5.4 Uniaxial Crystal

P5.2 Derive (5.35).

P5.3 A quartz plate (uniaxial crystal with the optic axis perpendicular to the surfaces) has thickness $d = 0.96$ mm. The indices of refraction are $n_\text{o} = 1.54424$ and $n_\text{e} = 1.55335$. A plane wave with wavelength $\lambda_\text{vac} = 633$ nm passes through the plate. After emerging from the crystal, there is a phase difference Δ between the two polarization components of the plane wave, and this phase difference depends on incident angle θ_i. Use a computer to plot Δ as a function of incident angle from zero to 90°.

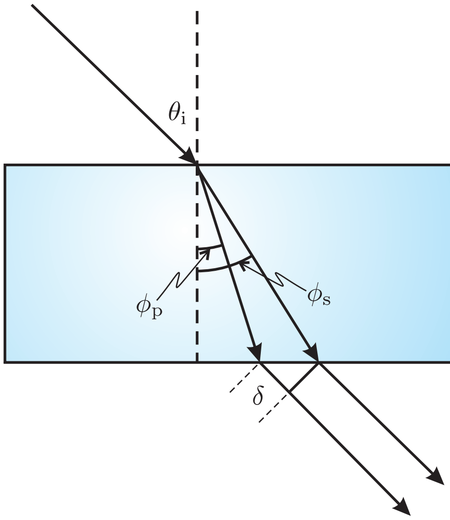


Figure 5.4 Diagram for P 5.3.

HINT: For s-polarized light, show that the number of wavelengths that fit in the plate is $\frac{d}{(\lambda_\text{vac}/n_\text{o})\cos\phi_s}$. For p-polarized light, show that the number of wavelengths that fit in the plate and the extra leg δ outside of the plate (see Fig. 5.4) is $\frac{d}{(\lambda_\text{vac}/n_p)\cos\phi_p} + \frac{\delta}{\lambda_\text{vac}}$, where $\delta = d\,[\tan\phi_s - \tan\phi_p]\sin\theta_\text{i}$ and n_p is given by (5.32). Find the difference between these expressions and multiply by 2π to find Δ.

L5.4 In the laboratory, send a HeNe laser ($\lambda_{\text{vac}} = 633$ nm) through two crossed polarizers, oriented at 45° and 135°. Place the quartz plate described in P 5.3 between the polarizers on a rotation stage. Now equal amounts of s- and p-polarized light strike the crystal as it is rotated from normal incidence.

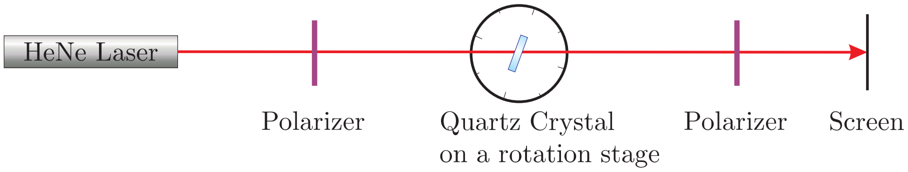


Figure 5.5 Schematic for L 5.4.

If the phase shift between the two paths is an odd integer times π, the crystal acts as a half wave plate and maximum transmission through the second polarizer results. If the phase shift is an even integer times π, then minimum transmission through the second polarizer results. Plot these measured maximum and minimum points on your computer-generated graph of the previous problem.

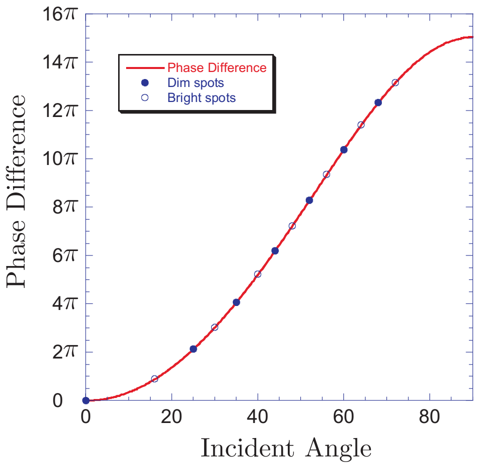


Figure 5.6 Plot for P 5.3 and L 5.4.

Review, Chapters 1–5

Students preparing for an exam will want to understand the following questions and problems thoroughly enough to be able to work them without referring back to previous chapters.

True and False Questions

R1 T or F: The optical index of *any* material (not vacuum) varies with frequency.

R2 T or F: The frequency of light can change as it enters a *crystal* (consider low intensity—no nonlinear effects).

R3 T or F: The entire expression $\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ associated with a light field (both the real part and the imaginary parts) is physically relevant.

R4 T or F: The *real* part of the refractive index cannot be less than one.

R5 T or F: s-polarized light and p-polarized light experience the *same* phase shift upon reflection from a material with complex index.

R6 T or F: When light is incident upon a material interface at Brewster's angle, only one polarization can *transmit.*

R7 T or F: When light is incident upon a material interface at Brewster's angle one of the polarizations stimulates dipoles in the material to oscillate with orientation along the *direction of the reflected* **k***-vector.*

R8 T or F: The critical angle for total internal reflection exists on both sides of a material interface.

R9 T or F: From any given location above a (smooth flat) surface of water, it is possible to see objects positioned anywhere under the water.

R10 T or F: From any given location beneath a (smooth flat) surface of water, it is possible to see objects positioned anywhere above the water.

R11 T or F: An evanescent wave travels *parallel* to the surface interface on the *transmitted* side.

R12 T or F: When p-polarized light enters a material at Brewster's angle, the *intensity* of the transmitted beam is the same as the intensity of the incident beam.

R13 T or F: For incident angles beyond the critical angle for total internal reflection, the Fresnel coefficients t_s and t_p are both zero.

R14 T or F: As light enters a crystal, the Poynting vector *always* obeys Snell's law.

R15 T or F: As light enters a crystal, the **k**-vector does *not* obey Snell's for the extraordinary wave.

Problems

R16 (a) Write down Maxwell's equations.

(b) Derive the wave equation for **E** under the assumptions that $\mathbf{J}_{\text{free}} = 0$ and $\mathbf{P} = \epsilon_0 \chi \mathbf{E}$. Note: $\nabla \times (\nabla \times \mathbf{f}) = \nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f}$.

(c) Show by direct substitution that $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$ is a solution to the wave equation. Find the resulting connection between k and ω. Give appropriate definitions for c and n, assuming that χ is real.

(d) If $\mathbf{k} = k\hat{\mathbf{z}}$ and $\mathbf{E}_0 = E_0 \hat{\mathbf{x}}$, find the associated **B**-field.

(e) The Poynting vector is $\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$, where the fields are real. Derive an expression for $I \equiv \langle S \rangle_t$.

R17 A horizontal and a vertical polarizer are placed in series, and horizontally polarized light with Jones vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ enters the system.

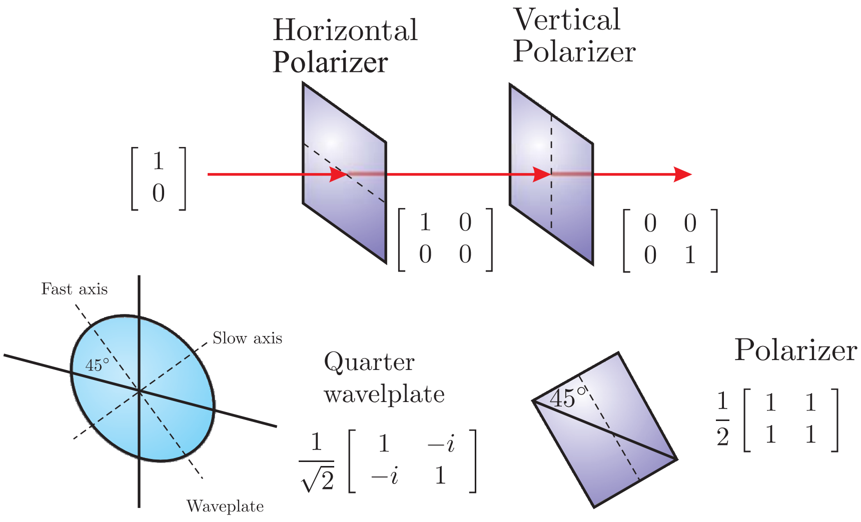

Horizontal Polarizer
Vertical Polarizer
$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
Fast axis
Slow axis
45°
Waveplate
Quarter wavelplate
$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$
45°
Polarizer
$\frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Figure 5.7

(a) What is the Jones vector of the transmitted field?

(b) Now a polarizer at 45° is inserted between the other two polarizers. What is the Jones vector of the transmitted field? How does the final intensity compare to initial intensity?

(c) Now a quarter wave plate with a fast-axis angle of 45° is inserted between the two polarizers (instead of the polarizer of part (b)). What is the Jones vector of the transmitted field? How does the final intensity compare to initial intensity?

R18 (a) Find the Jones matrix for half wave plate with its fast axis making an arbitrary angle θ with the x-axis.

HINT: Project an arbitrary polarization with E_x and E_y onto the fast and slow axes of the wave plate. Shift the slow axis phase by π, and then project the field components back onto the horizontal and vertical axes. The answer is

$$\begin{bmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \sin^2\theta - \cos^2\theta \end{bmatrix}$$

(b) We desire to attenuate continuously a polarized laser beam using a half wave plate and a polarizer aligned to the initial polarization of the beam (see figure). The fast axis of the half wave plate is initially aligned in the direction of polarization and then rotated through an angle θ. What is the ratio of the intensity exiting the polarizer to the incoming intensity as a function of θ?

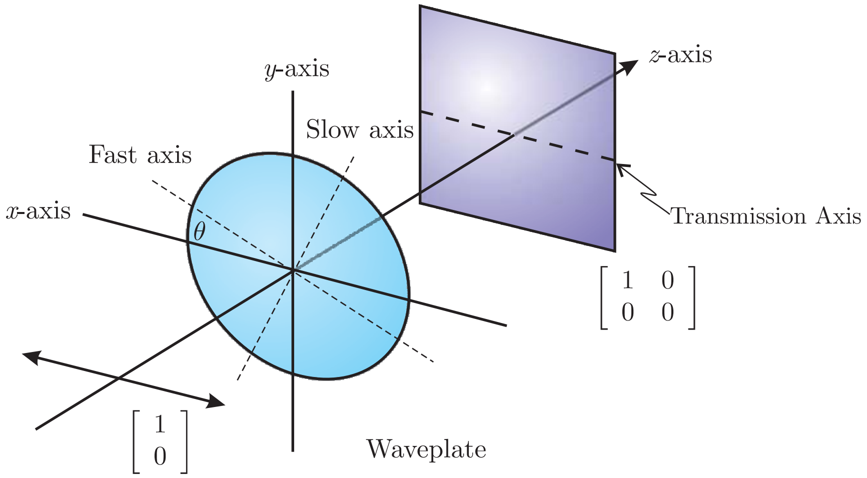

y-axis
z-axis
Slow axis
Fast axis
x-axis
θ
Transmission Axis
$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$
Waveplate

Figure 5.8 Polarizing Elements

R19 Consider an interface between two isotropic media where the incident field is defined by

$$\mathbf{E}_\mathrm{i} = \left[E_\mathrm{i}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_\mathrm{i} - \hat{\mathbf{z}}\sin\theta_\mathrm{i}\right) + \hat{\mathbf{x}}E_\mathrm{i}^{(s)}\right] e^{i[k_\mathrm{i}(y\sin\theta_\mathrm{i}+z\cos\theta_\mathrm{i})-\omega_\mathrm{i}t]}$$

The plane of incidence is shown in Fig. 5.9

(a) By inspection of the figure, write down similar expressions for the reflected and transmitted fields (i.e. $\mathbf{E}_\mathrm{r}$ and $\mathbf{E}_\mathrm{t}$).

(b) Find an expression relating $\mathbf{E}_\mathrm{i}$, $\mathbf{E}_\mathrm{r}$, and $\mathbf{E}_\mathrm{t}$ using the boundary condition at the interface. From this expression obtain the law of reflection and Snell's law.

(c) The boundary condition requiring that the tangential component of **B** must be continuous leads to

$$n_{\rm i}(E_{\rm i}^{(p)} - E_{\rm r}^{(p)}) = n_t E_{\rm t}^{(p)}$$

$$n_{\rm i}(E_{\rm i}^{(s)} - E_{\rm r}^{(s)})\cos\theta_{\rm i} = n_{\rm t} E_{\rm t}^{(s)}\cos\theta_{\rm t}$$

Use this and the results from part (b) to derive

$$r_p \equiv \frac{E_{\rm r}^{(p)}}{E_{\rm i}^{(p)}} = -\frac{\tan(\theta_{\rm i} - \theta_{\rm t})}{\tan(\theta_{\rm i} + \theta_{\rm t})}$$

You may use the identity

$$\frac{\sin\theta_{\rm i}\cos\theta_{\rm i} - \sin\theta_{\rm t}\cos\theta_{\rm t}}{\sin\theta_{\rm i}\cos\theta_{\rm i} + \sin\theta_{\rm t}\cos\theta_{\rm t}} = \frac{\tan(\theta_{\rm i} - \theta_{\rm t})}{\tan(\theta_{\rm i} + \theta_{\rm t})}$$

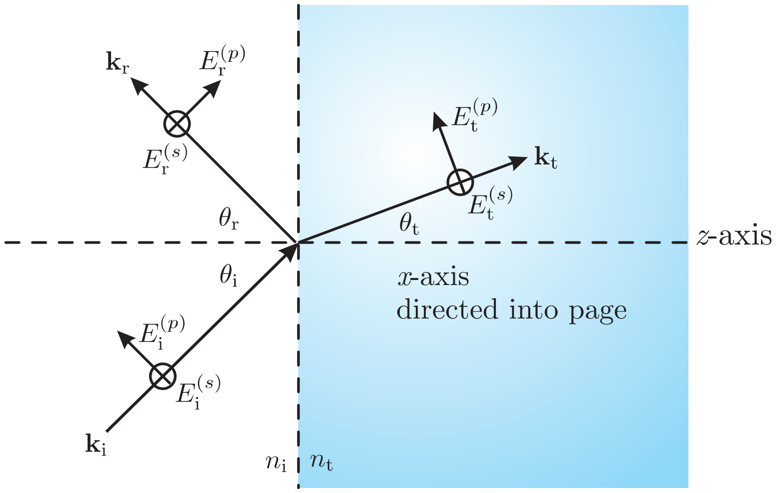


Figure 5.9

R20 The Fresnel equations are

$$r_s \equiv \frac{E_{\rm r}^{(s)}}{E_{\rm i}^{(s)}} = \frac{\sin\theta_{\rm t}\cos\theta_{\rm i} - \sin\theta_{\rm i}\cos\theta_{\rm t}}{\sin\theta_{\rm t}\cos\theta_{\rm i} + \sin\theta_{\rm i}\cos\theta_{\rm t}}$$

$$t_s \equiv \frac{E_{\rm t}^{(s)}}{E_{\rm i}^{(s)}} = \frac{2\sin\theta_{\rm t}\cos\theta_{\rm i}}{\sin\theta_{\rm t}\cos\theta_{\rm i} + \sin\theta_{\rm i}\cos\theta_{\rm t}}$$

$$r_p \equiv \frac{E_{\rm r}^{(p)}}{E_{\rm i}^{(p)}} = \frac{\cos\theta_{\rm t}\sin\theta_{\rm t} - \cos\theta_{\rm i}\sin\theta_{\rm i}}{\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}}$$

$$t_p \equiv \frac{E_{\rm t}^{(p)}}{E_{\rm i}^{(p)}} = \frac{2\cos\theta_{\rm i}\sin\theta_{\rm t}}{\cos\theta_{\rm t}\sin\theta_{\rm t} + \cos\theta_{\rm i}\sin\theta_{\rm i}}$$

(a) Find what each of these equations reduces to when $\theta_{\rm i} = 0$. Give your answer in terms of $n_{\rm i}$ and $n_{\rm t}$.

(b) What percent of light (intensity) reflects from a glass surface ($n = 1.5$) when light enters from air ($n = 1$) at normal incidence?

(c) What percent of light reflects from a glass surface when light exits into air at normal incidence?

R21 Light goes through a glass prism with optical index $n = 1.55$. The light enters at Brewster's angle and exits at normal incidence.

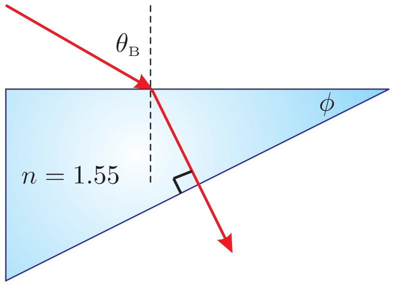


Figure 5.10

(a) *Derive* and calculate Brewster's angle $\theta_{\rm B}$. You may use the results of R19 (c).

(b) Calculate ϕ.

(c) What percent of the light (power) goes all the way through the prism if it is p-polarized? Ignore light that might make multiple reflections within the prism and come out with directions other than that shown by the arrow. You may use the Fresnel coefficients given in R20.

(d) What percent for s-polarized light?

R22 A 45°- 90°- 45° prism is a good device for reflecting a beam of light parallel to the initial beam. The exiting beam will be parallel to the entering beam even when the incoming beam is not normal to the front surface (although it needs to be in the plane of the drawing).

(a) How large an angle θ can be tolerated before there is no longer total internal reflection at both interior surfaces? Assume $n = 1$ outside of the prism and $n = 1.5$ inside.

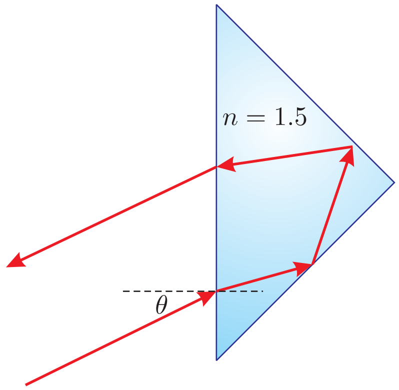


Figure 5.11

(b) If the light enters and leaves the prism at normal incidence, what will the difference in phase be between the s and p-polarizations? You may use the Fresnel coefficients given in R20.

R23 Second harmonic generation (the conversion of light with frequency ω into light with frequency 2ω) can occur when very intense laser light travels in a material. For good harmonic production, the laser light and the second harmonic light need to travel at the *same* speed in the material. In other words, both frequencies need to have the same index of refraction so that harmonic light produced down stream joins in phase with the harmonic light produced up stream, referred to as *phase matching*. This ensures a coherent building of the second harmonic field rather than destructive cancellations.

Unfortunately, the index of refraction is almost never the same for different frequencies in a given material, owing to dispersion. However, we can achieve phase matching in some crystals where one frequency propagates as an ordinary wave and the other propagates as an extraordinary wave. We cause the two indices to be precisely the same by *tuning* the angle of the crystal.

Consider a ruby laser propagating and generating the second harmonic in a uniaxial KDP crystal (potassium dihydrogen phosphate). The indices of refraction are given by n_o and

$$\frac{n_o n_e}{\sqrt{n_o^2 \sin^2\phi + n_e^2 \cos^2\phi}}$$

where ϕ is the angle made with the optic axis. At the frequency of a ruby laser, KDP has indices $n_o(\omega) = 1.505$ and $n_e(\omega) = 1.465$. At the frequency of the second harmonic, the indices are $n_o(2\omega) = 1.534$ and $n_e(2\omega) = 1.487$.

Show that phase matching can be achieved if the laser is polarized so that it experiences only the ordinary index and the second harmonic light is polarized perpendicular to that. At what angle ϕ does this phase matching occur?

Selected Answers

R17: (b) 1/4, (c) 1/2.

R20: (b) 4% (c) 4%.

R21: (b) 33°, (c) 95%, (d) 79%.

R22: (a) 4.8°, (b) 74°.

R23: 51.12°.

Chapter 6

Multiple Parallel Interfaces

6.1 Introduction

In chapter 3, we studied the transmission and reflection of light at a single interface between two isotropic and homogeneous materials with indices n_0 and n_2. We found that the percent of light reflected and transmitted depends on the incident angle θ_0 and on whether the light is s or p-polarized. The connection between the reflected and transmitted fields and the incident field is given by the Fresnel coefficients (3.18)–(3.21). The fraction of the incident power going into the reflected or transmitted beams is given by either R_s and T_s or R_p and T_p, depending on the polarization of the incident light (see (3.22) and (3.25)).

In this chapter we consider the overall transmission and reflection through two parallel interfaces, where a layer of a third material is inserted between the initial and final materials. This situation occurs frequently in optics. For example, lenses are often coated with a thin layer of material in an effort to reduce reflections. A metal mirror usually has a thin oxide layer or a protective coating between the metal and the air.

Section 6.2 introduces the general formalism for the double boundary problem. In section 6.3 the results are manipulated into an easier-to-interpret form, valid as long as the critical angle for total internal reflection is not exceeded at the first interface. In section 6.4 we examine the "tunneling" of evanescent waves across a gap between two parallel surfaces when the critical angle for total internal reflection is exceeded.

The formalism we develop for the double-boundary problem is useful for describing a simple instrument called a *Fabry-Perot etalon* (or *interferometer* if the instrument has the capability of variable spacing between the two surfaces). The Fabry-Perot etalon, which is useful for distinguishing closely spaced wavelengths, is constructed from two partially reflective surfaces separated by a fixed distance.

Beginning in section 6.8, we study multilayer coatings, where an arbitrary number of interfaces exist between many material layers. Multilayers are often used to make highly reflective mirror coatings from dielectric materials (as opposed to metallic materials). Such mirror coatings can reflect with efficiencies greater than 99.9% at certain wavelengths. In contrast, metallic mirrors typically reflect with $\sim 96\%$ efficiency, which can be a significant loss if there are many mirrors in an optical system. Dielectric multilayer coatings also have the advantage of being more durable and harder to damage with high-intensity lasers.

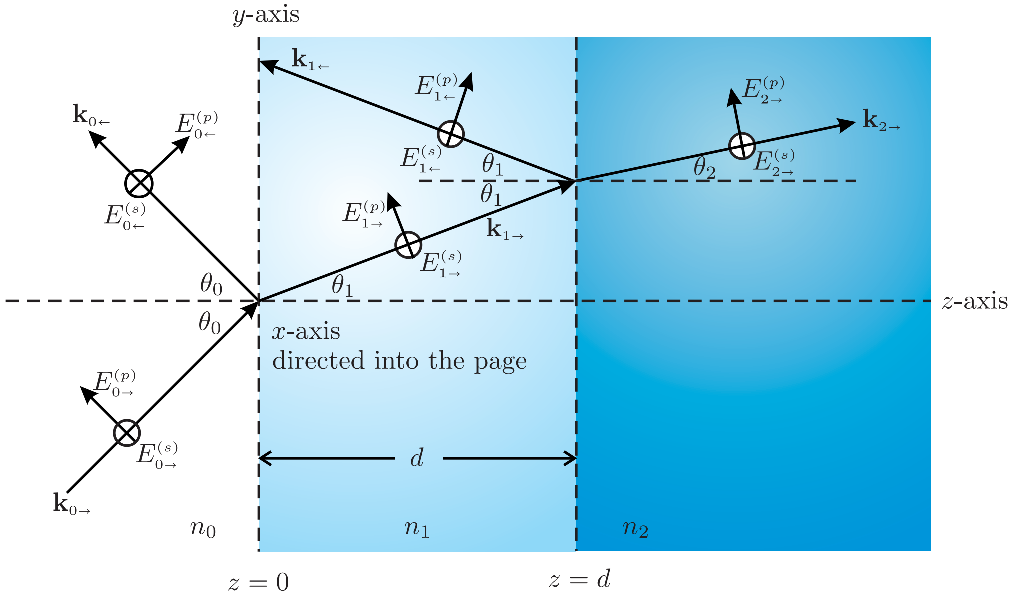


Figure 6.1 Waves propagating through a dual interface between materials.

6.2 Double Boundary Problem Solved Using Fresnel Coefficients

Consider a slab of material sandwiched between two other materials as depicted in Fig. 6.1. Because there are multiple reflections inside the middle layer, we have dropped the subscripts i, r, and t used in chapter 3 and instead use the symbols → and ← to indicate forward and backward traveling waves, respectively. Let n_1 stand for the refractive index of the middle layer. In preparation for our treatment of many-layer systems, we use n_0 and n_2 to represent the indices of the other two regions. For simplicity, we assume that indices are real. As with the single-boundary problem, we are interested in finding the transmitted fields $E^{(s)}_{2\rightarrow}$ and $E^{(p)}_{2\rightarrow}$ in terms of the incident fields $E^{(s)}_{0\rightarrow}$ and $E^{(p)}_{0\rightarrow}$. Similarly, we can also find the reflected fields $E^{(s)}_{0\leftarrow}$ and $E^{(p)}_{0\leftarrow}$ in terms the incident fields $E^{(s)}_{0\rightarrow}$ and $E^{(p)}_{0\rightarrow}$.

Both forward and backward-traveling plane waves exist in the middle material. Our intuition rightly tells us that in this region there are many reflections, bouncing both forward and backwards between the two surfaces. It might therefore seem that there should be an infinite number of fields represented, each corresponding to a different bounce. Fortunately, the forward-traveling plane waves arising from the many bounces in the middle layer all travel in the same direction. Similarly, the backwards-traveling plane waves arising from the many bounces travel in a single direction. Hence, these many fields join neatly into a net forward-moving and a net backwards-moving plane wave field.

As of yet, we do not know the amplitudes and phases of the two resulting plane waves in the middle layer, but we can denote them by $E^{(s)}_{1\rightarrow}$ and $E^{(s)}_{1\leftarrow}$ or by $E^{(p)}_{1\rightarrow}$ and $E^{(p)}_{1\leftarrow}$, separated into their s or p-components, as usual. Similarly, $E^{(s)}_{0\leftarrow}$ and $E^{(p)}_{0\leftarrow}$ as well as $E^{(s)}_{2\rightarrow}$ and $E^{(p)}_{2\rightarrow}$ are understood to include all fields which "leak" through the surfaces on each of the repeated bounces. All of these are included in the overall reflection and transmission of the fields. Thus, we need not concern ourselves with the infinite number of plane wave fields arising from the many bounces; we need only consider the five plane waves depicted in Fig. 6.1.

The fields at the boundaries are connected via the Fresnel coefficients (3.18)–(3.21), which are direct consequences of Maxwell's equations. At the first surface we define

$$\begin{aligned}
r_s^{0\to1} &\equiv \frac{\sin\theta_1\cos\theta_0 - \sin\theta_0\cos\theta_1}{\sin\theta_1\cos\theta_0 + \sin\theta_0\cos\theta_1}\\
t_s^{0\to1} &\equiv \frac{2\sin\theta_1\cos\theta_0}{\sin\theta_1\cos\theta_0 + \sin\theta_0\cos\theta_1}\\
r_p^{0\to1} &\equiv \frac{\cos\theta_1\sin\theta_1 - \cos\theta_0\sin\theta_0}{\cos\theta_1\sin\theta_1 + \cos\theta_0\sin\theta_0}\\
t_p^{0\to1} &\equiv \frac{2\cos\theta_0\sin\theta_1}{\cos\theta_1\sin\theta_1 + \cos\theta_0\sin\theta_0}
\end{aligned} \tag{6.1}$$

The notation 0 → 1 indicates the first surface from the perspective of starting on the incident side and propagating towards the middle layer. The coefficients (6.1) are written as though the problem involves only a single interface. They do not take into account any "feedback" from the second surface.

Similarly, the single-boundary Fresnel coefficients for light approaching the first interface from within the middle layer are

$$\begin{aligned}
r_s^{1\to0} &= -r_s^{0\to1}\\
t_s^{1\to0} &\equiv \frac{2\sin\theta_0\cos\theta_1}{\sin\theta_0\cos\theta_1 + \sin\theta_1\cos\theta_0}\\
r_p^{1\to0} &= -r_p^{0\to1}\\
t_p^{1\to0} &\equiv \frac{2\cos\theta_1\sin\theta_0}{\cos\theta_0\sin\theta_0 + \cos\theta_1\sin\theta_1}
\end{aligned} \tag{6.2}$$

The notation 1 → 0 indicates connections at the first interface, but from the perspective of beginning inside the middle layer. Finally, the single-boundary coefficients for light approaching the second interface are

$$\begin{aligned}
r_s^{1\to2} &\equiv \frac{\sin\theta_2\cos\theta_1 - \sin\theta_1\cos\theta_2}{\sin\theta_2\cos\theta_1 + \sin\theta_1\cos\theta_2}\\
t_s^{1\to2} &\equiv \frac{2\sin\theta_2\cos\theta_1}{\sin\theta_2\cos\theta_1 + \sin\theta_1\cos\theta_2}\\
r_p^{1\to2} &\equiv \frac{\cos\theta_2\sin\theta_2 - \cos\theta_1\sin\theta_1}{\cos\theta_2\sin\theta_2 + \cos\theta_1\sin\theta_1}\\
t_p^{1\to2} &\equiv \frac{2\cos\theta_1\sin\theta_2}{\cos\theta_2\sin\theta_2 + \cos\theta_1\sin\theta_1}
\end{aligned} \tag{6.3}$$

The notation 1 → 2 indicates connections made at the second interface from the perspective of beginning in the middle layer.

Our task is to connect the five plane waves depicted in Fig. 6.1 using the various Fresnel coefficients (6.1)–(6.3). For simplicity, we will consider s-polarized light, but the analysis can be extended to p-polarized light simply by changing the subscripts in the derivation. We begin at the second interface, which looks like a single-boundary problem (i.e. only one plane wave on the transmitted side). The field $E_{1\to}^{(s)}$ represents the forward-traveling field of the middle region evaluated at the origin $(y, z) = (0, 0)$, which we arbitrarily define to be located at the *first interface*. At the second interface, the forward traveling wave is given by $E_{1\rightarrow}^{(s)} e^{i\mathbf{k}_{1\rightarrow}\cdot\mathbf{r}}$, where $\mathbf{r} = \hat{\mathbf{z}}d$ and $\mathbf{k}_{1\rightarrow} = k_1(\hat{\mathbf{y}}\sin\theta_1 + \hat{\mathbf{z}}\cos\theta_1)$. The transmitted field in the third medium is related to the forward-traveling field of the middle region via

$$E_{2\rightarrow}^{(s)} = t_s^{1\rightarrow 2} E_{1\rightarrow}^{(s)} e^{ik_1 d\cos\theta_1} \tag{6.4}$$

where we have adjusted the phase of the field in (6.4) by $\mathbf{k}_{1\rightarrow}\cdot\mathbf{r} = k_1 d\cos\theta_1$.

Keep in mind that (6.4) represents the connection made at the point $(y, z) = (0, d)$ on the *second interface*. In the case of the transmitted field, we let $E_{2\rightarrow}^{(s)}$ stand for the transmitted field at the point $(y, z) = (0, d)$; its phase is built into its definition. The factor $t_s^{1\rightarrow 2}$ is the single-boundary Fresnel transmission coefficient at the interface (6.3), and we have used it in a manner consistent with our previous analysis in chapter 3.

We have written (6.4) for s-polarized light. The equation looks the same for p-polarized light; just replace the subscript s with p. Through the remainder of this section and the next, we will continue to economize by writing the equations only for s-polarized light with the understanding that they apply equally well to p-polarized light.

The backward-traveling plane wave in the middle region arises from the reflection of the forward-traveling plane wave in that same region. In this case, the connection using the appropriate Fresnel coefficient gives

$$E_{1\leftarrow}^{(s)} e^{-ik_1 d\cos\theta_1} = r_s^{1\rightarrow 2} E_{1\rightarrow}^{(s)} e^{ik_1 d\cos\theta_1} \tag{6.5}$$

Here again we have chosen to let $E_{0\leftarrow}^{(s)}$ represent a plane wave field referenced to the origin $(y, z) = (0, 0)$. Therefore, the factor $e^{-ik_1 d\cos\theta_1}$ is needed at $(y, z) = (0, d)$ (i.e. $\mathbf{r} = \hat{\mathbf{z}}d$) since the $\mathbf{k}$-vector for the reverse-traveling field in the middle region is $\mathbf{k}_{1\leftarrow} = k_1(\hat{\mathbf{y}}\sin\theta_1 - \hat{\mathbf{z}}\cos\theta_1)$.

We next connect the two plane waves in the middle region with the incident plane wave. In this case we must simultaneously connect $E_{1\rightarrow}^{(s)}$ with both $E_{0\rightarrow}^{(s)}$ and $E_{1\leftarrow}^{(s)}$ since they each give a contribution:

$$E_{1\rightarrow}^{(s)} = t_s^{0\rightarrow 1} E_{0\rightarrow}^{(s)} + r_s^{1\rightarrow 0} E_{1\leftarrow}^{(s)} \tag{6.6}$$

Since all fields in (6.6) are evaluated at the origin $(y, z) = (0, 0)$, there is no need for any phase factors like in (6.4) or (6.5). The relation (6.6) shows that the forward traveling wave in the middle region arises from both a transmission of the incident wave and a reflection of the backwards-traveling wave in the middle region. (We could also write an expression involving the overall reflected field $E_{0\leftarrow}^{(s)}$, but we refrain.) In summary, we have used the single-boundary Fresnel coefficients to construct the necessary connections in the double-boundary problem.

We next solve (6.4)–(6.6) to find the final transmitted field in terms of the incident field. We do this by eliminating $E_{1\rightarrow}^{(s)}$ and $E_{1\leftarrow}^{(s)}$ from the expressions. Equation (6.4) can be inverted as follows:

$$E_{1\rightarrow}^{(s)} = \frac{E_{2\rightarrow}^{(s)}}{t_s^{1\rightarrow 2} e^{ik_1 d\cos\theta_1}} \tag{6.7}$$

When this is substituted into (6.5), we obtain

$$E_{1\leftarrow}^{(s)} = \frac{r_s^{1\rightarrow 2} e^{ik_1 d\cos\theta_1}}{t_s^{1\rightarrow 2}} E_{2\rightarrow}^{(s)} \tag{6.8}$$

Substitution of (6.7) and (6.8) into (6.6) yields

$$\frac{E_{2\rightarrow}^{(s)}}{t_s^{1\rightarrow 2}e^{ik_1 d\cos\theta_1}} = t_s^{0\rightarrow 1}E_{0\rightarrow}^{(s)} + r_s^{1\rightarrow 0}\frac{r_s^{1\rightarrow 2}e^{ik_1 d\cos\theta_1}}{t_s^{1\rightarrow 2}}E_{2\rightarrow}^{(s)} \tag{6.9}$$

This can be simplified to

$$\frac{E_{2\rightarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = \frac{t_s^{0\rightarrow 1}t_s^{1\rightarrow 2}}{e^{-ik_1 d\cos\theta_1} - r_s^{1\rightarrow 0}r_s^{1\rightarrow 2}e^{ik_1 d\cos\theta_1}} \tag{6.10}$$

where the factor

$$k_1 d\cos\theta_1 = \frac{2\pi n_1 d\cos\theta_1}{\lambda_{\text{vac}}} \tag{6.11}$$

represents the phase acquired by either plane wave in traversing the middle region (see (2.24) and (2.26)).

Actually, we are mainly interested in the fraction of the power that emerges through the final surface. As in (3.29), the fraction of power transmitted is given by

$$T_s^{\text{tot}} = \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\left|\frac{E_{2\rightarrow}^{(s)}}{E_{0\rightarrow}^{(s)}}\right|^2 \qquad (\theta_2 \text{ real}). \tag{6.12}$$

Of course the relationship

$$T_s^{\text{tot}} + R_s^{\text{tot}} = 1 \tag{6.13}$$

still applies, but it is convenient for us to compute T_s^{tot} directly through (6.12) instead of indirectly from R_s^{tot}.

When the transmitted angle θ_2 is real, we may write the fraction of the transmitted power as

$$T_s^{\text{tot}} = \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\rightarrow 1}\right|^2\left|t_s^{1\rightarrow 2}\right|^2}{\left|e^{-ik_1 d\cos\theta_1} - r_s^{1\rightarrow 0}r_s^{1\rightarrow 2}e^{ik_1 d\cos\theta_1}\right|^2} \qquad (\theta_2 \text{ real}) \tag{6.14}$$

in accordance with (6.10) and (6.12). As was mentioned, (6.14) applies equally well to p-polarized light (just change the subscripts). Equation (6.14) is valid also even if the angle θ_1 is complex. Thus, it can be applied to the case of evanescent waves "tunneling" through a gap where θ_0 is beyond the critical angle for total internal reflection from the middle layer. This will be studied further in section 6.4. Note that even if θ_1 is complex, the angle θ_2 is still real if the critical angle *in the absence of the middle layer* is not exceeded.

6.3 Double Boundary Problem at Sub Critical Angles

In the case that θ_1 is real (or in other words when $\cos\theta_1$ is real so that no evanescent wave is to be considered), we may simplify (6.14) as follows:

$$\begin{aligned}T_s^{\text{tot}} &= \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{\left(e^{-ik_1d\cos\theta_1} - r_s^{1\to0}r_s^{m\to t}e^{ik_1d\cos\theta_1}\right)\left(e^{ik_1d\cos\theta_1} - \left(r_s^{1\to0}\right)^*\left(r_s^{1\to2}\right)^*e^{-ik_1d\cos\theta_1}\right)}\\ &= \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{1+\left|r_s^{1\to0}\right|^2\left|r_s^{1\to2}\right|^2 - 2\text{Re}\left\{r_s^{1\to0}r_s^{1\to2}e^{2ik_1d\cos\theta_1}\right\}}\\ &= \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{1+\left|r_s^{1\to0}\right|^2\left|r_s^{1\to2}\right|^2 - 2\text{Re}\left\{\left|r_s^{1\to0}\right|e^{i\delta_{r_s^{1\to0}}}\left|r_s^{1\to2}\right|e^{i\delta_{r_s^{1\to2}}}e^{2ik_1d\cos\theta_1}\right\}}\\ &= \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{1+\left|r_s^{1\to0}\right|^2\left|r_s^{1\to2}\right|^2 - 2\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|\cos\left(\delta+\delta_{r_s}\right)} \qquad (\theta_2 \text{ and } \theta_1 \text{ real})\end{aligned} \tag{6.15}$$

On the last line we have introduced the definitions

$$\delta \equiv 2k_1d\cos\theta_1 \tag{6.16}$$

and

$$\delta_{r_s} \equiv \delta_{r_s^{1\to0}} + \delta_{r_s^{1\to2}} \tag{6.17}$$

The phase terms $\delta_{r_s^{1\to0}}$ and $\delta_{r_s^{1\to2}}$ are defined indirectly and may be extracted from the relationships

$$r_s^{1\to0} = \left|r_s^{1\to0}\right|e^{i\delta_{r_s^{1\to0}}} \tag{6.18}$$

and

$$r_s^{1\to2} = \left|r_s^{1\to2}\right|e^{i\delta_{r_s^{1\to2}}} \tag{6.19}$$

We can continue our simplification of (6.15) by using the following identity:

$$\cos\Phi = 1 - 2\sin^2\frac{\Phi}{2} \tag{6.20}$$

where $\Phi \equiv \delta + \delta_{r_s}$. With this, (6.15) can be written as

$$\begin{aligned}T_s^{\text{tot}} &= \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{1+\left|r_s^{1\to0}\right|^2\left|r_s^{1\to2}\right|^2 - 2\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|\left[1-2\sin^2\left(\frac{\Phi}{2}\right)\right]}\\ &= \frac{n_2\cos\theta_2}{n_0\cos\theta_0}\frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{\left(1-\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|\right)^2 + 4\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|\sin^2\left(\frac{\Phi}{2}\right)}\\ &= \frac{T_s^{\max}}{1+F_{\text{s}}\sin^2\left(\frac{\Phi}{2}\right)} \qquad (\theta_2 \text{ and } \theta_1 \text{ real})\end{aligned} \tag{6.21}$$

where

$$T_s^{\max} \equiv \frac{n_2\cos\theta_2\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{n_0\cos\theta_0\left(1-\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|\right)^2}, \tag{6.22}$$

$$F_s \equiv \frac{4\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|}{\left(1-\left|r_s^{1\to0}\right|\left|r_s^{1\to2}\right|\right)^2} \tag{6.23}$$

The quantity $T_s^{\max}$ is the maximum possible transmittance of power through the surfaces, and F_{s} is called the *coefficient of finesse* (not to be confused with *reflecting finesse* discussed in section 6.7), which determines how strongly the transmittance is influenced by varying the spacing d or the wavelength λ_{vac} (causing Φ to vary).

The maximum transmittance $T_s^{\max}$ can be manipulated as follows:

$$T_s^{\max} = \frac{\frac{n_1 \cos\theta_1}{n_0 \cos\theta_0} \left|t_s^{0\to1}\right|^2 \frac{n_2 \cos\theta_2}{n_1 \cos\theta_1} \left|t_s^{1\to2}\right|^2}{\left(1 - \left|r_s^{1\to0}\right| \left|r_s^{1\to2}\right|\right)^2} = \frac{T_s^{0\to1} T_s^{1\to2}}{\left(1 - \sqrt{R_s^{1\to0} R_s^{\mathrm{m}\to\mathrm{t}}}\right)^2} \tag{6.24}$$

where we have introduced the familiar single-boundary reflectance and transmittance of the power at each of the interfaces. Similarly, we can simplify the expression for the finesse coefficient:

$$F_{\mathrm{s}} = \frac{4\sqrt{R_s^{1\to0} R_s^{1\to2}}}{\left(1 - \sqrt{R_s^{\mathrm{m}\to\mathrm{i}} R_s^{1\to2}}\right)^2} \tag{6.25}$$

Please note that $R_s^{1\to0} = R_s^{0\to1}$, as verified from (6.2). Again, although the above equations have been written expressly for s-polarized light, they can be used for p-polarized light by changing all subscripts to p.

Example 6.1

You desire to make a "beam splitter" for s-polarized light as shown in Fig. 6.2 by coating a piece of glass ($n = 1.5$) with a thin film of zinc sulfide ($n = 2.32$). The idea is to get about half of the light to reflect from the front of the glass. An anti-reflection coating is applied to the back surface of the glass. The light is incident at 45° as shown in Fig. 6.2.

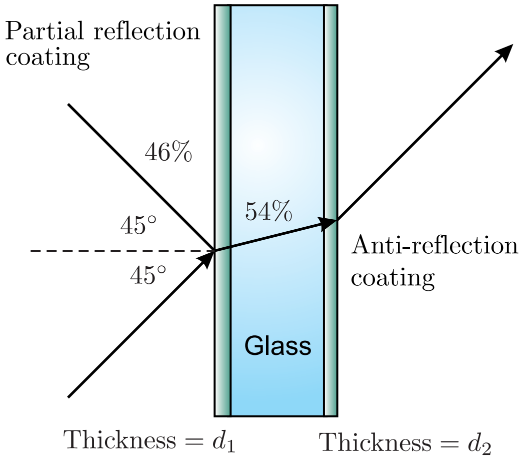


Figure 6.2

Find the highest transmittance possible through an antireflection film of magnesium fluoride ($n = 1.38$) at the back surface of the "beam splitter." Find the smallest possible d_2 that accomplishes this for light with wavelength $\lambda_{\mathrm{vac}} = 633$ nm. (In P 6.3 you will consider the reflection from the front coating.)

NOTE: Since the antireflection films are usually imperfect, beam splitter substrates are often slightly wedged so that unwanted reflections from the second surface exit in a different direction.

Solution:

We have

$$
\begin{aligned}
&n_0 = 1.5\\
&n_1 = 1.38\\
&n_2 = 1\\
&\theta_2 = 45^\circ\\
&n_1 \sin\theta_1 = \sin\theta_2 \Rightarrow \theta_1 = \sin^{-1}\left(\frac{\sin 45^\circ}{1.38}\right) = 30.82^\circ\\
&r_s^{1\to2} = -\frac{\sin(\theta_1-\theta_2)}{\sin(\theta_1+\theta_2)} = -\frac{\sin(30.82^\circ - 45^\circ)}{\sin(30.82^\circ + 45^\circ)} = 0.253\\
&n_0 \sin\theta_0 = \sin\theta_2 \Rightarrow \theta_0 = \sin^{-1}\left(\frac{\sin 45^\circ}{1.5}\right) = 28.13^\circ\\
&r_s^{1\to0} = -\frac{\sin(\theta_1-\theta_0)}{\sin(\theta_1+\theta_0)} = -\frac{\sin(30.82^\circ - 28.13^\circ)}{\sin(30.82^\circ + 28.13^\circ)} = -0.0549\\
&R_s^{1\to0} \equiv |-0.0549|^2 = 0.0030\\
&R_s^{1\to2} \equiv |0.253|^2 = 0.0640\\
&T_s^{0\to1} = T_s^{1\to0} = 1 - R_s^{1\to0} = 1 - 0.0030 = 0.997\\
&T_s^{1\to2} = 1 - R_s^{1\to2} = 1 - 0.0640 = 0.936\\
&\delta_{r_s} \equiv \delta_{r_s^{1\to0}} + \delta_{r_s^{1\to2}} = \pi + 0 = \pi\\
&F = \frac{4\sqrt{R_{1\to0}R_{1\to2}}}{\left(1-\sqrt{R_{1\to0}R_{1\to2}}\right)^2} = \frac{4\sqrt{(0.0030)(0.0640)}}{\left(1-\sqrt{(0.0030)(0.0640)}\right)^2} = 0.0570\\
&T_s^{\max} = \frac{T_s^{0\to1}T_s^{1\to2}}{\left(1-\sqrt{R_s^{1\to0}R_s^{1\to2}}\right)^2} = \frac{(0.997)(0.936)}{\left(1-\sqrt{(0.0030)(0.0640)}\right)^2} = 0.960\\
&T_s^{\text{tot}} = \frac{0.960}{1+0.0570\sin^2\left(\frac{\delta+\pi}{2}\right)}
\end{aligned}
$$

The maximum transmittance occurs when $\sin^2\left(\frac{\delta+\pi}{2}\right) = 0$. In that case, $T_{\text{tot}} = 0.960$, meaning that 96% of the light is transmitted.

$$
\delta + \pi = 2k_1 d_2 \cos\theta_1 + \pi = 2\pi \Rightarrow d_2 = \frac{\lambda_{\text{vac}}}{4n_1\cos\theta_1} = \frac{633\text{ nm}}{4(1.38)\cos 30.82^\circ} = 134\text{ nm}
$$

Without the coating, (i.e. $d_2 = 0$), the transmittance through the antireflection coating would be 0.908, so the coating does give an improvement.

6.4 Beyond Critical Angle: Tunneling of Evanescent Waves

The formula (6.14) for the transmittance holds, even if the middle angle θ_1 doesn't exist in a physical sense (i.e. if it is complex). We can use (6.14) to describe frustrated total internal reflection where θ_0 and θ_2 exceed the critical angle. In this case an evanescent wave occurs in the middle region. If the second surface is brought close to the first and the spacing between the two surfaces is small enough, the evanescent wave stimulates the second surface and a transmitted wave results. It is often inconvenient to deal with a complex angle θ_1

when calculating the single-boundary Fresnel coefficients, so we rewrite $\sin\theta_1$ using Snell's law:

$$\sin\theta_1 = \frac{n_0}{n_1}\sin\theta_0 = \frac{n_2}{n_1}\sin\theta_2 \tag{6.26}$$

and $\cos\theta_1$ as

$$\cos\theta_1 = i\sqrt{\sin^2\theta_1 - 1} \tag{6.27}$$

Note that beyond the critical angle, $\sin\theta_1$ is greater than one.

Example 6.2

Calculate the transmittance of p-polarized light through the region between two closely spaced 45° right prisms as a function of the vacuum wavelength λ_{vac} and the prism spacing d, as shown in Fig. 6.3 (see P 6.4 for the s-polarized case). Take the index of refraction of the prisms to be $n = 1.5$, surrounded by index $n = 1$ and use $\theta_0 = \theta_2 = 45°$. Neglect possible reflections from the exterior surfaces of the prisms.

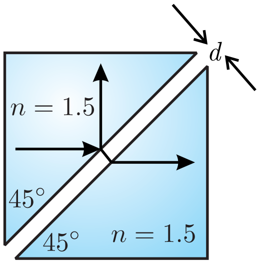


Figure 6.3 Frustrated total internal reflection in two prisms.

Solution: First we must compute the Fresnel coefficients appearing in (6.14). From (6.1)–(6.3) we compute the various necessary Fresnel coefficients, using (6.26) and (6.27) to handle the complex angles:

$$\begin{aligned}\left|t_p^{0\to1}\right|^2 &= \left|\frac{2\cos\theta_0\sin\theta_1}{\cos\theta_1\sin\theta_1 + \cos\theta_0\sin\theta_0}\right|^2 \\ &= \left|\frac{2\cos\theta_0(n\sin\theta_0)}{\left(i\sqrt{n^2\sin^2\theta_0 - 1}\right)(n\sin\theta_0) + \cos\theta_0\sin\theta_0}\right|^2 = 5.76\end{aligned} \tag{6.28}$$

$$\begin{aligned}\left|t_p^{1\to2}\right|^2 &= \left|\frac{2\cos\theta_1\sin\theta_2}{\cos\theta_2\sin\theta_2 + \cos\theta_1\sin\theta_1}\right|^2 \\ &= \left|\frac{2\left(i\sqrt{n^2\sin^2\theta_2 - 1}\right)\sin\theta_2}{\cos\theta_2\sin\theta_2 + \left(i\sqrt{n^2\sin^2\theta_0 - 1}\right)(n\sin\theta_0)}\right|^2 = 0.64\end{aligned} \tag{6.29}$$

$$\begin{aligned}r_p^{1\to2} = r_p^{1\to0} = -r_p^{0\to1} &= -\frac{\cos\theta_1\sin\theta_1 - \cos\theta_0\sin\theta_0}{\cos\theta_1\sin\theta_1 + \cos\theta_0\sin\theta_0} \\ &= -\frac{\left(i\sqrt{n^2\sin^2\theta_0 - 1}\right)(n\sin\theta_0) - \cos\theta_0\sin\theta_0}{\left(i\sqrt{n^2\sin^2\theta_0 - 1}\right)(n\sin\theta_0) + \cos\theta_0\sin\theta_0} \\ &= e^{-i1.287}\end{aligned} \tag{6.30}$$

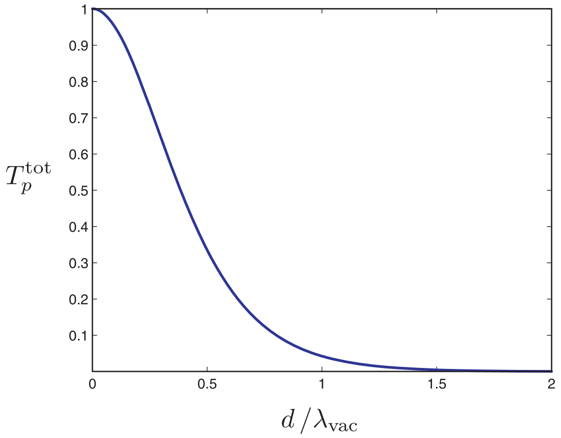


Figure 6.4 Transmittance of p-polarized light through a gap between two 45° prisms with $n = 1.5$ as the gap width is varied (Example 6.2).

We also need

$$\begin{aligned} k_1 d\cos\theta_1 &= \frac{2\pi}{\lambda_{\text{vac}}} d\cos\theta_1 \\ &= 2\pi\left(i\sqrt{n^2\sin^2\theta_0 - 1}\right)\frac{d}{\lambda_{\text{vac}}} \\ &= i2.22\left(\frac{d}{\lambda_{\text{vac}}}\right) \end{aligned} \tag{6.31}$$

Now we are ready to compute the net transmittance (6.14). Since $\theta_0 = \theta_2$ and $n_0 = n_2$, we have

$$\begin{aligned} T_p^{\text{tot}} &= \frac{\left|t_s^{0\to1}\right|^2\left|t_s^{1\to2}\right|^2}{\left|e^{-ik_1 d\cos\theta_1} - r_s^{1\to0} r_s^{1\to2} e^{ik_1 d\cos\theta_1}\right|^2} \\ &= \frac{(5.76)(0.64)}{\left|e^{-i\left[i2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)\right]} - e^{-i1.287}e^{-i1.287}e^{i\left[i2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)\right]}\right|^2} \\ &= \frac{3.69}{\left(e^{2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)} - e^{-2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)-i2.574}\right)\left(e^{2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)} - e^{-2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)+i2.574}\right)} \\ &= \frac{3.69}{e^{4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + e^{-4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} - 2\left(\frac{e^{i2.574}+e^{-i2.574}}{2}\right)} \\ &= \frac{3.69}{e^{4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + e^{-4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} - 2\cos(2.574)} \\ &= \frac{3.69}{e^{4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + e^{-4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + 1.69} \end{aligned} \tag{6.32}$$

Figure 6.4 shows a plot of the transmittance (6.32) calculated in Example 6.2. Notice that the transmittance goes to one as expected when the two prisms are brought together: $T_p^{\text{tot}}(d/\lambda_{\text{vac}} = 0) = 1$. When the prisms get to be about a wavelength apart, the transmittance is significantly reduced, and as the distance gets large compared to a wavelength, the transmittance quickly goes to zero ($T_p^{\text{tot}}(d/\lambda_{\text{vac}} \gg 1) \approx 0$).

6.5 Fabry-Perot

Marie Paul Auguste Charles Fabry (1867-1945) and Jean Baptiste Gaspard Gustave Alfred Perot (1863-1925) realized that a double interface could be used to distinguish wavelengths of light that are very close together. The Fabry-Perot instrument consists simply of two identical (parallel) surfaces separated by spacing d. Our analysis in section 6.3 applies. For simplicity, we choose the refractive index before the initial surface and after the final surface to be the same (i.e. $n_0 = n_2$). We assume that the transmission angles are such that total internal reflection is avoided. Whether the double-boundary setup transmits light well or poorly depends on the exact spacing between the two boundaries and on the reflectivity of the surfaces, as well as on the wavelength of the light.

If the spacing d separating the two parallel surfaces is adjustable (scanned), the instrument is called a *Fabry-Perot interferometer*. If the spacing is fixed while the angle of the incident light is varied, the instrument is called a *Fabry-Perot etalon*. An etalon can therefore be as simple as a piece of glass with parallel surfaces. Sometimes, a thin optical membrane called a *pellicle* is used as an etalon (occasionally inserted into laser cavities to discriminate against certain wavelengths). However, to achieve sharp discrimination between closely-spaced wavelengths, a large spacing d is desirable. The two surfaces should also reflect relatively well, much better than, say, a simple air-glass interface.

As we previously derived (6.21), the transmittance through a double boundary is

$$T^{\text{tot}} = \frac{T^{\max}}{1 + F\sin^2\left(\frac{\Phi}{2}\right)} \tag{6.33}$$

In the case of identical interface on the incident and transmitted sides, the transmittance and reflecance coefficients are the same at each surface (i.e. $T = T^{0\to1} = T^{1\to2}$ and $R = R^{1\to0} = R^{1\to2}$). In this case, the maximum transmittance and the finesse coefficient are

$$T^{\max} = \frac{T^2}{(1-R)^2} \tag{6.34}$$

and

$$F = \frac{4R}{(1-R)^2} \tag{6.35}$$

In principle, these equations should be evaluated for either s or p-polarized light. However, a Fabry-Perot interferometer or etalon is usually operated near normal incidence so that there is little difference between the two polarizations.

When using a Fabry-Perot instrument, one observes the transmittance T^{tot} as the parameter Φ is varied (see (6.16) and (6.20)). The parameter Φ can be varied by altering d, θ_1, or λ as prescribed by

$$\Phi = \frac{4\pi n_1 d}{\lambda_{\text{vac}}}\cos\theta_1 + \delta_{\text{r}} \tag{6.36}$$

To increase the sensitivity of the instrument, it is desirable to have the transmittance T^{tot} vary strongly when Φ is varied. By inspection of (6.33), we see that T^{tot} varies strongest if the *finesse coefficient* F is large. We achieve a large finesse coefficient by increasing the reflectance R.

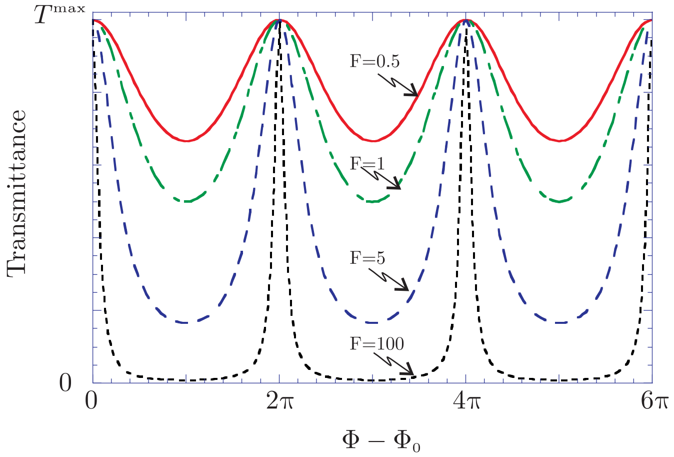


Figure 6.5 Transmittance as the phase Φ is varied. The different curves correspond to different values of the finesse coefficient. Φ_0 represents a large multiple of 2π.

The total transmittance T^{tot} (6.33) through a Fabry-Perot instrument is depicted in Fig. 6.5 as a function of Φ. The various curves correspond to different values of F.

Typical values of Φ can be extremely large. For example, suppose that the instrument is used at near-normal incidence (i.e. $\cos\theta_1 \cong 1$) with a wavelength of $\lambda_{\text{vac}} = 500$ nm and an interface separation of $d_0 = 1$ cm. From (6.36) the value of Φ (ignoring the constant phase term δ_r) is approximately

$$\Phi_0 = \frac{4\pi\,(1\text{ cm})}{500\text{ nm}} = 80,000\pi \tag{6.37}$$

As we vary d, λ, or θ_1 by small amounts, we can easily cause Φ to change by 2π as depicted in Fig. 6.5. The figure shows small changes in Φ above a value Φ_0, which represents a large multiple of 2π.

The basic setup of a Fabry-Perot instrument is shown in Fig. 6.6. In order to achieve a relatively high finesse coefficient F, we require fairly high reflectivities at the two surfaces. To accomplish this, special coatings can be applied to the surfaces, for example, a thin layer of silver (or some other coating) to achieve a partial reflection, say 90%. Typically, two glass substrates are separated by distance d, with the coated surfaces facing each other as shown in the figure. The substrates are aligned so that the interior surfaces are parallel to each other. It is typical for each substrate to be slightly wedge-shaped so that unwanted reflections from the outer surfaces do not interfere with the double boundary situation between the two plates.

Actually, each interior coating may be thought of as its own double-boundary problem (or multiple-boundary as the case may be). However, without regard for the details of the coatings, we can say that each coating has a certain overall transmittance T and a certain overall reflection R. As light goes through the coating, it can also be attenuated through absorption. Therefore, at each coating surface we have

$$R + T + A = 1 \tag{6.38}$$

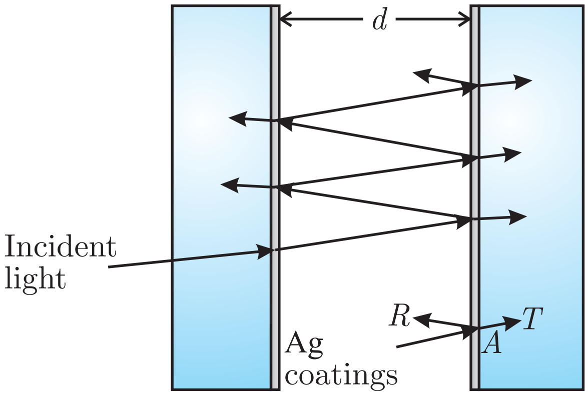


Figure 6.6 Typical Fabry-Perot setup. If the spacing d is variable, it is called an interferometer; otherwise, it is called an etalon.

where A represents the amount of light absorbed at a coating. Notice from (6.38) that when we increase the value of R, the value of T must decrease. Thus, to the extent that A is non zero, there is an apparent tradeoff between increasing the finesse coefficient F and maintaining a bright (observable) transmittance $T^{\max}$ through the instrument (see (6.34) and (6.35)). However, in Fig. 6.5, each curve is plotted in terms of its own $T^{\max}$.

The reflection phase δ_r in (6.36) depends on the exact nature of the coatings in the Fabry-Perot instrument. However, we do not need to know the value of δ_r (depending on both the complex index of the coating material and its thickness). Whatever the value of δ_r, we only care that it is constant. Experimentally, we can always compensate for the δ_r by "tweaking" the spacing d. Note that the required "tweak" on the spacing need only be a fraction of a wavelength, which is tiny when compared to the overall spacing d, typically many thousands of wavelengths.

In the next section, we examine the transmittance (6.33) in detail as the spacing d and the angle θ_1 are adjusted. We also discuss typical experimental arrangements for a Fabry-Perot interferometer or etalon. In section 6.7, we examine how a Fabry-Perot instrument is able to distinguish closely spaced wavelengths, and we will introduce the concept of *free spectral range* and *resolving power* of the instrument.

6.6 Setup of a Fabry-Perot Instrument

Figure 6.7 shows the typical experimental setup for a Fabry-Perot interferometer. A collimated beam of light is sent through the instrument. The beam is aligned so that it is normal to the surfaces. It is critical for the two surfaces of the interferometer to be very close to parallel. For initial alignment, the back-reflected beams from each surface can be monitored to ensure rough alignment. Then as fringes appear, the alignment is further adjusted until the entire transmitted beam becomes one large fringe, which blinks all together as the spacing d changes (by tiny amounts). A mechanical actuator is then used to vary the spacing between the plates, and the transmittance of the light is observed with a detector connected to an oscilloscope. The sweep of the oscilloscope must be synchronized with the period of the (oscillating) mechanical driver. To make the alignment of the instrument less

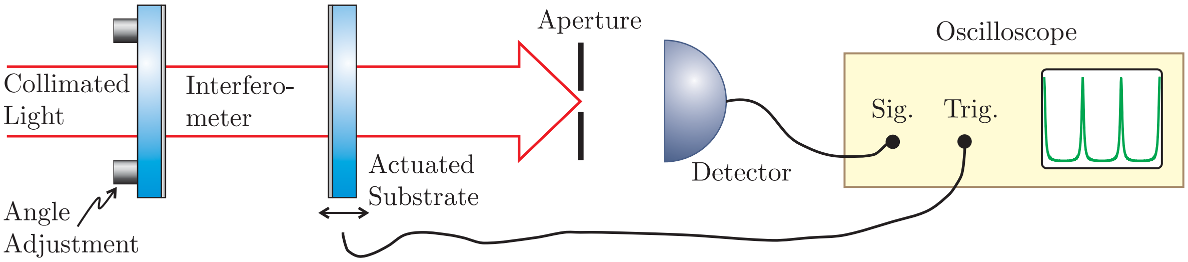


Figure 6.7 Setup for a Fabry-Perot interferometer.

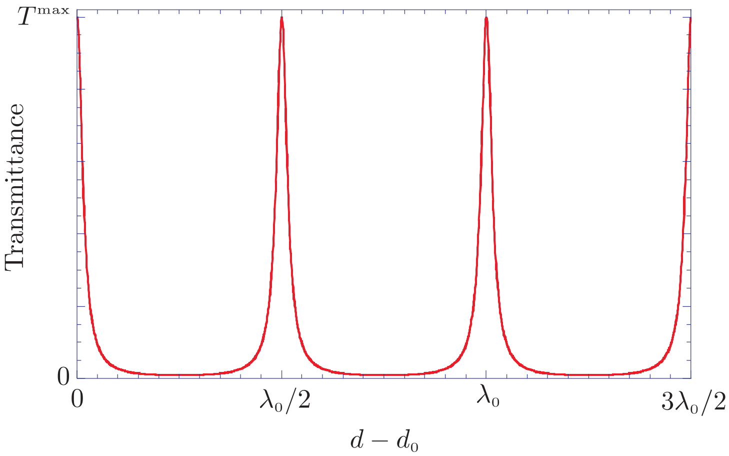


Figure 6.8 Transmittance as the separation d is varied ($F = 100$). d_0 represents a large distance for which Φ is a multiple of 2π.

critical, a small aperture can be placed in front of the detector so that it observes only a small portion of the beam.

The transmittance as a function of plate separation is shown in Fig. 6.8. In this case, Φ varies via changes in d only (see (6.36) with $\cos\theta_1 = 1$ and fixed wavelength). As the spacing is increased by only a half wavelength, the transmittance changes through a complete period. Figure 6.8 shows what is seen on an oscilloscope when the mechanical driver travels at constant velocity. The various peaks in the figure are called *fringes*.

The setup for a Fabry-Perot etalon is similar to that of the interferometer. The key difference is that the angle of the incident light is varied rather than the plate separation. One way to do this is to observe light from a "point source" which forms a conical beam that transverses the device, as depicted in Fig. 6.9. Different portions of the beam go through the device at different angles. When aligned straight on, the transmitted light forms a "bull's-eye" pattern on a screen, as will be described below. Often the two surfaces in the etalon are held parallel to each other by a precision ring spacer to eliminate the need for alignment.

In Fig. 6.10 we graph the transmittance T^{tot} (6.33) as a function of angle (holding wavelength and plate separation fixed). Since $\cos\theta_1$ is not a linear function, the spacing of the peaks varies with angle. Actually, as θ_1 increases from zero, the cosine steadily decreases, causing Φ to decrease. Each time Φ decreases by 2π we get a new peak. Not surprisingly, only a modest change in angle is necessary to cause the transmittance to vary from maximum to minimum, or vice versa. In Fig. 6.10, we have again assumed $\lambda_{\text{vac}} = 500$ nm and $d_0 = 1$ cm. The advantage to the Fabry-Perot etalon (as opposed to the interferometer) is that no moving parts are needed. The disadvantage is that light must be sent through the instrument at many angles to see the variation in the transmittance. The peaks in the figure are called *fringes*.

An example of the bull's-eye pattern observed with this setup is shown in Fig. 6.10(b). An increase in radius corresponds to an increase in the cone angle. Thus, the bull's-eye pattern can be understood as the curve in Fig. 6.10(a) rotated about a circle. If the wavelength or the spacing between the plates were to vary, the radii (or angles) where the fringes appear would shift accordingly. For example, the center spot could become dark.

Finally, consider the setup shown in Fig. 6.11, which is used to observe light from a diffuse source. The earlier setup shown in Fig. 6.9 won't work for a diffuse source unless all of the light is blocked except for a small "point source." This is impractical if there remains insufficient illumination at the final screen for observation. In order to preserve as much light as possible we can sandwich the etalon between two lenses. We place the diffuse source at the focal point of the first lens. We place the screen at the focal point of the second lens. This causes an image of the source to appear on the screen. (If the diffuse source has the shape of Mickey Mouse, then an image of Mickey Mouse appears on the screen.) Each point of the diffuse source is mapped to a corresponding point on the screen; the orientation of the points is preserved (albeit inverted). In addition, the light associated with any particular point of the source travels as a collimated beam in the region between the lenses. Each collimated beam traverses the etalon with a unique angle. Because of the differing angles, the light associated with each point traverses the etalon with higher or lower transmittance. The result is that the Bull's eye pattern seen in Fig. 6.10 becomes superimposed on the image of the diffuse source. One can observe the pattern directly by substituting the lens and retina of the eye for the final lens and screen.

6.7 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument

Thus far, we have examined how the transmittance through a Fabry-Perot instrument varies with surface separation d and angle θ_1. However, the main purpose of a Fabry-Perot instrument is to measure small changes in the wavelength of light, which similarly affects the value of Φ (see (6.36)).

Consider a Fabry-perot interferometer where the transmittance through the instrument is plotted as a function of surface separation d. (For purposes of the following discussion, we could have instead chosen a Fabry-Perot etalon at various transmittance angles.) Let the spacing d_0 correspond to the case when Φ is a multiple of 2π for the wavelength λ_{vac}. Next suppose we adjust the wavelength of the light from $\lambda_{\text{vac}} = \lambda_0$ to $\lambda_{\text{vac}} = \lambda_0 + \Delta\lambda$ while observing the transmittance. As we do this, the value of Φ changes. Fig. 6.12 shows what happens as we scan the spacing d of the interferometer in the neighborhood of d_0. A change in wavelength causes the position of the fringes to shift so that a peak no longer occurs

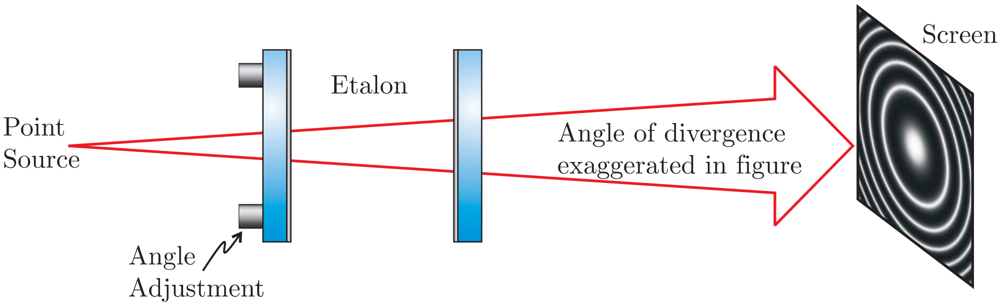


Figure 6.9 A diverging monochromatic beam traversing a Fabry-Perot etalon.

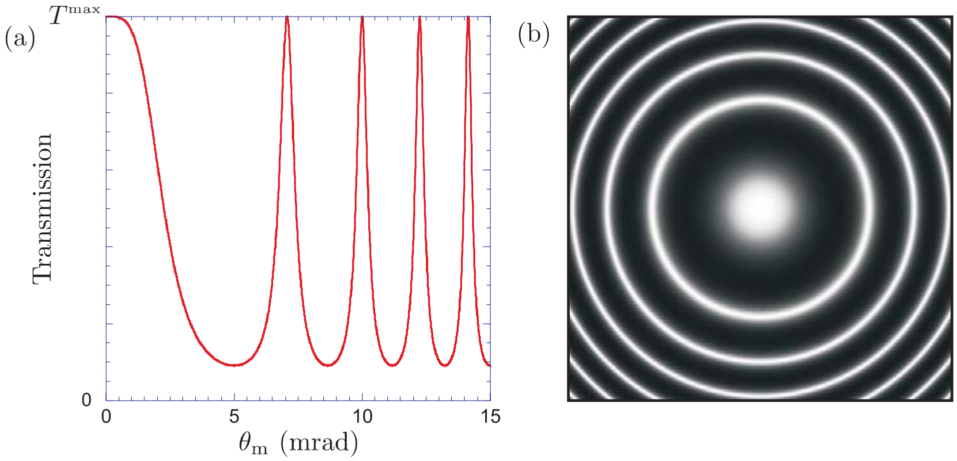


Figure 6.10 (a) Transmittance as the angle θ_1 is varied. It is assumed that the distance d is chosen such that Φ is a multiple of 2π when the angle is zero. (b) Pattern on the screen of a diverging monochromatic beam traversing a Fabry-Perot etalon with $F = 10$.

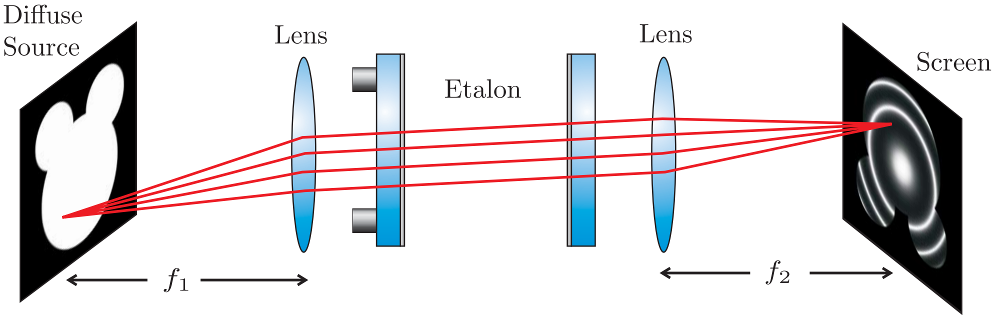


Figure 6.11 Setup of a Fabry-Perot etalon for looking at a diffuse source.

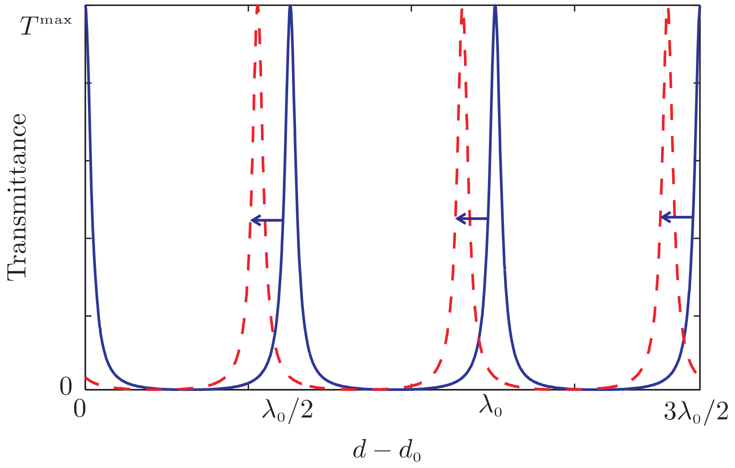


Figure 6.12 Transmittance as the spacing d is varied for two different wavelengths ($F = 100$). The solid line plots the transmittance of light with a wavelength of λ_0, and the dashed line plots the transmittance of a wavelength shorter than λ_0. Note that the fringes shift positions for different wavelengths.

when the spacing is d_0. The dashed line corresponds to a different wavelength.

We now find the connection between a change in wavelength and the amount that Φ changes, giving rise to the fringe shift seen in Fig. 6.12. Suppose that the transmittance through the Fabry-Perot instrument is maximum at the wavelength λ_0. That is, we have

$$\Phi_0 = \frac{4\pi n_1 d_0 \cos\theta_1}{\lambda_0} + \delta_{\rm r} \tag{6.39}$$

where Φ_0 is an integer multiple of 2π. Now consider what happens to Φ as the wavelength increases. At a new wavelength (all else remaining the same) we have

$$\Phi = \frac{4\pi n_1 d_0 \cos\theta_1}{\lambda_0 + \Delta\lambda} + \delta_{\rm r} \tag{6.40}$$

The change in wavelength $\Delta\lambda$ is usually very small compared to λ_0, so we can represent the denominator with the first two terms of a Taylor-series expansion:

$$\frac{1}{\lambda_0 + \Delta\lambda} = \frac{1}{\lambda_0\left(1 + \Delta\lambda/\lambda_0\right)} \cong \frac{1 - \Delta\lambda/\lambda_0}{\lambda_0} \tag{6.41}$$

Then (6.40) can be rewritten as

$$\Phi_0 - \Phi = \frac{4\pi n_1 d_0 \cos\theta_1}{\lambda_0^2}\Delta\lambda \tag{6.42}$$

Equation (6.42) enables us to compute the amount of fringe shift (like those seen in Fig. 6.12) for a given change in wavelength. Conversely, if we observe a certain shift in the location of the fringes we can say by what amount the wavelength must have changed. If the change in wavelength is enough to cause Φ to decrease by 2π, the fringes in Fig. 6.12 shift through a whole period, and the picture looks the same.

This behavior shows an important limitation of the instrument. If the fringes shift by too much, we might become confused as to whether anything has changed at all, owing to the periodic nature of the fringes. We can avoid this confusion if we are able to watch the fringes shift as we continuously vary the wavelength, but for many applications we do not have continuous control over the wavelength. For example, we may want to send simultaneously two nearby wavelengths through the instrument to make a comparison. If the wavelengths are separated by too much, we may be confused. The fringes of one wavelength may be shifted past several fringes of the other wavelength, and we will not be able to tell by how much they are different.

This introduces the concept of *free spectral range*, which is the wavelength change $\Delta\lambda_{\text{FSR}}$ that causes the fringes to shift through one period. We find this by setting (6.42) equal to 2π. After rearranging, we get

$$\Delta\lambda_{\text{FSR}} = \frac{\lambda_{\text{vac}}^2}{2n_1 d_0 \cos\theta_1} \tag{6.43}$$

If the wavelength is $\lambda_{\text{vac}} = 500$ nm and the spacing is $d_0 = 1$ cm, the free spectral range is $\Delta\lambda_{\text{FSR}} = (500 \text{ nm})^2/2(1 \text{ cm}) = 0._{0\rightarrow 1}3$ nm, assuming near normal incidence and an index $n_1 = 1$. This extremely narrow wavelength range is the widest that should be examined for the given parameters. In summary, the free spectral range is the largest change in wavelength permissible while avoiding confusion. To convert this wavelength difference $\Delta\lambda_{\text{FSR}}$ into a corresponding frequency difference, one differentiates $\omega = 2\pi c/\lambda_{\text{vac}}$ to get

$$|\Delta\omega| = \frac{2\pi c\Delta\lambda}{\lambda_{\text{vac}}^2} \tag{6.44}$$

We next consider the *smallest* change in wavelength that can be noticed, or *resolved* with a Fabry-Perot instrument. For example, if two very near-by wavelengths are sent through the instrument simultaneously, we can distinguish them only if the separation between their corresponding fringe peaks is at least as large as the width of individual peaks. This situation of two barely resolvable fringe peaks is shown on the left of Fig. 6.13. We will look for the wavelength change that causes a peak to shift by its own width.

We define the width of a peak by its *full width at half maximum* (FWHM). Again, let Φ_0 be a multiple of 2π so that a peak in transmittance occurs when $\Phi = \Phi_0$. In this case, we have from (6.33) that

$$T^{\text{tot}} = \frac{T^{\max}}{1 + F\sin^2\left(\frac{\Phi_0}{2}\right)} = T^{\max} \tag{6.45}$$

If Φ varies from Φ_0 to $\Phi_0 \pm \Phi_{\text{FWHM}}/2$, then, by definition, the transmittance drops to one half. Therefore, we may write

$$T^{\text{tot}} = \frac{T^{\max}}{1 + F\sin^2\left(\frac{\Phi_0 \pm \Phi_{\text{FWHM}}/2}{2}\right)} = \frac{T^{\max}}{2} \tag{6.46}$$

We solve (6.46) for Φ_{FWHM}, and we see that this equation requires

$$F\sin^2\left(\frac{\Phi_{\text{FWHM}}}{4}\right) = 1 \tag{6.47}$$

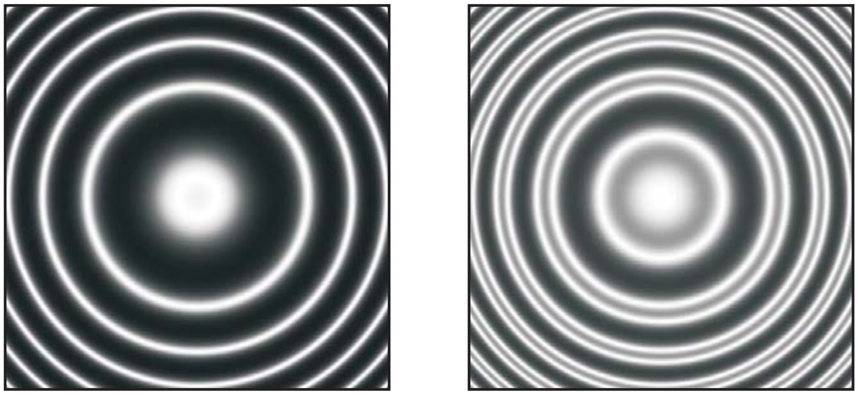


Figure 6.13 Transmittance as function of angle through a Fabry-Perot etalon. Two nearby wavelengths are sent through the instrument simultaneously, (left) barely resolved and (right) easily resolved.

where we have taken advantage of the fact that Φ_0 is a multiple of 2π. Next, we suppose that Φ_{FWHM} is rather small so that we may represent the sine by its argument. This approximation is okay if the finesse coefficient F is rather large (say, 100). With this approximation, (6.47) simplifies to

$$\Phi_{\text{FWHM}} \cong \frac{4}{\sqrt{F}}. \tag{6.48}$$

The ratio of the period between peaks 2π to the width Φ_{FWHM} of individual peaks is called the *reflecting finesse* (or just *finesse*).

$$f \equiv \frac{2\pi}{\Phi_{\text{FWHM}}} = \frac{\pi\sqrt{F}}{2} \tag{6.49}$$

This parameter is often used to characterize the performance of a Fabry-Perot instrument. Note that a higher finesse f implies sharper fringes in comparison to the fringe spacing.

Finally, we are ready to compute the minimum wavelength difference that can be resolved using the instrument. The free spectral range $\Delta\lambda_{\text{FSR}}$ compared to the minimum wavelength $\Delta\lambda_{\text{FWHM}}$ is the same as a whole period 2π compared to Φ_{FWHM}, or the reflecting finesse f. Therefore, we have

$$\Delta\lambda_{\text{FWHM}} = \frac{\Delta\lambda_{\text{FSR}}}{f} = \frac{\lambda_{\text{vac}}^2}{\pi n_1 d_0 \cos\theta_1 \sqrt{F}} \tag{6.50}$$

For $\lambda_{\text{vac}} = 500$ nm, $d_0 = 1$ cm, and $F = 100$ (again assuming near normal incidence and $n_1 = 1$), this minimum resolvable wavelength change is

$$\Delta\lambda_{\text{FWHM}} = \frac{(500\text{ nm})^2}{\pi\,(1\text{ cm})\sqrt{100}} = 0.00080\text{ nm} \tag{6.51}$$

This means that a wavelength spread of 0.00080 nm centered on $\lambda_0 = 500$ nm looks about the same in the Fabry-Perot instrument as a pure wavelength at $\lambda_0 = 500$ nm. However, a wavelength variation larger than this will be noticed.

As a final note, a common characterization of how well an instrument distinguishes close-together wavelengths is given by the ratio of λ_0 to $\Delta\lambda_{\text{min}}$, where $\Delta\lambda_{\text{min}}$ is the minimum change of wavelength that the instrument can distinguish in the neighborhood of λ_0. (We are not as impressed when $\Delta\lambda_{\text{min}}$ is small, if λ_0 also is small.) This ratio is called the *resolving power* of the instrument:

$$\text{RP} \equiv \frac{\lambda_0}{\Delta\lambda_{\text{min}}} \tag{6.52}$$

In the case of a Fabry-Perot instrument, we have $\Delta\lambda_{\text{min}} = \Delta\lambda_{\text{FWHM}}$. Fabry-Perot instruments tend to have very high resolving powers since they respond to very small differences in wavelength. When $\Delta\lambda_{\text{min}} = 0.00080$ nm and $\lambda_0 = 500$ nm, the resolving power of a Fabry-Perot instrument is an impressive RP $= 600,000$. For comparison, the resolving power of a typical grating spectrometer is much less (a few thousand). However, a spectrometer has the advantage that it can observe a much wider range of wavelengths at once (not confined within the narrow free spectral range of a Fabry-Perot instrument).

6.8 Multilayer Coatings

In this section, we generalize our previous analysis of a double interface to an arbitrary number of parallel interfaces (i.e. multilayer coatings). As we saw in section 6.3, a single coating applied to an optical surface is often insufficient to accomplish the desired effect, especially if the goal is to make a highly reflective mirror. For example, if we want to make a mirror surface using a dielectric coating (with the advantage of being less fragile and more reflective than a metal coating), a single layer is insufficient to reflect the majority of the light, even if a relatively high index is used. In P 6.3 we compute that a single dielectric layer deposited on glass can reflect at most about 46% of the light. We would like to do much better (e.g. >99%), and this can be accomplished with multilayer dielectric coatings which can have considerably better reflectivities than metal surfaces such as silver.

We now proceed to develop the formalism of the general multi-boundary problem. Rather than incorporate the single-interface Fresnel coefficients into the problem as we did in section 6.2, we return to the basic boundary conditions for the electric and magnetic fields at each interface between the layers.

We examine p-polarized light incident on an arbitrary multilayer coating (all interfaces parallel to each other). We leave it as an exercise to re-derive the formalism for s-polarized light (see P 6.11). The upcoming derivation is valid also for complex refractive indices, although our notation suggests real indices. The ability to deal with complex indices is very important if, for example, we want to make mirror coatings work in the extreme ultraviolet wavelength range where virtually every material is absorptive. Consider the diagram of a multilayer coating in Fig. 6.14 for which the angle of light propagation in each region may be computed from Snell's law:

$$n_0 \sin\theta_0 = n_1 \sin\theta_1 = \cdots = n_N \sin\theta_N = n_{N+1} \sin\theta_{N+1} \tag{6.53}$$

where N denotes the number of layers in the coating. The subscript 0 represents the initial medium outside of the multilayer, and the subscript $N+1$ represents the final material, or the substrate on which the layers are deposited.

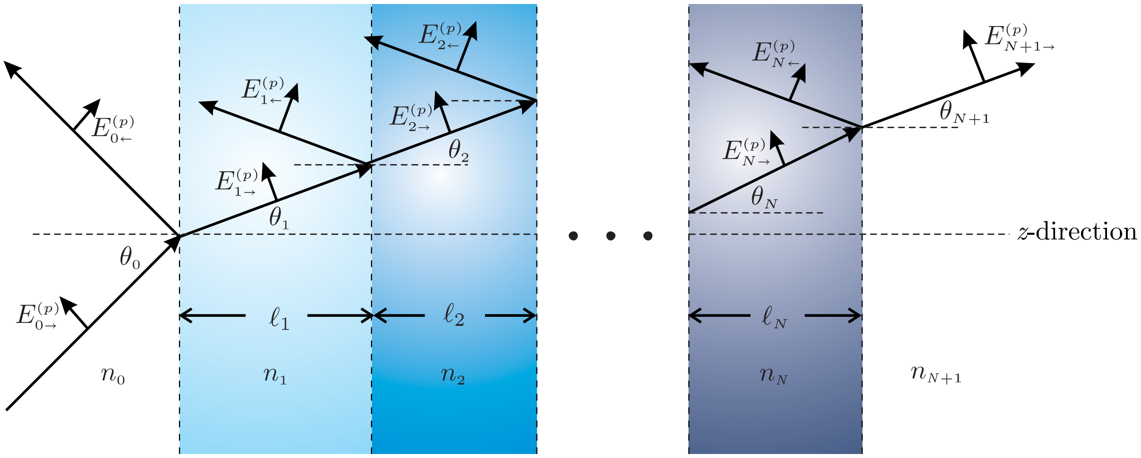


Figure 6.14 Light propagation through multiple layers.

In each layer, only two plane waves exist, each of which is composed of light arising from the many possible bounces from various layer interfaces. The subscript i indicates plane wave fields in individual layers that travel roughly in the incident direction, and the subscript r indicates plane wave fields that travel roughly in the reflected direction. In the final region, there is only one plane wave traveling with a forward or transmitted direction. We will re-label it as $E^{(p)}_{\mathrm{t}_{N+1}} \equiv E^{(p)}_{\mathrm{i}_{N+1}}$ since it is the overall transmitted field.

As we have studied in chapter 3 (see (3.9) and (3.13)), the boundary conditions for the parallel components of the **E** field and for the parallel components of the **B** field lead respectively to

$$\cos\theta_0 \left(E^{(p)}_{0\rightarrow} + E^{(p)}_{0\leftarrow}\right) = \cos\theta_1 \left(E^{(p)}_{1\rightarrow} + E^{(p)}_{1\leftarrow}\right) \tag{6.54}$$

and

$$n_0 \left(E^{(p)}_{0\rightarrow} - E^{(p)}_{0\leftarrow}\right) = n_1 \left(E^{(p)}_{1\rightarrow} - E^{(p)}_{1\leftarrow}\right) \tag{6.55}$$

These equations are applicable only for p-polarized light. Similar equations give the field connection for s-polarized light (see (3.8) and (3.14)).

We have applied these boundary conditions at the first interface only. Of course there are many more interfaces in the multilayer. For the connection between the j^{th} layer and the next, we may similarly write

$$\cos\theta_j \left(E^{(p)}_{j\rightarrow} e^{ik_j\ell_j\cos\theta_j} + E^{(p)}_{j\leftarrow} e^{-ik_j\ell_j\cos\theta_j}\right) = \cos\theta_{j+1} \left(E^{(p)}_{j+1\rightarrow} + E^{(p)}_{j+1\leftarrow}\right) \tag{6.56}$$

and

$$n_j \left(E^{(p)}_{j\rightarrow} e^{ik_j\ell_j\cos\theta_j} - E^{(p)}_{j\leftarrow} e^{-ik_j\ell_j\cos\theta_j}\right) = n_{j+1} \left(E^{(p)}_{j+1\rightarrow} - E^{(p)}_{j+1\leftarrow}\right) \tag{6.57}$$

Here we have set the origin within each layer at the left surface. Then when making the connection with the subsequent layer at the right surface, we must specifically take into account the phase $\mathbf{k}_j \cdot (\ell_j \hat{\mathbf{z}}) = k_j \ell_j \cos\theta_j$. This corresponds to the phase acquired by the plane wave field in traversing the layer with thickness ℓ_j. The right-hand sides of (6.56) and (6.57) need no phase adjustment since the $(j+1)^{\mathrm{th}}$ field is evaluated on the left side of its layer.

At the final interface, the boundary conditions reduce to

$$\cos\theta_N\left(E^{(p)}_{N\rightarrow}e^{ik_N\ell_N\cos\theta_N}+E^{(p)}_{N\leftarrow}e^{-ik_N\ell_N\cos\theta_N}\right)=\cos\theta_{N+1}E^{(p)}_{N+1\rightarrow} \tag{6.58}$$

and

$$n_N\left(E^{(p)}_{N\rightarrow}e^{ik_N\ell_N\cos\theta_N}-E^{(p)}_{N\leftarrow}e^{-ik_N\ell_N\cos\theta_N}\right)=n_{N+1}E^{(p)}_{N+1\rightarrow} \tag{6.59}$$

These equations are the same as (6.56) and (6.57) when $j = N$. However, we have written them here explicitly since they are unique in that $E^{(p)}_{N+1\leftarrow}\equiv 0$.

At this point we are ready to solve (6.54)–(6.59). We would like to eliminate all fields besides $E^{(p)}_{0\rightarrow}$, $E^{(p)}_{0\leftarrow}$, and $E^{(p)}_{N+1\rightarrow}$. Then we will be able to find the overall reflectance and transmittance of the multilayer coating. In solving (6.54)–(6.59), we must proceed with care, or the algebra can quickly get out of hand. Fortunately, most students have had training in linear algebra, and this is a case where that training pays off.

We first write a general matrix equation that summarizes the mathematics in (6.54)–(6.59), as follows:

$$\begin{bmatrix}\cos\theta_j e^{i\beta_j} & \cos\theta_j e^{-i\beta_j}\\ n_j e^{i\beta_j} & -n_j e^{-i\beta_j}\end{bmatrix}\begin{bmatrix}E^{(p)}_{j\rightarrow}\\ E^{(p)}_{j\leftarrow}\end{bmatrix}=\begin{bmatrix}\cos\theta_{j+1} & \cos\theta_{j+1}\\ n_{j+1} & -n_{j+1}\end{bmatrix}\begin{bmatrix}E^{(p)}_{j+1\rightarrow}\\ E^{(p)}_{j+1\leftarrow}\end{bmatrix} \tag{6.60}$$

where

$$\beta_j\equiv\begin{cases}0 & j=0\\ k_j\ell_j\cos\theta_j & 1\le j\le N\end{cases} \tag{6.61}$$

and

$$\begin{aligned}E^{(p)}_{N+1\rightarrow}&\equiv E^{(p)}_{N+1\rightarrow}\\ E^{(p)}_{N+1\leftarrow}&\equiv 0\end{aligned} \tag{6.62}$$

Then we solve (6.60) for the incident fields as follows:

$$\begin{bmatrix}E^{(p)}_{j\rightarrow}\\ E^{(p)}_{j\leftarrow}\end{bmatrix}=\begin{bmatrix}\cos\theta_j e^{i\beta_j} & \cos\theta_j e^{-i\beta_j}\\ n_j e^{i\beta_j} & -n_j e^{-i\beta_j}\end{bmatrix}^{-1}\begin{bmatrix}\cos\theta_{j+1} & \cos\theta_{j+1}\\ n_{j+1} & -n_{j+1}\end{bmatrix}\begin{bmatrix}E^{(p)}_{j+1\rightarrow}\\ E^{(p)}_{j+1\leftarrow}\end{bmatrix} \tag{6.63}$$

We can use (6.63) to connect the fields in the initial and final layers. If we write (6.63) for the $j = 0$ case, and then substitute using (6.63) again with $j = 1$ we find

$$\begin{aligned}\begin{bmatrix}E^{(p)}_{0\rightarrow}\\ E^{(p)}_{0\leftarrow}\end{bmatrix}&=\begin{bmatrix}\cos\theta_0 & \cos\theta_0\\ n_0 & -n_0\end{bmatrix}^{-1}\begin{bmatrix}\cos\theta_1 & \cos\theta_1\\ n_1 & -n_1\end{bmatrix}\begin{bmatrix}E^{(p)}_{1\rightarrow}\\ E^{(p)}_{1\leftarrow}\end{bmatrix}\\ &=\begin{bmatrix}\cos\theta_0 & \cos\theta_0\\ n_0 & -n_0\end{bmatrix}^{-1}M^{(p)}_1\begin{bmatrix}\cos\theta_2 & \cos\theta_2\\ n_2 & -n_2\end{bmatrix}\begin{bmatrix}E^{(p)}_{2\rightarrow}\\ E^{(p)}_{2\leftarrow}\end{bmatrix}\end{aligned} \tag{6.64}$$

where we have grouped the matrices related to the $j = 1$ layer together via

$$M^{(p)}_1\equiv\begin{bmatrix}\cos\theta_1 & \cos\theta_1\\ n_1 & -n_1\end{bmatrix}\begin{bmatrix}\cos\theta_1 e^{i\beta_1} & \cos\theta_1 e^{-i\beta_1}\\ n_1 e^{i\beta_1} & -n_1 e^{-i\beta_1}\end{bmatrix}^{-1} \tag{6.65}$$

By repeating this procedure for all N layers, we connect the fields in the initial medium with the final medium as follows:

$$\begin{bmatrix}E^{(p)}_{0\rightarrow}\\ E^{(p)}_{0\leftarrow}\end{bmatrix}=\begin{bmatrix}\cos\theta_0 & \cos\theta_0\\ n_0 & -n_0\end{bmatrix}^{-1}\left(\prod_{j=1}^{N}M^{(p)}_j\right)\begin{bmatrix}\cos\theta_{N+1} & \cos\theta_{N+1}\\ n_{N+1} & -n_{N+1}\end{bmatrix}\begin{bmatrix}E^{(p)}_{N+1\rightarrow}\\ 0\end{bmatrix} \tag{6.66}$$

where the matrices related to the j^{th} layer are grouped together according to

$$\begin{aligned} M_j^{(p)} &\equiv \begin{bmatrix} \cos\theta_j & \cos\theta_j \\ n_j & -n_j \end{bmatrix} \begin{bmatrix} \cos\theta_j e^{i\beta_j} & \cos\theta_j e^{-i\beta_j} \\ n_j e^{i\beta_j} & -n_j e^{-i\beta_j} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \cos\beta_j & -i\sin\beta_j\cos\theta_j/n_j \\ -in_j\sin\beta_j/\cos\theta_j & \cos\beta_j \end{bmatrix} \end{aligned} \tag{6.67}$$

The matrix inversion in the first line was performed using (0.44). The symbol Π signifies the product of the matrices with the lowest subscripts on the left:

$$\prod_{j=1}^{N} M_j^{(p)} \equiv M_1^{(p)} M_2^{(p)} \cdots M_N^{(p)} \tag{6.68}$$

As a finishing touch, we divide (6.64) by the incident field $E_{0\rightarrow}^{(p)}$ and perform the matrix inversion using (0.44) to obtain

$$\begin{bmatrix} 1 \\ E_{0\leftarrow}^{(p)}/E_{0\rightarrow}^{(p)} \end{bmatrix} = A^{(p)} \begin{bmatrix} E_{N+1\rightarrow}^{(p)}/E_{0\rightarrow}^{(p)} \\ 0 \end{bmatrix} \tag{6.69}$$

where

$$A^{(p)} \equiv \begin{bmatrix} a_{11}^{(p)} & a_{12}^{(p)} \\ a_{21}^{(p)} & a_{22}^{(p)} \end{bmatrix} = \frac{1}{2n_0\cos\theta_0} \begin{bmatrix} n_0 & \cos\theta_0 \\ n_0 & -\cos\theta_0 \end{bmatrix} \left(\prod_{j=1}^{N} M_j^{(p)} \right) \begin{bmatrix} \cos\theta_{N+1} & 0 \\ n_{N+1} & 0 \end{bmatrix} \tag{6.70}$$

In the final matrix after the product in (6.70) we have replaced the entries in the right column with zeros. This is permissable since the column vector that $A^{(p)}$ operates on in (6.69) has a zero in the bottom component. (Having zeros in the matrix can save computation time when calculating with large N.)

Equation (6.69) represents two equations, which must be solved simultaneously to find the ratios $E_{0\leftarrow}^{(p)}/E_{0\rightarrow}^{(p)}$ and $E_{N+1\rightarrow}^{(p)}/E_{0\rightarrow}^{(p)}$. Once the matrix $A^{(p)}$ is computed, this is a relatively simple task:

$$t_p \equiv \frac{E_{N+1\rightarrow}^{(p)}}{E_{0\rightarrow}^{(p)}} = \frac{1}{a_{11}^{(p)}} \qquad \text{(Multilayer)} \tag{6.71}$$

$$r_p \equiv \frac{E_{0\leftarrow}^{(p)}}{E_{0\rightarrow}^{(p)}} = \frac{a_{21}^{(p)}}{a_{11}^{(p)}} \qquad \text{(Multilayer)} \tag{6.72}$$

The convenience of this notation lies in the fact that we can deal with an arbitrary number of layers N with varying thickness and index. The essential information for each layer is contained succinctly in its respective 2×2 matrix. To find the overall effect of the many layers, we need only multiply the matrices for each layer together to find A, and then we can use (6.71) and (6.72) to compute the reflection and transmission coefficients for the whole system.

The derivation for s-polarized light is similar to the derivation for p-polarized light. The equation corresponding to (6.69) for s-polarized light turns out to be

$$\begin{bmatrix} 1 \\ E_{0\leftarrow}^{(s)}/E_{0\rightarrow}^{(s)} \end{bmatrix} = A^{(s)} \begin{bmatrix} E_{N+1\rightarrow}^{(s)}/E_{0\rightarrow}^{(s)} \\ 0 \end{bmatrix} \tag{6.73}$$

where

$$A^{(s)} \equiv \begin{bmatrix} a_{11}^{(s)} & a_{12}^{(s)} \\ a_{21}^{(s)} & a_{22}^{(s)} \end{bmatrix} = \frac{1}{2n_0 \cos\theta_0} \begin{bmatrix} n_0 \cos\theta_0 & 1 \\ n_0 \cos\theta_0 & -1 \end{bmatrix} \left(\prod_{j=1}^{N} M_j^{(s)} \right) \begin{bmatrix} 1 & 0 \\ n_{N+1} \cos\theta_{N+1} & 0 \end{bmatrix} \tag{6.74}$$

and

$$M_j^{(s)} = \begin{bmatrix} \cos\beta_j & -i \sin\beta_j / (n_j \cos\theta_j) \\ -i n_j \cos\theta_j \sin\beta_j & \cos\beta_j \end{bmatrix} \tag{6.75}$$

We can then compute the transmission and reflection coefficients in the same manner that we found the p-components

$$t_s \equiv \frac{E_{N+1\rightarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = \frac{1}{a_{11}^{(s)}} \qquad \text{(Multilayer)} \tag{6.76}$$

$$r_s \equiv \frac{E_{0\leftarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = \frac{a_{21}^{(s)}}{a_{11}^{(s)}} \qquad \text{(Multilayer)} \tag{6.77}$$

6.9 Repeated Multilayer Stacks

In general high-reflection coatings are designed with alternating high and low refractive indices. For high reflectivity, each layer should have a quarter-wave thickness. That is, we need

$$\beta_j = \frac{\pi}{2} \qquad \text{(high reflector)} \tag{6.78}$$

This amounts to the condition on the thickness of

$$\ell_j = \frac{\lambda_{\text{vac}}}{4 n_j \cos\theta_j} \qquad \text{(high reflector)} \tag{6.79}$$

Since the layers alternate high and low indices, at every other boundary there is a phase shift of π upon reflection from the interface. Hence, the quarter wavelength spacing gives maximum reflectivity since the reflected wave in each layer meets the wave in the previous layer in phase. In this situation, the matrix for each layer becomes

$$M_j^{(p)} = \begin{bmatrix} 0 & -i \cos\theta_j / n_j \\ -i n_j / \cos\theta_j & 0 \end{bmatrix} \qquad \text{(high reflector, } p\text{-polarized)} \tag{6.80}$$

The matrices for a high and a low refractive index layer are multiplied together in the usual manner. Each layer pair takes the form

$$\begin{bmatrix} 0 & -\frac{i\cos\theta_H}{n_H} \\ -\frac{i n_H}{\cos\theta_H} & 0 \end{bmatrix} \begin{bmatrix} 0 & -\frac{i\cos\theta_L}{n_L} \\ -\frac{i n_L}{\cos\theta_L} & 0 \end{bmatrix} = \begin{bmatrix} -\frac{n_L \cos\theta_H}{n_H \cos\theta_L} & 0 \\ 0 & -\frac{n_H \cos\theta_L}{n_L \cos\theta_H} \end{bmatrix} \tag{6.81}$$

To extend to $q = N/2$ layer pairs, we have

$$\begin{aligned} \prod_{j=1}^{N} M_j^{(p)} &= \begin{bmatrix} -\frac{n_L \cos\theta_H}{n_H \cos\theta_L} & 0 \\ 0 & -\frac{n_H \cos\theta_L}{n_L \cos\theta_H} \end{bmatrix}^q \\ &= \begin{bmatrix} \left(-\frac{n_L \cos\theta_H}{n_H \cos\theta_L}\right)^q & 0 \\ 0 & \left(-\frac{n_H \cos\theta_L}{n_L \cos\theta_H}\right)^q \end{bmatrix} \end{aligned} \tag{6.82}$$

and using (6.82) we can compute $A^{(p)}$

$$A^{(p)} = \frac{1}{2}\left[\begin{array}{cc} \left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} + \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0} & 0 \\ \left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} - \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0} & 0 \end{array}\right] \tag{6.83}$$

This stack of q periods can achieve extraordinarily high reflectivity. In the limit of $q \to \infty$, we have $t_p \to 0$ and $r_p \to -1$ from (6.71) and (6.72), giving 100% reflection.

Sometimes multilayer coatings are made with repeated stacks of layers. In general, if the same series of layers in (6.82) is repeated many times, say q times, the following formula known as Sylvester's theorem (see appendix 0.4) comes in handy:

$$\left[\begin{array}{cc} A & B \\ C & D \end{array}\right]^q = \frac{1}{\sin\theta}\left[\begin{array}{cc} A\sin q\theta - \sin(q-1)\theta & B\sin q\theta \\ C\sin q\theta & D\sin q\theta - \sin(q-1)\theta \end{array}\right] \tag{6.84}$$

where

$$\cos\theta \equiv \frac{1}{2}(A+D). \tag{6.85}$$

This formula relies on the condition $AD - BC = 1$, which is true for matrices of the form (6.67) and (6.75) or any product of them. Here, A, B, C, and D represent the elements of a matrix composed of a block of matrices corresponding to a repeated pattern within the stack.

Many different types of multilayer coatings are possible. For example, a Brewster's-angle polarizer has a coating designed to transmit with high efficiency p-polarized light while simultaneously reflecting s-polarized light with high efficiency. The backside of the substrate is left uncoated where p-polarized light passes with 100% efficiency at Brewster's angle.

Exercises

6.2 Double Boundary Problem Solved Using Fresnel Coefficients

P6.1 You have a 1 micron thick coating of dielectric material ($n = 2$) on a piece of glass ($n = 1.5$). Use a computer to plot the magnitude of the Fresnel coefficient (6.10) from air into the glass at normal incidence. Plot as a function of wavelength for wavelengths between 200 nm and 800 nm (assume the index remains constant over this range).

6.3 Double Boundary Problem at Sub Critical Angles

P6.2 A light wave impinges at normal incidence on a thin glass plate with index n and thickness d.

(a) Show that the transmittance through the plate as a function of wavelength is

$$T^{\text{tot}} = \frac{1}{1 + \frac{\left(n^2-1\right)^2}{4n^2}\sin^2\left(\frac{2\pi nd}{\lambda_{\text{vac}}}\right)}$$

HINT: Find

$$r^{1\to2} = r^{1\to0} = -r^{0\to1} = \frac{n-1}{n+1}$$

and then use

$$T^{\text{i}\to\text{m}} = 1 - R^{0\to1}$$

$$T^{1\to2} = 1 - R^{1\to2}$$

(b) If $n = 1.5$, what is the maximum and minimum transmittance through the plate?

(c) If the plate thickness is $d = 150\ \mu$m, what wavelengths transmit with maximum efficiency?

HINT: Give a formula involving an integer N.

P6.3 Consider the "beam splitter" introduced in Example 6.2. Show that the maximum reflectance possible from the single coating at the first surface is 46%. Find the smallest possible d_1 that accomplishes this for light with wavelength $\lambda_{\text{vac}} =$ 633 nm.

6.4 Beyond Critical Angle: Tunneling of Evanescent Waves

P6.4 Re-compute (6.32) in the case of s-polarized light. Write the result in the same form as the last expression in (6.32). HINT: You need to redo (6.28)–(6.30).

L6.5 Consider s-polarized microwaves ($\lambda_{\text{vac}} = 3$ cm) encountering an air gap separating two paraffin wax prisms ($n = 1.5$). The 45° right-angle prisms are arranged with the geometry shown in Fig. 6.3. The presence of the second prism frustrates the total internal reflection that would have occurred if the first prism were by itself. This occurs because "feedback" from the second surface disrupts the evanescent waves.

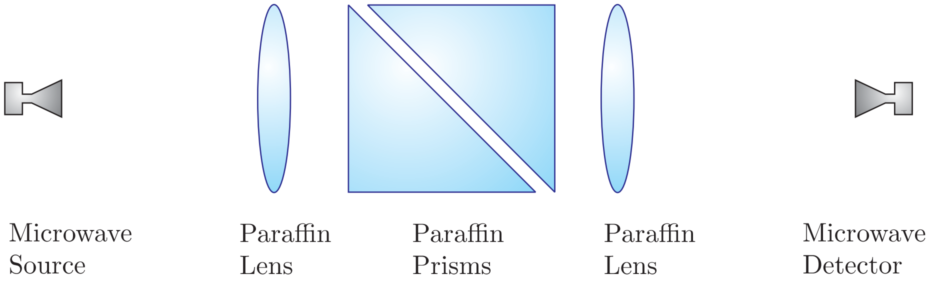


Figure 6.15

(a) Use a computer to plot the transmittance through the gap as a function of separation d (normal to gap surface). Do not consider reflections from other surfaces of the prisms.

HINT: Plot the result of P 6.4.

(b) Measure the transmittance of the microwaves through the prisms as function of spacing d (normal to the surface) and superimpose the results on the graph of part (a).

RESULT: See the graph below. Presumably experimental error causes some discrepancy, but the trend is clear.

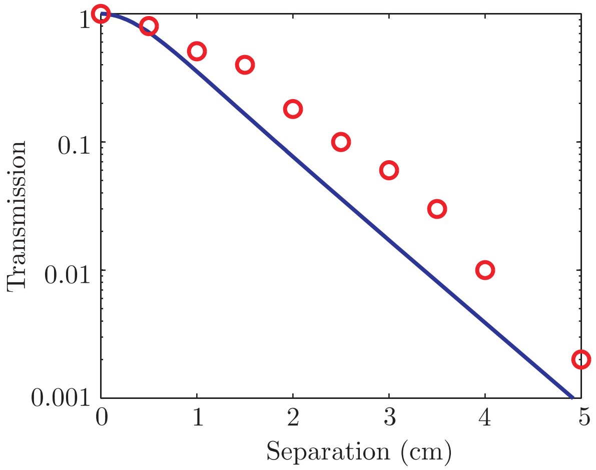


Figure 6.16

6.7 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument

P6.6 A Fabry-Perot interferometer has silver-coated plates each with reflectance $R = 0.9$, transmittance $T = 0.05$, and absorbance $A = 0.05$. The plate separation is $d = 0.5$ cm with interior index $n_1 = 1$. Suppose that the wavelength being observed near normal incidence is 587 nm.

(a) What is the maximum and minimum transmittance through the interferometer?

(b) What are the free spectral range $\Delta\lambda_{\text{FSR}}$ and the fringe width $\Delta\lambda_{\text{FWHM}}$?

(c) What is the resolving power?

P6.7 Generate a plot like Fig. 6.10(a), showing the fringes you get in a Fabry-Perot etalon when θ_1 is varied. Let $T_{\max} = 1$, $F = 10$, $\lambda = 500$ nm, $d = 1$ cm, and $n_1 = 1$.

(a) Plot T vs. θ_1 over the angular range used in Fig. 6.10(a).

(c) Suppose d was slightly different, say 1.00002 cm. Make a plot of T vs θ_1 for this situation.

P6.8 Consider the configuration depicted in Fig. 6.9, where the center of the diverging light beam $\lambda_{\text{vac}} = 633$ nm approaches the plates at normal incidence. Suppose that the spacing of the plates (near $d = 0.5$ cm) is just right to cause a bright fringe to occur at the center. Let $n_1 = 1$. Find the angle for the m^{th} circular bright fringe surrounding the central spot (the 0^{th} fringe corresponding to the center). HINT: $\cos\theta \cong 1 - \theta^2/2$. The answer has the form $a\sqrt{m}$; find the value of a.

L6.9 Characterize a Fabry-Perot etalon in the laboratory using a HeNe laser ($\lambda_{\text{vac}} = 633$ nm). Assume that the bandwidth $\Delta\lambda_{\text{HeNe}}$ of the HeNe laser is very narrow compared to the fringe width of the etalon $\Delta\lambda_{\text{FWHM}}$. Assume two identical reflective surfaces separated by 5.00 mm. Deduce the free spectral range $\Delta\lambda_{\text{FSR}}$, the fringe width $\Delta\lambda_{\text{FWHM}}$, the resolving power, and the reflecting finesse (small f).

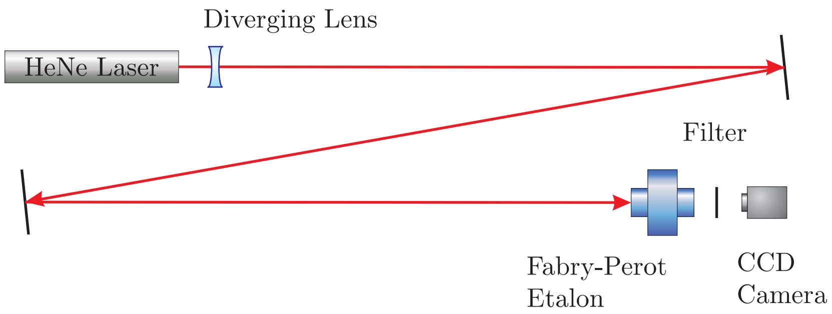


Figure 6.17

L6.10 Use the same Fabry-Perot etalon to observe the Zeeman splitting of the yellow line $\lambda = 587.4$ nm emitted by a krypton lamp when a magnetic field is applied. As the line splits and moves through half of the free spectral range, the peak of the decreasing wavelength and the peak of the increasing wavelength meet on the screen. When this happens, by how much has each wavelength shifted?

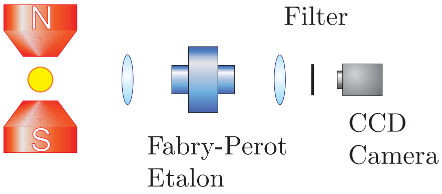


Figure 6.18

6.8 Multilayer Coatings

P6.11 (a) Write (6.54) through (6.59) for s-polarized light.

(b) From these equations, derive (6.73)–(6.75).

P6.12 Beginning with (6.76) for a single layer between two materials (i.e. two interfaces), derive (6.21). WARNING: This is more work than it may appear at first.

6.9 Repeated Multilayer Stacks

P6.13 (a) What should be the thickness of the high and the low index layers in a periodic high-reflector mirror? Let the light be p-polarized and strike the mirror surface at 45°. Take the indices of the layers be $n_{\mathrm{H}} = 2.32$ and $n_{\mathrm{L}} = 1.38$, deposited on a glass substrate with index $n = 1.5$. Let the wavelength be $\lambda_{\mathrm{vac}} = 633$ nm.

(b) Find the reflectance R with 1, 2, 4, and 8 periods in the high-low stack.

P6.14 Find the high-reflector matrix for s-polarized light that corresponds to (6.82).

P6.15 Design an anti-reflection coating for use in air (assume the index of air is 1):

(a) Show that for normal incidence and $\lambda/4$ films (thickness= $\frac{1}{4}$ the wavelength of light inside the material), the reflectance of a single layer (n_1) coating on a glass is

$$R = \left(\frac{n_g - n_1^2}{n_g + n_1^2}\right)^2$$

(b) Show that for a two coating setup (air-n_1-n_2-glass; n_1 and n_2 are each a $\lambda/4$ film), that

$$R = \left(\frac{n_2^2 - n_g n_1^2}{n_2^2 + n_g n_1^2}\right)^2$$

(c) If $n_g = 1.5$, and you have a choice of these common coating materials: ZnS ($n = 2.32$), CeF ($n = 1.63$) and MgF ($n = 1.38$), find the combination that gives you the lowest R for part (b). (Be sure to specify which material is n_1 and which is n_2.) What R does this combination give?

P6.16 Suppose you design a two-coating "anti-reflection optic" (each coating set for $\lambda/4$, as in the last problem) using $n_1 = 1.6$ and $n_2 = 2.1$. Assume you've got $n_g = 1.5$ and normal incidence. If you design your coatings to be quarter-wave for $\lambda = 550$ nm (in the middle of the visible range) the R that you found in P 6.15(b) will be true only for that specific wavelength for two reasons: the index changes with λ, but more importantly, the thicknesses used in the coatings will not be $\lambda/4$ for other wavelengths. Let's ignore the index change with λ and focus on the wavelength dependence. Use the matrix techniques and a computer to plot $R(\lambda_{\text{air}}$ for 400 to 700 nm (visible range). Do this for a single bilayer (one layer of each coating, two bilayers, four bilayers, and 25 bilayers.

Chapter 7

Superposition of Quasi-Parallel Plane Waves

7.1 Introduction

Through the remainder of our study of optics, we will be interested in the superposition of many plane waves, which interfere to make an overall waveform. Such a waveform can be represented as follows:

$$\mathbf{E}(\mathbf{r}, t) = \sum_j \mathbf{E}_j e^{i(\mathbf{k}_j \cdot \mathbf{r} - \omega_j t)} \tag{7.1}$$

The corresponding magnetic field (see (2.55)) is

$$\mathbf{B}(\mathbf{r}, t) = \sum_j \mathbf{B}_j e^{i(\mathbf{k}_j \cdot \mathbf{r} - \omega_j t)} = \sum_j \frac{\mathbf{k}_j \times \mathbf{E}_j}{\omega_j} e^{i(\mathbf{k}_j \cdot \mathbf{r} - \omega_j t)} \tag{7.2}$$

In section 7.2, we show that the intensity of this overall field under certain assumptions can be expressed as

$$I(\mathbf{r}, t) = \frac{n\epsilon_0 c}{2} \mathbf{E}(\mathbf{r}, t) \cdot \mathbf{E}^* (\mathbf{r}, t) \tag{7.3}$$

where $\mathbf{E}(\mathbf{r}, t)$ represents the entire complex expression for the electric field rather than just the real part. Although this expression is reminiscent of (2.61), it should be kept in mind that we previously considered only a single plane wave (perhaps with two distinct polarization components). It may not be immediately obvious, but (7.3) automatically time-averages over rapid oscillations so that I retains only a *slowly varying* time dependence.

Equation (7.3) is exact only if the vectors $\mathbf{k}_j$ are all parallel. This is not as serious a restriction as might seem at first. For example, the output of a Michelson interferometer (studied in the next chapter) is the superposition of two fields, each composed of a range of frequencies with parallel $\mathbf{k}_j$'s. We can relax the restriction of parallel $\mathbf{k}_j$'s slightly and apply (7.3) also to plane waves with *nearly* parallel $\mathbf{k}_j$'s such as occurs in a Young's two-slit diffraction experiment (studied in the next chapter). In such diffraction problems, (7.3) is viewed as an approximation valid to the extent that the vectors $\mathbf{k}_j$ are close to parallel.

In section 7.3 we introduce the concept of *group velocity*, which is distinct from *phase velocity* that we encountered previously. As we saw in chapter 2, the real part of refractive index in certain situations can be less than one, indicating *superluminal* wave crest propagation (i.e. greater than c)! In this case, the group velocity is usually less than c. Group velocity tracks the speed of the interference or "ripples" resulting from the superposition of multiple waves. Thus, the intensity of a waveform is more connected with the group velocity, rather than the phase velocity.

Nevertheless, it is possible for the group velocity also to become superluminal when absorption or amplification is involved. Group velocity tracks the *presence* or *locus* of field energy, which is indirectly influenced by an exchange of energy with the medium. For a complete picture, one must consider the energy stored in both the field and the medium. So-called superluminal pulse propagation occurs when a magician invites the audience to look only at the field energy while energy transfers into and out of the "unwatched" domain of the medium. Extra field energy can seemingly appear prematurely downstream, but only if there is already non-zero field energy downstream to stimulate a transfer of energy between the field and the medium. As is explained in Appendix 7.A, the actual transport of energy is strictly bounded by c; superluminal *signal* propagation is impossible.

In section 7.4, we reconsider waveforms composed of a continuum of plane waves, each with a distinct frequency ω. We discuss superpositions of plane waves in terms of Fourier theory. (For an introductory overview of Fourier transforms, see section 0.3.) Essentially, a Fourier transform enables us to determine which plane waves are necessary to construct a given wave from $\mathbf{E}(\mathbf{r}_1, t)$. This is important if we want to know what happens to a waveform as it traverses from point $\mathbf{r}_1$ to $\mathbf{r}_2$ in a material with a frequency-dependent index. Different frequency components of the waveform experience different phase velocities, causing the waveform to undergo distortion as it propagates, a phenomenon called *dispersion.* Since we already know how individual plane waves propagate in a material, we can reassemble them at the end of propagation to obtain the new overall pulse $\mathbf{E}(\mathbf{r}_2, t)$ (i.e. by performing an inverse Fourier transform). This procedure is examined in section 7.6 specifically for a light pulse with a Gaussian temporal profile. We shall see that the group velocity tracks the movement of the center of the wave packet. The arguments are presented in a *narrowband* context where the pulse maintains its characteristic shape while spreading. In section 7.7, we examine group velocity in a generalized *broadband* context where the wave packet can become severely distorted during propagation.

7.2 Intensity

In this section we justify the expression for intensity given in (7.3). The Poynting vector (2.51) is

$$\mathbf{S}(\mathbf{r}, t) = \frac{\mathrm{Re}\{\mathbf{E}(\mathbf{r}, t)\} \times \mathrm{Re}\{\mathbf{B}(\mathbf{r}, t)\}}{\mu_0} \tag{7.4}$$

Upon substitution of (7.1) and (7.2) into the above expression, we obtain

$$\mathbf{S}(\mathbf{r}, t) = \sum_{j,m} \frac{1}{\omega_m \mu_0} \mathrm{Re}\left\{\mathbf{E}_j e^{i(\mathbf{k}_j \cdot \mathbf{r} - \omega_j t)}\right\} \times \left[\mathbf{k}_m \times \mathrm{Re}\left\{\mathbf{E}_m e^{i(\mathbf{k}_m \cdot \mathbf{r} - \omega_m t)}\right\}\right] \tag{7.5}$$

For simplicity, we assume that all vectors $\mathbf{k}_j$ are real. If the wave vectors are complex, the same upcoming result can be obtained. In that case, as in (2.61), the field amplitudes $\mathbf{E}_j$ would correspond to local amplitudes (as energy is absorbed or amplified during propagation).

Next we apply the BAC-CAB rule (P 0.12) to (7.5) and obtain

$$\begin{aligned}\mathbf{S}(\mathbf{r},t) = \sum_{j,m} \frac{1}{\omega_m \mu_0} \Big[\mathbf{k}_m \Big(\mathrm{Re}\Big\{\mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}\Big\} \cdot \mathrm{Re}\Big\{\mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}\Big\}\Big) \\ - \mathrm{Re}\Big\{\mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}\Big\} \Big(\mathrm{Re}\Big\{\mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}\Big\} \cdot \mathbf{k}_m\Big)\Big]\end{aligned} \tag{7.6}$$

The last term in (7.6) can be dismissed if all of the $\mathbf{k}_m$ are perpendicular to each of the $\mathbf{E}_j$. This can only be ensured if all $\mathbf{k}$-vectors are parallel to each other. Let us make this rather stringent assumption and drop the last term in (7.6). The magnitude of the Poynting vector then becomes

$$S(\mathbf{r},t) = \epsilon_0 c \sum_{j,m} n_m \mathrm{Re}\Big\{\mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}\Big\} \cdot \mathrm{Re}\Big\{\mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}\Big\} \quad \text{(parallel } \mathbf{k}\text{-vectors)} \tag{7.7}$$

where in accordance with (1.44) and (2.21) we have introduced

$$\frac{k_m}{\omega_m \mu_0} = n_m \epsilon_0 c. \tag{7.8}$$

Here n_m refers to the refractive index associated with the frequency ω_m. If we assume that the index does not vary dramatically with frequency, we may approximate it as a constant. We usually measure intensity outside of materials (in air or in vacuum), so this approximation is often quite fine. With these approximations the magnitude of the Poynting vector becomes (with the help of (0.17))

$$\begin{aligned}S(\mathbf{r},t) &= n\epsilon_0 c \sum_{j,m} \left[\frac{\mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} + \mathbf{E}_j^* e^{-i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}}{2}\right] \cdot \left[\frac{\mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)} + \mathbf{E}_m^* e^{-i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}}{2}\right] \\ &= \frac{nc\epsilon_0}{4} \sum_{j,m} \Big[\mathbf{E}_j \cdot \mathbf{E}_m e^{i[(\mathbf{k}_j+\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j+\omega_m)t]} + \mathbf{E}_j^* \cdot \mathbf{E}_m^* e^{-i[(\mathbf{k}_j+\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j+\omega_m)t]} \\ &\qquad + \mathbf{E}_j \cdot \mathbf{E}_m^* e^{i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j-\omega_m)t]} + \mathbf{E}_j^* \cdot \mathbf{E}_m e^{-i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j-\omega_m)t]}\Big] \\ &\qquad \text{(parallel } \mathbf{k}\text{-vectors, constant } n\text{)}\end{aligned} \tag{7.9}$$

Notice that each of the first two terms in (7.9) oscillates very rapidly (at frequency $\omega_j + \omega_m$). The time average of these terms goes to zero. The second two terms oscillate slowly or not at all if $j = m$. Taking the time average over the rapid oscillation in (7.9), we then get

$$\begin{aligned}
\langle S(\mathbf{r},t)\rangle_{\text{osc.}} &= \frac{n\epsilon_0 c}{2}\sum_{j,m}\frac{\mathbf{E}_j\cdot\mathbf{E}_m^* e^{i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j-\omega_m)t]}+\mathbf{E}_j^*\cdot\mathbf{E}_m e^{-i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_n-\omega_m)t]}}{2}\\
&= \frac{n\epsilon_0 c}{2}\text{Re}\left\{\sum_j \mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}\cdot\sum_m \mathbf{E}_m^* e^{-i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}\right\}\\
&= \frac{n\epsilon_0 c}{2}\text{Re}\left[\mathbf{E}\left(\mathbf{r},t\right)\cdot\mathbf{E}^*(\mathbf{r},t)\right].
\end{aligned}$$

$$\text{(parallel } \mathbf{k}\text{-vectors, constant } n\text{, time-averaged over rapid oscillations)} \tag{7.10}$$

In writing the final line we have again invoked (7.1). Notice that the expression $\mathbf{E}(\mathbf{r},t)\cdot\mathbf{E}^*(\mathbf{r},t)$ is already real. Therefore, we may drop the function Re [], and (7.3) is verified. The assumptions behind (7.3) are now clear.

In dropping the vector symbol from all $\mathbf{k}_m$ to get (7.7) we assumed that all $\mathbf{k}_m$ are nearly parallel to each other. If some of the $\mathbf{k}_m$ point in an anti-parallel direction, we can still proceed with the above approximations but with negative signs entered explicitly into (7.7) for those components. For example, a standing wave has no net flow of energy and the net Poynting vector is zero. This brings out the distinction between irradiance S and intensity I. Intensity is a measure of what atoms "feel", which is not zero for standing waves. On the other hand, $\langle S\rangle$ is identically zero for standing waves because there is no net flow of energy. Thus, we often apply (7.10) to standing waves (technically incorrect in the above context), but we refer to the result as intensity instead of irradiance or Poynting flux. We do this because for many experiments it is not important whether the field is traveling or standing, but it is only important that atoms locally experience an oscillating electric field. At extreme intensities, however, where the influence of the magnetic field becomes comparable to that of the electric field, the distinction between propagating and standing fields can become important.

In summary, the intensity of the field (time-averaged over rapid oscillations) may be expressed approximately as

$$I(\mathbf{r},t)\cong\frac{n\epsilon_0 c}{2}\mathbf{E}(\mathbf{r},t)\cdot\mathbf{E}^*\left(\mathbf{r},t\right)\qquad\text{(parallel or antiparallel } \mathbf{k}\text{-vectors, constant } n\text{)} \tag{7.11}$$

where $\mathbf{E}(\mathbf{r},t)$ is entered in complex format.

7.3 Group vs. Phase Velocity: Sum of Two Plane Waves

Consider the sum of two plane waves with equal amplitudes:

$$\mathbf{E}(\mathbf{r},t)=\mathbf{E}_0 e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)}+\mathbf{E}_0 e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)} \tag{7.12}$$

As we previously studied (see P 1.10), the velocities of the individual wave crests are

$$\begin{aligned}
v_{p1} &= \omega_1/k_1\\
v_{p2} &= \omega_2/k_2
\end{aligned} \tag{7.13}$$

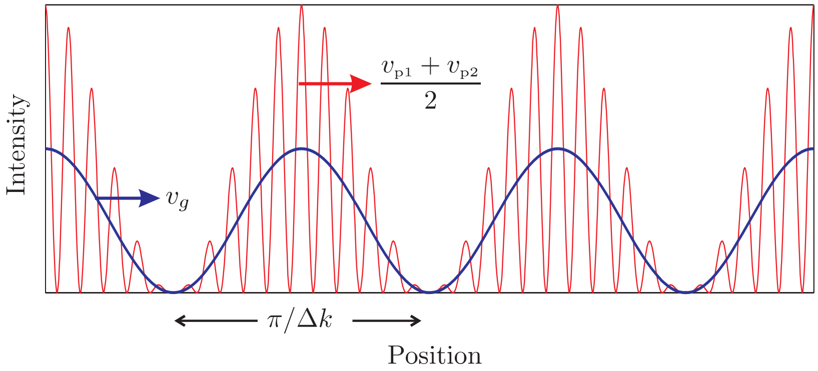


Figure 7.1 Intensity of two interfering plane waves. The solid line shows intensity averaged over rapid oscillations.

These are known as the *phase velocities* of the individual plane waves. As the two plane waves propagate, they interfere, giving regions of higher and lower intensity.

As we now show, the peaks in the intensity distribution (7.11) can move at a velocity quite different from the phase velocities in (7.13). The intensity associated with (7.12) is computed as follows:

$$\begin{aligned} I(\mathbf{r},t) &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0\cdot\mathbf{E}_0^*\left[e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)}+e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)}\right]\left[e^{-i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)}+e^{-i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)}\right]\\ &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0\cdot\mathbf{E}_0^*\left[2+e^{i[(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}-(\omega_2-\omega_1)t]}+e^{-i[(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}-(\omega_2-\omega_1)t]}\right]\\ &= n\epsilon_0 c\mathbf{E}_0\cdot\mathbf{E}_0^*\left[1+\cos\left[(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}-(\omega_2-\omega_1)t\right]\right]\\ &= n\epsilon_0 c\mathbf{E}_0\cdot\mathbf{E}_0^*\left[1+\cos\left(\Delta\mathbf{k}\cdot\mathbf{r}-\Delta\omega t\right)\right] \end{aligned} \tag{7.14}$$

where

$$\begin{aligned}\Delta\mathbf{k} &\equiv \mathbf{k}_2-\mathbf{k}_1\\ \Delta\omega &\equiv \omega_2-\omega_1\end{aligned} \tag{7.15}$$

Keep in mind that this intensity is averaged over rapid oscillations. The solid line in Fig. 7.1 shows this time-averaged version of the intensity given by the above expression. The dashed line shows the intensity with the rapid oscillations retained, according to (7.9). It is left as an exercise (see P 7.3) to show that the rapid-oscillation peaks in Fig. 7.1 (dashed) move at the average of the phase velocities in (7.13).

An examination of (7.14) reveals that the time-averaged curve in Fig. 7.1 (solid) travel with speed

$$v_g \equiv \frac{\Delta\omega}{\Delta k} \tag{7.16}$$

This is known as the *group velocity*. Essentially, v_g may be thought of as the velocity for the envelope that encloses the rapid oscillations.

In general, v_g and v_p are not the same. This means that as the waveform propagates, the rapid oscillations move within the larger modulation pattern, for example, continually disappearing at the front and reappearing at the back of each modulation. The *presence* of field energy (which gives rise to intensity) is clearly tied more to v_g than to v_p. The group velocity is identified with the propagation of overall waveforms.

As an example of the behavior of group velocity, consider the propagation of two plane waves in a plasma (see P 2.8) for which the index is real over a range of frequencies. The index of refraction is given by

$$n_{\text{plasma}}(\omega) = \sqrt{1 - \omega_p^2/\omega^2} < 1 \quad \text{(assuming } \omega > \omega_p\text{)} \tag{7.17}$$

The phase velocity for each frequency is computed by

$$\begin{aligned} v_{p1} &= c/n_{\text{plasma}}(\omega_1) \\ v_{p2} &= c/n_{\text{plasma}}(\omega_2) \end{aligned} \tag{7.18}$$

Since $n_{\text{plasma}} < 1$, both of these velocities exceed c. However, the group velocity is

$$v_g = \frac{\Delta\omega}{\Delta k} \cong \frac{d\omega}{dk} = \left[\frac{dk}{d\omega}\right]^{-1} = \left[\frac{d}{d\omega}\frac{\omega n_{\text{plasma}}(\omega)}{c}\right]^{-1} = n_{\text{plasma}}(\omega)\, c \tag{7.19}$$

which is clearly less than c (deriving the final expression in (7.19) from the previous one is left as an exercise). For convenience, we have taken ω_1 and ω_2 to lie very close to each other.

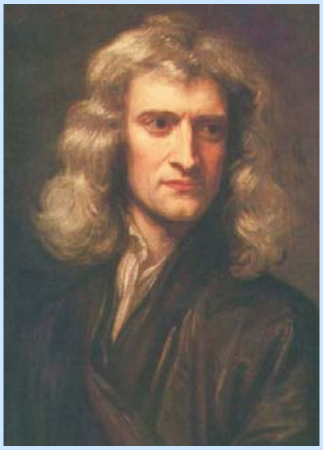
This example shows that in an environment where the index of refraction is real (i.e. no net exchange of energy with the medium), the group velocity does not exceed c, although the phase velocity does. The group velocity tracks the *presence* of field energy, whether that energy propagates or is extracted from a material. The universal speed limit c is always obeyed in energy transportation.

The fact that the phase velocity can exceed c should not disturb students. In the above example, the “fast-moving” phase oscillations result merely from an interplay between the field and the plasma. In a similar sense, the intersection of an ocean wave with the shoreline can also exceed c, if different points on the wave front happen to strike the shore nearly simultaneously. The point of intersection between the wave and the shoreline does not constitute an actual object under motion. Similarly, wave crests of individual plane waves do not necessarily constitute actual objects that are moving; in general, v_p is not the relevant speed at which events up stream influence events down stream. From another perspective, individual plane waves have infinite length and infinite duration. They do not exist in isolation except in our imagination. All real waveforms are comprised of a range of frequency components, and so interference always happens. Energy is associated with regions of constructive interference between those waves.

If there is an exchange of energy between the field and the medium (i.e. if the index of refraction is complex), v_g still describes where field energy may be found, but it does not give the whole story in terms of energy flow (addressed in Appendix 7.A).

7.4 Frequency Spectrum of Light

We continue our study of waveforms. An arbitrary waveform can be constructed from a superposition of plane waves. The discrete summation in (7.1) is of limited use, since a waveform constructed from a discrete sum must eventually repeat over and over. To create a waveform that does not repeat (e.g. a single laser pulse or, technically speaking,



Isaac Newton

(1643–1727, English)

Newton demonstrated that "white" light is composed of many different colors. He realized that the amount of refraction experienced by light depends on its color, so that refracting telescopes would suffer from chromatic abberation. He advanced a "corpuscular" theory of light, although his notion of light particles bears little resemblance to the modern notion of light quanta.

any waveform that exists in the physical world) a continuum of plane waves is necessary. Several examples of waveforms are shown in Fig. 7.2. To construct non-repeating waveforms, the summation in (7.1) must be replaced by an integral, and the waveform at a point $\mathbf{r}$ can be expressed as

$$\mathbf{E}(\mathbf{r},t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}\left(\mathbf{r},\omega\right) e^{-i\omega t} d\omega \tag{7.20}$$

The function $\mathbf{E}\left(\mathbf{r},\omega\right)$ has units of field per frequency. It gives the contribution of each frequency component to the overall waveform and includes all spatial dependence such as the factor $\exp\left\{i\mathbf{k}\left(\omega\right)\cdot\mathbf{r}\right\}$. The function $\mathbf{E}\left(\mathbf{r},\omega\right)$ is distinguished from the function $\mathbf{E}(\mathbf{r},t)$ by its argument (i.e. ω instead of t). The factor $1/\sqrt{2\pi}$ is introduced to match our Fourier transform convention.

Given knowledge of $\mathbf{E}\left(\mathbf{r},\omega\right)$, the waveform $\mathbf{E}(\mathbf{r},t)$ can be constructed. Similarly, if the waveform $\mathbf{E}(\mathbf{r},t)$ is known, the field per frequency can be obtained via

$$\mathbf{E}\left(\mathbf{r},\omega\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}\left(\mathbf{r},t\right) e^{i\omega t} dt \tag{7.21}$$

This operation, which produces $\mathbf{E}\left(\mathbf{r},\omega\right)$ from $\mathbf{E}(\mathbf{r},t)$, is called a *Fourier transform.* The operation (7.20) is called the *inverse Fourier transform.* For a review of Fourier theory, see section 0.3.

Even though $\mathbf{E}(\mathbf{r},t)$ can be written as a real function (since, after all, only the real part is relevant), $\mathbf{E}\left(\mathbf{r},\omega\right)$ is in general complex. The real and imaginary parts of $\mathbf{E}\left(\mathbf{r},\omega\right)$keep track of how much cosine and how much sine, respectively, make up $\mathbf{E}(\mathbf{r},t)$. Keep in mind that both positive and negative frequency components go into the cosine and sine according to (0.6). Therefore, it should not seem strange that we integrate (7.20) over all frequencies, both positive and negative. If $\mathbf{E}\left(\mathbf{r},t\right)$ is taken to be a real function, then we have the symmetry relation

$$\mathbf{E}\left(\mathbf{r},-\omega\right) = \mathbf{E}^{*}\left(\mathbf{r},\omega\right) \quad \text{(if } \mathbf{E}(\mathbf{r},t) \text{ is real)} \tag{7.22}$$

However, often $\mathbf{E}(\mathbf{r}, t)$ is written in complex notation, where taking the real part is implied. For example, the real waveform

$$\mathbf{E}_r(\mathbf{r}, t) = \mathbf{E}_{0r}(\mathbf{r})\, e^{-t^2/2\tau^2} \cos(\omega_0 t - \phi) \tag{7.23}$$

is usually written as

$$\mathbf{E}_c(\mathbf{r}, t) = \mathbf{E}_{0c}(\mathbf{r})\, e^{-t^2/2\tau^2} e^{-i\omega_0 t} \tag{7.24}$$

where $\mathbf{E}_r(\mathbf{r}, t) = \mathrm{Re}\{\mathbf{E}_c(\mathbf{r}, t)\}$. The phase ϕ is hidden within the complex amplitude $\mathbf{E}_{0c}(\mathbf{r})$, where in writing (7.23) we have assumed (for simplicity) that each field vector component contains the same phase. This waveform is shown in Fig. 7.2 for various parameters τ.

Consider the Fourier transforms of the waveform (7.23). Upon applying (7.21) we get (see P 0.27)

$$\mathbf{E}_r(\mathbf{r}, \omega) = \tau \mathbf{E}_{0r}(\mathbf{r}) \frac{e^{-i\phi} e^{-\frac{\tau^2(\omega+\omega_0)^2}{2}} + e^{i\phi} e^{-\frac{\tau^2(\omega-\omega_0)^2}{2}}}{2} \tag{7.25}$$

Similarly, the Fourier transform of (7.24), i.e. the complex version of the same waveform, is

$$\mathbf{E}_c(\mathbf{r}, \omega) = \tau \mathbf{E}_{0c}(\mathbf{r})\, e^{-\frac{\tau^2(\omega-\omega_0)^2}{2}} \tag{7.26}$$

The latter transform is less cumbersome to perform, and for this reason more often used.

Figure 7.3 shows graphs of $|E_r(\mathbf{r}, \omega)|^2$ associated with the waveforms in Fig. 7.2. Figure 7.4 shows graphs of $\mathbf{E}_c(\mathbf{r}, \omega) \cdot \mathbf{E}_c^*(\mathbf{r}, \omega)/2$ obtained from the complex versions of the same waveforms. The graphs show the *power spectra* of the field (aside from some multiplicative constants). A waveform that lasts for a brief interval of time (i.e. small τ) has the widest spectral distribution in the *frequency domain.* In Figs. 7.3a and 7.4a, we have chosen an extremely short waveform (perhaps even physically difficult to create, with $\tau = \pi/(2\omega_0)$, see Fig. 7.2a) to illustrate the distinction between working with the real and the complex representations of the field. Notice that the Fourier transform (7.25) of the real field depicted in Fig. 7.3 obeys the symmetry relation (7.22), whereas the Fourier transform of the complex field (7.26) does not.

Essentially, the power spectrum of the complex representation of the field can be understood to be twice the power spectrum of the real representation, but plotted only for the positive frequencies. This works well as long as the spectrum is well localized so that there is essentially no spectral amplitude near $\omega = 0$ (i.e. no DC component). This is not the case in Figs. 7.3a and 7.4a. Because the waveform is extremely short in time, the extraordinarily wide spectral peaks spread to the origin, and Fig. 7.4a does not accurately depict the positive-frequency side of Fig. 7.3a since the two peaks merge into each other. In practice, we almost never run into this problem in optics (i.e. waveforms are typically much longer in time). For one thing, in the above examples, the waveform or pulse duration τ is so short that there is only about one oscillation within the pulse. Typically, there are several oscillations within a waveform and no DC component. Throughout the remainder of this book, we shall assume that the frequency spread is localized around ω_0, so that we can use the complex representation with impunity.

The intensity defined by (7.3) is also useful for the continuous superposition of plane waves as defined by the inverse Fourier transform (7.20). We can plug in the expression for the field in complex format. The intensity in (7.3) takes care of the time-average over rapid

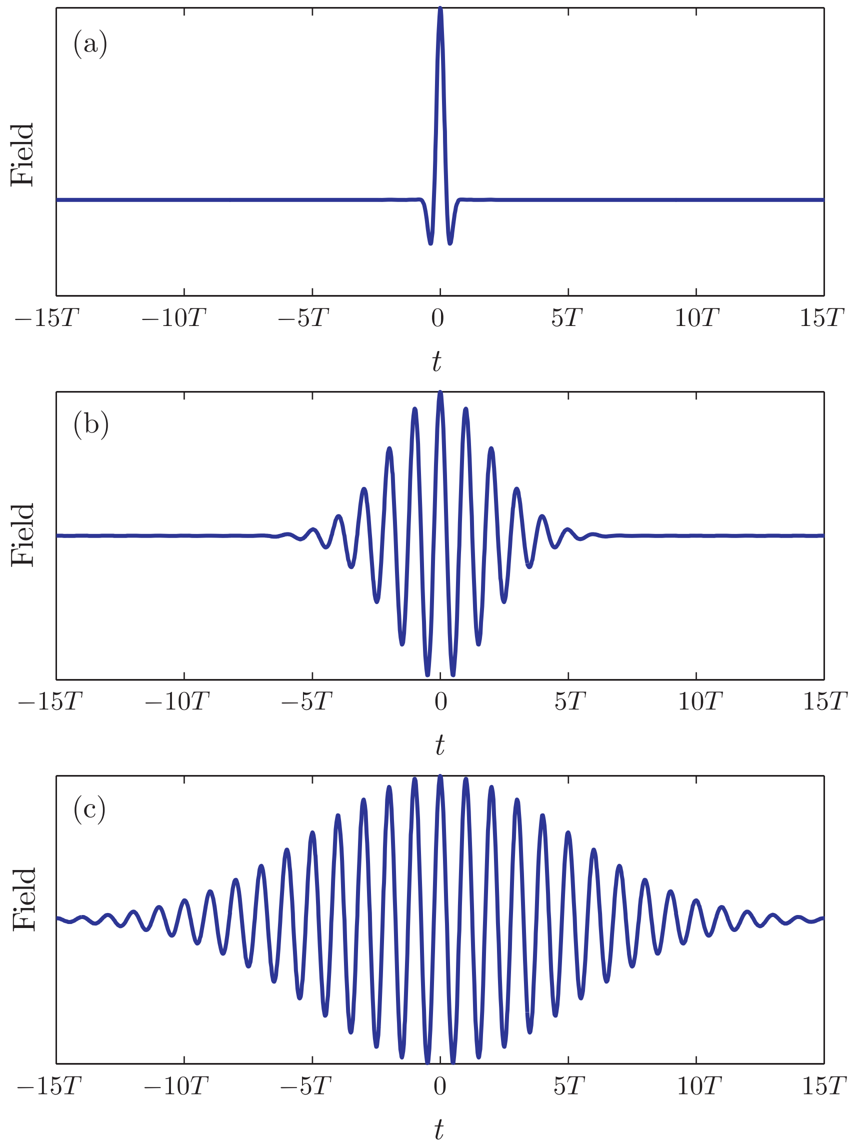


Figure 7.2 (a) Electric field (7.23) with $\tau = T/4$, where T is the period of the carrier frequency: $T = 2\pi/\omega_0$. (b) Electric field (7.23) with $\tau = 2T$. (c) Electric field (7.23) with $\tau = 5T$.

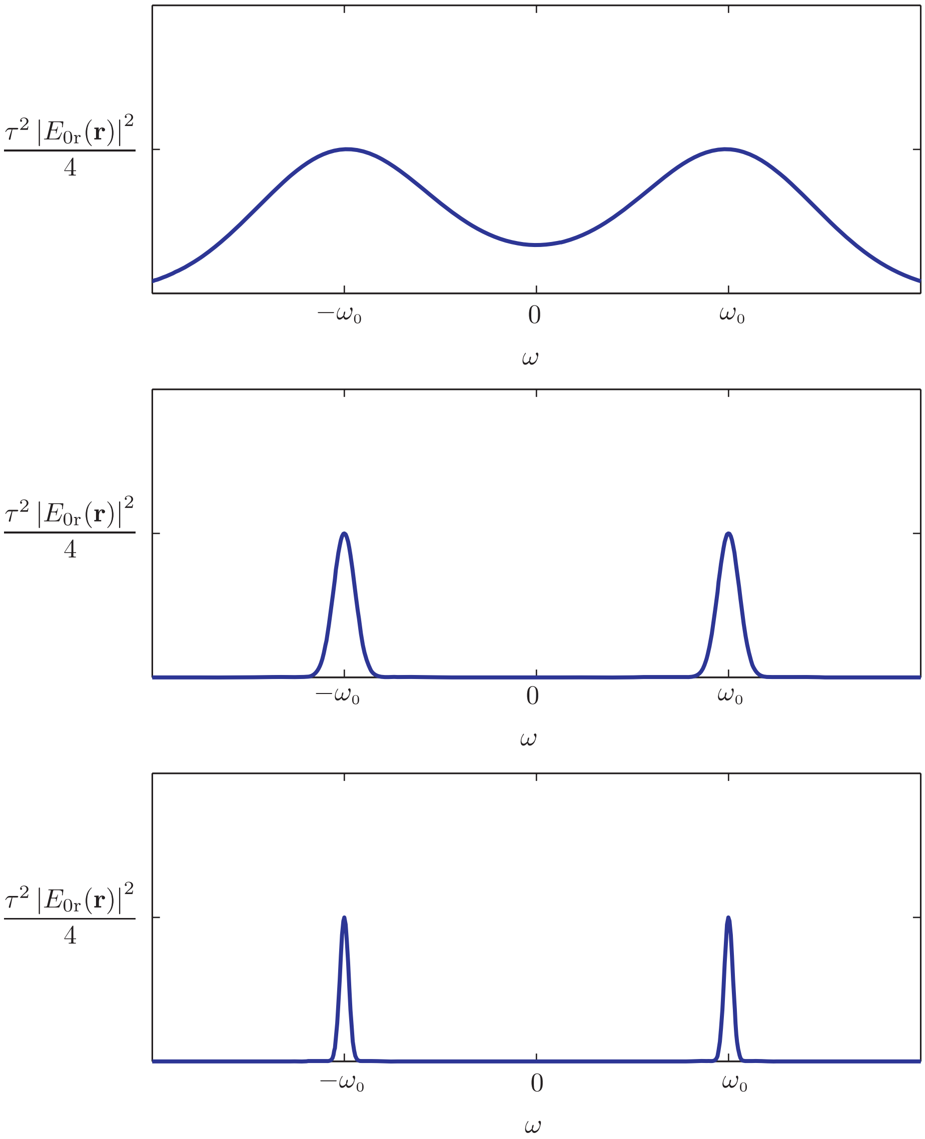


Figure 7.3 (a) Power spectrum based on (7.25) with $\tau = T/4$, where T is the period of the carrier frequency: $T = 2\pi/\omega_0$. (b) Power spectrum based on (7.25) with $\tau = 2T$. (c) Power spectrum based on (7.25) with $\tau = 5T$.

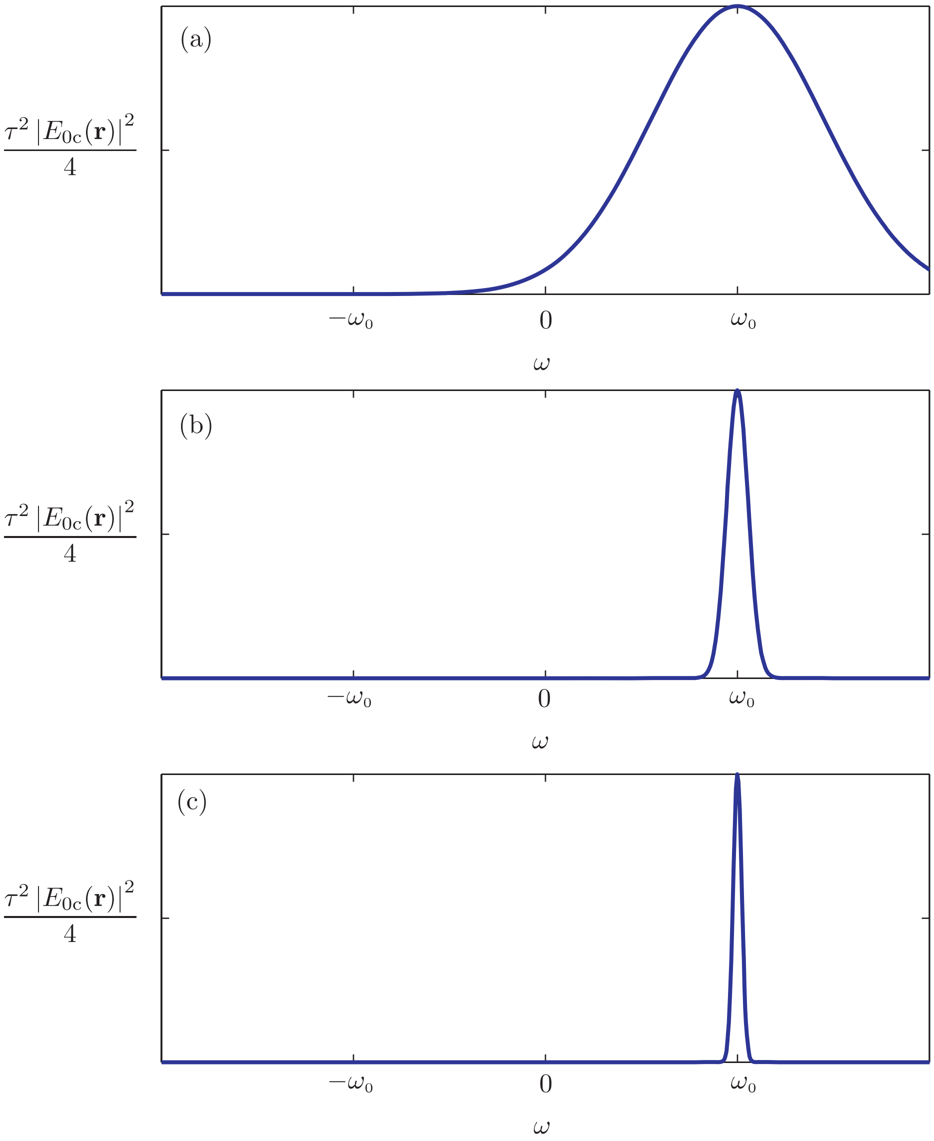


Figure 7.4 (a) Power spectrum based on (7.26) with $\tau = T/4$, where T is the period of the carrier frequency: $T = 2\pi/\omega_0$. (b) Power spectrum based on (7.26) with $\tau = 2T$. (c) Power spectrum based on (7.26) with $\tau = 5T$.

John Strutt (3rd Baron Rayleigh)
(1842–1919, British)

As head of the Cavendish laboratory, Rayleigh studied a wide variety of subjects. He developed the notion of group velocity and used it to understand the propagation of vibration in numerous systems. He won the Nobel prize in physics in 1904.

oscillations. While this is very convenient, this also points out why the complex notation should not be used for extremely short waveforms (e.g. for optical pulses a few femtoseconds long): There needs to be a sufficient number of oscillations within the waveform to make the rapid time average meaningful (as opposed to that in Fig. 7.2a).

Parseval's theorem (see P 0.31) imposes an interesting connection between the time-integral of the intensity and the frequency-integral of the power spectrum:

$$\int_{-\infty}^{\infty} I(\mathbf{r},t)dt = \int_{-\infty}^{\infty} I\left(\mathbf{r},\omega\right) d\omega \tag{7.27}$$

where

$$\begin{aligned} I(\mathbf{r},t) &\equiv \frac{n\epsilon_0 c}{2}\mathbf{E}(\mathbf{r},t)\cdot\mathbf{E}^*\left(\mathbf{r},t\right) \\ I\left(\mathbf{r},\omega\right) &\equiv \frac{n\epsilon_0 c}{2}\mathbf{E}\left(\mathbf{r},\omega\right)\cdot\mathbf{E}^*\left(\mathbf{r},\omega\right) \end{aligned} \tag{7.28}$$

The power spectrum $I\left(\mathbf{r},\omega\right)$ is observed when the waveform is sent into a spectral analyzer such as a diffraction spectrometer. Please excuse the potentially confusing notation (in wide usage): $I\left(\mathbf{r},\omega\right)$ is not the Fourier transform of $I(\mathbf{r},t)$!

7.5 Group Delay of a Wave Packet

When all **k**-vectors associated with a waveform point in the same direction, it becomes straightforward to predict the form of a pulse at different locations given knowledge of the waveform at another. Being able to predict the shape and arrival time of waveform is very important since a waveform traversing a material such as glass can undergo significant temporal *dispersion* as different frequency components experience different indices of refraction. For example, an ultra-short laser pulse traversing a glass window or a lens can emerge with significantly longer duration, owing to this effect. An example of this is given in the next section.

The fourier transform (7.21) gives the amplitudes of the individual plane wave components making up a waveform. We already know how to propagate individual plane waves through a material (see (2.22)). A phase shift associated with a displacement $\Delta\mathbf{r}$ modifies the field according to

$$\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)=\mathbf{E}\left(\mathbf{r}_0,\omega\right)e^{i\mathbf{k}(\omega)\cdot\Delta\mathbf{r}} \tag{7.29}$$

The $\mathbf{k}$-vector contains the pertinent information about the material via $k=n(\omega)\omega/c$. (A complex wave vector $\mathbf{k}$ may also be used if absorption or amplification is present.)

The procedure for finding what happens to a pulse when it propagates through a material is clear. Take the Fourier transform of the known incident pulse $\mathbf{E}\left(\mathbf{r}_0,t\right)$ to find the plane-wave coefficients $\mathbf{E}\left(\mathbf{r}_0,\omega\right)$ at the beginning of propagation. Apply the phase adjustment in (7.29) to find the plane wave coefficients $\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)$ at the end of propagation. Then take the inverse Fourier transform to determine the waveform $\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},t\right)$ at the new position:

$$\begin{aligned}\mathbf{E}(\mathbf{r}_0+\Delta\mathbf{r},t) &= \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}_0+\Delta\mathbf{r},\omega)e^{-i\omega t}\,d\omega \\ &= \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}_0,\omega)e^{i(\mathbf{k}(\omega)\cdot\Delta\mathbf{r}-\omega t)}\,d\omega \end{aligned}\tag{7.30}$$

The exponent in (7.29) is called the *phase delay* for the pulse propagation. It is often expanded in a Taylor series about a *carrier frequency* $\bar{\omega}$:

$$\mathbf{k}\cdot\Delta\mathbf{r}\cong\left[\mathbf{k}|_{\bar{\omega}}+\left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\bar{\omega}}(\omega-\bar{\omega})+\frac{1}{2}\left.\frac{\partial^2\mathbf{k}}{\partial\omega^2}\right|_{\bar{\omega}}(\omega-\bar{\omega})^2+\cdots\right]\cdot\Delta\mathbf{r} \tag{7.31}$$

The $\mathbf{k}$-vector has a sometimes-complicated frequency dependence through the functional form of $n(\omega)$. If we retain only the first two terms in this expansion then (7.30) becomes

$$\begin{aligned}\mathbf{E}(\mathbf{r}_0+\Delta\mathbf{r},t) &= \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}_0,\omega)e^{i\left(\left[\mathbf{k}(\bar{\omega})+\frac{\partial\mathbf{k}}{\partial\omega}\big|_{\bar{\omega}}(\omega-\bar{\omega})\right]\cdot\Delta\mathbf{r}-\omega t\right)}\,\mathrm{d}\omega \\ &= e^{i\left[\mathbf{k}(\bar{\omega})-\bar{\omega}\frac{\partial\mathbf{k}}{\partial\omega}\big|_{\bar{\omega}}\right]\cdot\Delta\mathbf{r}}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\mathbf{E}\left(\mathbf{r}_0,\omega\right)e^{-i\omega\left(t-\frac{\partial\mathbf{k}}{\partial\omega}\big|_{\bar{\omega}}\cdot\Delta\mathbf{r}\right)}\,\mathrm{d}\omega \\ &= e^{i[\mathbf{k}(\bar{\omega})\cdot\Delta\mathbf{r}-\bar{\omega}t']}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\mathbf{E}\left(\mathbf{r}_0,\omega\right)e^{-i\omega(t-t')}\,d\omega\end{aligned}\tag{7.32}$$

where in the last line we have used the definition

$$t'\equiv\left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\bar{\omega}}\cdot\Delta\mathbf{r}. \tag{7.33}$$

If we assume that the imaginary part of $\mathbf{k}$ is constant near $\bar{\omega}$ so that t' is real, i.e.

$$t'=\left.\frac{\partial\,\mathrm{Re}\,\mathbf{k}}{\partial\omega}\right|_{\bar{\omega}}\cdot\Delta\mathbf{r} \tag{7.34}$$

then the last integral in (7.32) is simply the Fourier transform of the original pulse with a new time argument, so we can carry out the integral to obtain

$$\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},t\right)=e^{i[\mathbf{k}(\bar{\omega})\cdot\Delta\mathbf{r}-\bar{\omega}t']}\mathbf{E}\left(\mathbf{r}_0,t-t'\right) \tag{7.35}$$

The first term in (7.35) gives an overall phase shift due to propagation, and is related to the phase velocity of the carrier frequency (see (7.18)):

$$v_p^{-1}\left(\bar{\omega}\right)=\frac{k\left(\bar{\omega}\right)}{\bar{\omega}} \tag{7.36}$$

To compare the intensity profile of the pulse at $\mathbf{r}_0+\Delta\mathbf{r}$ with the profile at $\mathbf{r}_0$ we compute the square magnitude of (7.35)

$$I\left(\mathbf{r}_0+\Delta\mathbf{r},t\right)\propto\left|\mathbf{E}\left(\mathbf{r}_0,t-t'\right)\right|^2e^{-2\,\mathrm{Im}\,\mathbf{k}(\bar{\omega})\cdot\Delta\mathbf{r}}. \tag{7.37}$$

In (7.37) we see that (to first order) t' is the time required for the pulse to traverse the displacement $\Delta\mathbf{r}$. The exponential in (7.37) describes the amplitude of the pulse at the new point, which may have changed during propagation due to absorption. The function $\partial\,\mathrm{Re}\,\mathbf{k}\,/\partial\omega\cdot\Delta\mathbf{r}$ is known as the *group delay function*, and in (7.34) it is evaluated only at the carrier frequency $\bar{\omega}$. Traditional group velocity is obtained by dividing the displacement $\Delta\mathbf{r}$ by the group delay time t' to obtain

$$v_g^{-1}\left(\bar{\omega}\right)=\left.\frac{\partial\mathrm{Re}\{k(\omega)\}}{\partial\omega}\right|_{\bar{\omega}} \tag{7.38}$$

Group delay (or group velocity) essentially tracks the center of the packet.

In our derivation we have assumed that the phase delay $\mathbf{k}(\omega)\cdot\Delta\mathbf{r}$ could be well-represented by the first two terms of the expansion (7.31). While this assumption gives results that are often useful, the other terms also play a role. In section 7.6 we'll study what happens if you keep the next higher order term in the expansion. We'll find that this term controls the rate at which the wave packet spreads as it travels. We should also note that there are times when the expansion (7.31) fails to converge (usually when $\bar{\omega}$ is near a resonance of the medium), and the expansion approach is not valid. We'll address how to analyze pulse propagation for these situations in section 7.7.

7.6 Quadratic Dispersion

A light pulse traversing a material in general undergoes dispersion because different frequency components take on different phase velocities. As an example, consider a short laser pulse traversing an optical component such as a lens or window, as depicted in Fig. 7.5. The light can undergo temporal dispersion, where a short light pulse spreads out in time with the different frequency components becoming separated (often called stretching or chirping). Dispersion can occur even if the optic absorbs very little of the light. Dispersion does not alter the power spectrum of the light pulse (7.28), ignoring absorption or reflections at the surfaces of the component. This is because the amplitude of $\mathbf{E}(\mathbf{r},\omega)$ does not change, but merely its phase according to (7.29). In other words, the plane-wave components that

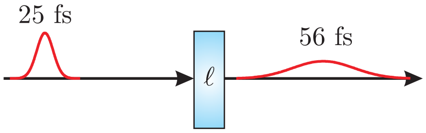


Figure 7.5 A 25 fs pulse traversing a 1 cm piece of BK7 glass.

make up the pulse can have their relative phases adjusted, while their individual amplitudes remain unchanged.

To compute the effect of dispersion on a pulse after it travels a distance in glass, we need to choose a specific pulse form. Suppose that just before entering the glass, the pulse has a Gaussian temporal profile given by (7.24). We'll place $\mathbf{r}_0$ at the start of the glass at $z = 0$ and assume that all plane-wave components travel in the $\hat{\mathbf{z}}$-direction, so that $\mathbf{k} \cdot \Delta\mathbf{r} = kz$. The polarization of the field will be the same for all frequencies. The Fourier transform of the Gaussian pulse is given in (7.26). Hence we have

$$\begin{aligned} \mathbf{E}(0,t) &= \mathbf{E}_0 e^{-t^2/2\tau^2} e^{-i\omega_0 t} \\ \mathbf{E}(0,\omega) &= \tau \mathbf{E}_0 e^{-\frac{\tau^2(\omega-\omega_0)^2}{2}} \end{aligned} \tag{7.39}$$

To find the field downstream we invoke (7.29), which gives the appropriate phase shift for each plane wave component:

$$\mathbf{E}(z,\omega) = \mathbf{E}(0,\omega) e^{ik(\omega)z} = \tau \mathbf{E}_0 e^{-\frac{\tau^2(\omega-\omega_0)^2}{2}} e^{ik(\omega)z} \tag{7.40}$$

To find the waveform at the new position z (where the pulse presumably has just exited the glass), we take the inverse Fourier transform of (7.40). However, before doing this we must specify the function $k(\omega)$. For example, if the glass material is replaced by vacuum, the wave number is simply $k_{\text{vac}}(\omega) = \omega/c$. In this case, the final waveform is

$$\mathbf{E}(z,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}_0 \tau e^{-\frac{\tau^2(\omega-\omega_0)^2}{2}} e^{i\frac{\omega}{c}z} e^{-i\omega t}\, d\omega = \mathbf{E}_0 e^{-\frac{1}{2}\left(\frac{t-z/c}{\tau}\right)^2} e^{i(k_0 z - \omega_0 t)} \quad \text{(vacuum)} \tag{7.41}$$

where $k_0 \equiv \omega_0/c$. Not surprisingly, after traveling a distance z though vacuum, the pulse looks identical to the original pulse, only its peak occurs at a later time z/c. The term $k_0 z$ appropriately adjusts the phase at different points in space so that at the time z/c the overall phase at z goes to zero.

Of course the functional form of the **k**-vector is different (and more complicated) in glass than in vacuum. One could represent the index with a multi-resonant Sellmeier equation with coefficients appropriate to the particular material (even more complicated than in P 2.2). For this example, however, we again resort to an expansion of the type (7.31), but this time we keep three terms. Let us choose the carrier frequency to be $\bar{\omega} = \omega_0$, so the expansion is

$$\begin{aligned} k(\omega) z &\cong k(\omega_0) z + \left.\frac{\partial k}{\partial \omega}\right|_{\omega_0} (\omega - \omega_0) z + \frac{1}{2} \left.\frac{\partial^2 k}{\partial \omega^2}\right|_{\omega_0} (\omega - \omega_0)^2 z + \cdots \\ &\cong k_0 z + v_g^{-1} (\omega - \omega_0) z + \alpha (\omega - \omega_0)^2 z \end{aligned} \tag{7.42}$$

where

$$k_0 \equiv k(\omega_0) = \frac{\omega_0 n(\omega_0)}{c} \tag{7.43}$$

$$v_g^{-1} \equiv \left.\frac{\partial k}{\partial \omega}\right|_{\omega_0} = \frac{n(\omega_0)}{c} + \frac{\omega_0 n'(\omega_0)}{c} \tag{7.44}$$

$$\alpha \equiv \frac{1}{2}\left.\frac{\partial^2 k}{\partial \omega^2}\right|_{\omega_0} = \frac{n'(\omega_0)}{c} + \frac{\omega_0 n''(\omega_0)}{2c} \tag{7.45}$$

With this approximation for $k(\omega)$, we are now able to perform the inverse Fourier transform on (7.40):

$$\begin{aligned}\mathbf{E}(z,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathbf{E}_0 \tau e^{-\frac{\tau^2(\omega-\omega_0)^2}{2}} e^{ik_0 z + iv_g^{-1}(\omega-\omega_0)z + i\alpha(\omega-\omega_0)^2 z} e^{-i\omega t}\, d\omega \\ &= \frac{\tau \mathbf{E}_0 e^{i(k_0 z - \omega_0 t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(\tau^2/2 - i\alpha z\right)(\omega-\omega_0)^2} e^{iv_g^{-1}(\omega-\omega_0)z - i(\omega-\omega_0)t}\, d\omega\end{aligned} \tag{7.46}$$

We can avoid considerable clutter if we change variables to $\omega' \equiv \omega - \omega_0$. Then the inverse Fourier transform becomes

$$\mathbf{E}(z,t) = \frac{\tau \mathbf{E}_0 e^{i(k_0 z - \omega_0 t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\tau^2}{2}\left(1 - i2\alpha z/\tau^2\right)\omega'^2 - i(t - z/v_g)\omega'}\, d\omega' \tag{7.47}$$

The above integral can be performed with the aid of (0.52). The result is

$$\begin{aligned}\mathbf{E}(z,t) &= \frac{\tau \mathbf{E}_0 e^{i(k_0 z - \omega_0 t)}}{\sqrt{2\pi}} \sqrt{\frac{\pi}{\frac{\tau^2}{2}\left(1 - i2\alpha z/\tau^2\right)}}\, e^{-\frac{(t - z/v_g)^2}{4\frac{\tau^2}{2}(1 - i2\alpha z/\tau^2)}} \\ &= \mathbf{E}_0 e^{i(k_0 z - \omega_0 t)} \frac{e^{\frac{i}{2}\tan^{-1}\frac{2\alpha z}{\tau^2}}}{\sqrt[4]{1 + \left(2\alpha z/\tau^2\right)^2}}\, e^{-\frac{(t - z/v_g)^2}{2\tau^2\left(1 + \left(2\alpha z/\tau^2\right)^2\right)}\left(1 + i2\alpha z/\tau^2\right)}\end{aligned} \tag{7.48}$$

Next, we spruce up the appearance of this rather cumbersome formula as follows:

$$\mathbf{E}(z,t) = \frac{\mathbf{E}_0}{\sqrt{\mathrm{T}(z)/\tau}} e^{-\frac{1}{2}\left[\frac{t - z/v_g}{\mathrm{T}(z)}\right]^2} e^{-\frac{i}{2}\left[\frac{t - z/v_g}{\mathrm{T}(z)}\right]^2 \Phi(z) + i(k_0 z - \omega_0 t) + i\frac{1}{2}\tan^{-1}\Phi(z)} \tag{7.49}$$

where

$$\Phi(z) \equiv \frac{2\alpha}{\tau^2} z \tag{7.50}$$

and

$$\mathrm{T}(z) \equiv \tau\sqrt{1 + \Phi^2(z)} \tag{7.51}$$

We can immediately make a few observation about (7.49). First, note that at $z = 0$ (i.e. zero thickness of glass), (7.49) reduces to the input pulse given in (7.39), as we would expect. Secondly, the peak of the pulse moves at speed v_g since the term

$$e^{-\frac{1}{2}\left[\frac{t - z/v_g}{\mathrm{T}(z)}\right]^2}$$

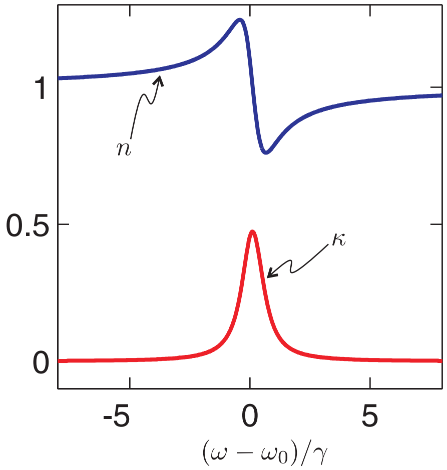


Figure 7.6 Index of refraction in the neighborhood of a resonance.

controls the pulse amplitude, while the other terms (multiplied by i) in the exponent of (7.49) merely alter the phase. Also note that the duration of the pulse increases and its peak intensity decreases as it travels, since $\mathrm{T}(z)$ increases with z. In P 7.9 we will find that (7.49) also predicts that for large z, the field of the spread-out pulse oscillates less rapidly at the beginning of the pulse than at the end (assuming $\alpha > 0$). This phenomenon is known as "chirp", and indicates that red frequencies get ahead of blue frequencies during propagation since they experience a lower index of refraction.

While we have derived these results for the specific case of a Gaussian pulse, the results are applicable to other pulse shapes also. Although the exact details will vary by pulse shape, all short pulses eventually broaden and chirp as they propagate through a dispersive medium such as glass (as long as the medium responds linearly to the field). Higher order terms in the expansion (7.31) to the spreading, chirping, and other deformation of the pulse as it propagates, but the become progressively more cumbersome to study analytically.

7.7 Generalized Context for Group Delay

The expansion of $\mathbf{k}(\omega)$ in (7.31) is inconvenient if the frequency content (bandwidth) of a waveform encompasses a substantial portion of a resonance structure such as shown in Fig. 7.6. In this case, it becomes necessary to retain a large number of terms in (7.31) to describe accurately the phase delay $\mathbf{k}(\omega)\cdot\Delta\mathbf{r}$. Moreover, if the bandwidth of the waveform is wider than the spectral resonance of the medium (as shown in Fig. 7.7), the series altogether fails to converge. These difficulties have led to the traditional viewpoint that group velocity loses meaning for broadband waveforms (interacting with a resonance in a material) since it is associated with the second term in the expansion (7.31), evaluated at a carrier frequency $\bar{\omega}$. In this section, we study a broader context for group velocity (or rather its inverse, group delay), which is always valid, even for broadband pulses where the expansion (7.31) utterly fails. The analysis avoids the expansion and so is not restricted to a narrowband context.

We are interested in the *arrival time* of a waveform (or pulse) to a point, say, where a

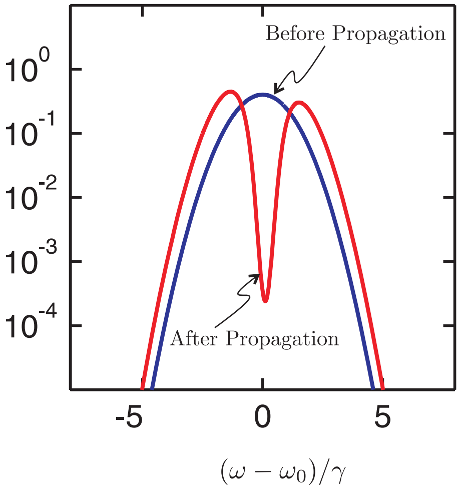


Figure 7.7 Normalized spectrum of a broadband pulse before and after propagation through an absorbing medium.

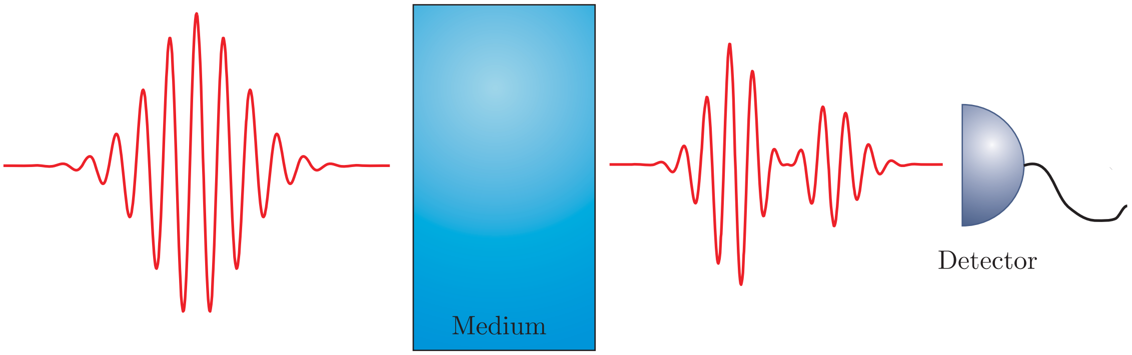


Figure 7.8 Pulse undergoing distortion during transit.

detector is located. The definition of the arrival time of pulse energy need only involve the Poynting flux (or the intensity), since it alone is responsible for energy transport. To deal with arbitrary broadband pulses, the arrival time should avoid presupposing a specific pulse shape, since the pulse may evolve in complicated ways during propagation. For example, the pulse peak or the midpoint on the rising edge of a pulse are poor indicators of arrival time if the pulse contains multiple peaks or a long and non-uniform rise time.

For the reasons given, we use a time expectation integral (or time "center-of-mass") to describe the arrival time of the pulse:

$$\langle t \rangle_{\mathbf{r}} \equiv \int_{-\infty}^{\infty} t\rho(\mathbf{r}, t)dt \qquad (7.52)$$

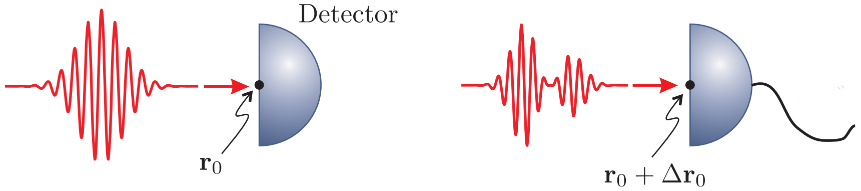


Figure 7.9 Transit time defined as the difference between arrival time at two points.

Here $\rho(\mathbf{r}, t)$ is a normalized distribution function associated with the intensity:

$$\rho(\mathbf{r}, t) \equiv \frac{I(\mathbf{r}, t)}{\int\limits_{-\infty}^{\infty} I(\mathbf{r}, t)\; dt} \tag{7.53}$$

For simplification, we assume that the light travels in a uniform direction. As we shall see, the function $dk/d\omega$ (inverse of group velocity) is linked to this temporal expectation of the incoming intensity.

Consider a pulse as it travels from point $\mathbf{r}_0$ to point $\mathbf{r} = \mathbf{r}_0 + \Delta\mathbf{r}$ in a homogeneous medium (see Fig. 7.9). The difference in arrival times at the two points is

$$\Delta t \equiv \langle t \rangle_{\mathbf{r}} - \langle t \rangle_{\mathbf{r}_0} \tag{7.54}$$

The pulse shape can evolve in complicated ways between the two points, spreading with different portions being absorbed (or amplified) during transit. Nevertheless, (7.54) renders an unambiguous time interval between the passage of the pulse center at each point.

This difference in arrival time can be shown to consist of two terms (see P 7.12):

$$\Delta t = \Delta t_G\,(\mathbf{r}) + \Delta t_R\,(\mathbf{r}_0) \tag{7.55}$$

The first term, called the *net group delay*, dominates if the field waveform is initially symmetric in time (e.g. an unchirped Gaussian). It amounts to a spectral average of the group delay function taken with respect to the spectral content of the pulse arriving at the final point $\mathbf{r} = \mathbf{r}_0 + \Delta\mathbf{r}$:

$$\Delta t_G\,(\mathbf{r}) = \int\limits_{-\infty}^{\infty} \rho\,(\mathbf{r}, \omega) \left(\frac{\partial \mathrm{Re}\mathbf{k}}{\partial \omega} \cdot \Delta\mathbf{r} \right) d\omega \tag{7.56}$$

where the spectral weighting function is

$$\rho(\mathbf{r}, \omega) \equiv \frac{I(\mathbf{r}, \omega)}{\int\limits_{-\infty}^{\infty} I(\mathbf{r}, \omega')\; d\omega'} \tag{7.57}$$

and $I\,(\mathbf{r}, \omega)$ is given in (7.28). The two curves in Fig. 7.7 show $\rho\,(\mathbf{r}_0, \omega)$ (before propagation) and $\rho\,(\mathbf{r}, \omega)$ (after propagation) for an initially Gaussian pulse. As seen in (7.57), the pulse travel time depends on the spectral shape of the pulse at the end of propagation.

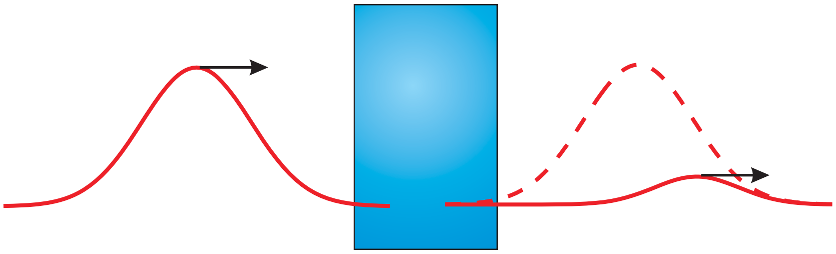


Figure 7.10 Narrowband pulse traversing an absorbing medium.

Note the close resemblance between the formulas (7.52) and (7.56). Both are expectation integrals. The former is executed as a "center-of-mass" integral on time; the latter is executed in the frequency domain on $\partial \mathrm{Re}\mathbf{k} \cdot \Delta\mathbf{r}/\partial\omega$, the group delay function. The group delay at every frequency present in the pulse influences the result. If the pulse has a narrow bandwidth in the neighborhood of $\bar{\omega}$, the integral reduces to $\partial \mathrm{Re}\mathbf{k}/\partial\omega|_{\bar{\omega}} \cdot \Delta\mathbf{r}$, in agreement with (7.38) (see P 7.10). The net group delay depends only on the spectral content of the pulse, independent of its temporal organization (i.e., the phase of $\mathbf{E}(\mathbf{r},\omega)$ has no influence). Only the real part of the $\mathbf{k}$-vector plays a direct role in (7.56).

The second term in (7.55), called the reshaping delay, represents a delay that arises solely from a reshaping of the spectral amplitude. This term takes into account how the pulse time center-of-mass shifts as portions of the spectrum are removed (or added). It is computed at $\mathbf{r}_0$ *before propagation takes place*:

$$\Delta t_R(\mathbf{r}_0) = \langle t\rangle_{\mathbf{r}_0}\Big|_{\text{altered}} - \langle t\rangle_{\mathbf{r}_0} \tag{7.58}$$

Here $\langle t\rangle_{\mathbf{r}_0}$ represents the usual arrival time of the pulse at the initial point $\mathbf{r}_0$, according to (7.52). The intensity at this point is associated with a field $\mathbf{E}(\mathbf{r}_0, t)$, connected to $\mathbf{E}(\mathbf{r}_0,\omega)$ through an inverse Fourier transform (7.20). On the other hand, $\langle t\rangle_{\mathbf{r}_0}\Big|_{\text{altered}}$ is the arrival time of a pulse associated with the modified field $\mathbf{E}(\mathbf{r}_0,\omega)\, e^{-\mathrm{Im}\mathbf{k}\cdot\Delta\mathbf{r}}$. Notice that $\mathbf{E}(\mathbf{r}_0,\omega)\, e^{-\mathrm{Im}\mathbf{k}\cdot\Delta\mathbf{r}}$ is still evaluated at the initial point $\mathbf{r}_0$. Only the spectral amplitude (not the phase) is modified, according to what is anticipated to be lost (or gained) during the trip. In contrast to the net group delay, the reshaping delay is sensitive to how a pulse is organized. The reshaping delay is negligible if the pulse is initially symmetric (in amplitude and phase) before propagation. The reshaping delay also goes to zero in the narrowband limit, and the total delay reduces to the net group delay.

As an example, consider the Gaussian pulse (7.24) with duration either $\tau_1 = 10/\gamma$ (narrowband) or $\tau_2 = 1/\gamma$ (broadband), where γ is the damping term in the Lorentz model described in section 2.3. Let the pulse travel a distance $\Delta\mathbf{r} = \hat{\mathbf{z}}c/(10\gamma)$through the absorbing medium (as depicted in Fig. 7.10), which has a resonance at frequency ω_0. The index of refraction is shown in Fig. 7.11. Its resonance has a width of γ. Fig. 7.12 shows the delay between the pulse arrival times at $\mathbf{r}_0$ and $\mathbf{r} = \mathbf{r}_0 + \Delta\mathbf{r}$ as the pulse's central frequency $\mathbf{r} = \mathbf{r}_0 + \Delta\mathbf{r}$ is varied in the neighborhood of the resonance. The solid line gives the total delay $\Delta t \cong \Delta t_G(\mathbf{r})$ experienced by the narrowband pulse in traversing the displacement. The reshaping delay in this case is negligible (i.e. $\Delta t_R(\mathbf{r}) \cong 0$) and is shown by the dotted line. Near resonance, *superluminal* behavior results as the transit time for the pulse becomes small and even negative. The peak of the attenuated pulse exits the medium even

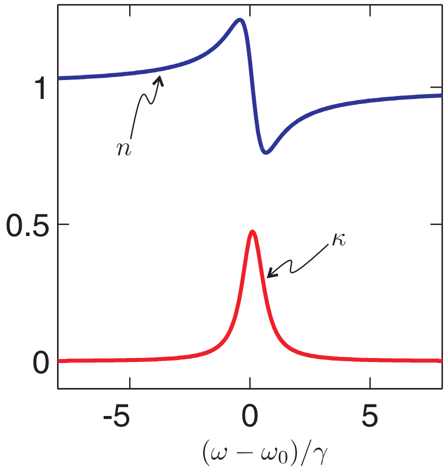


Figure 7.11 Real and imaginary parts of the refractive index for an absorptive medium.

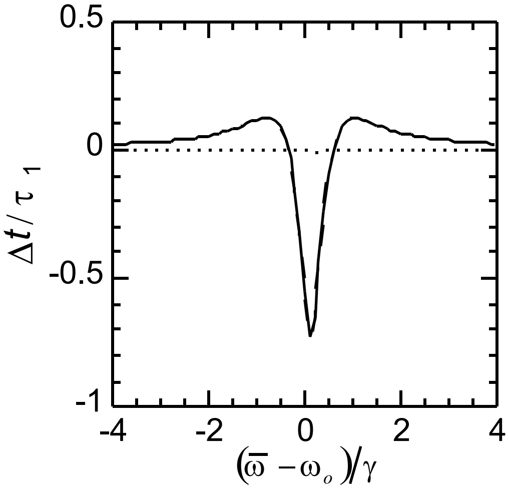


Figure 7.12 Pulse transit time for a narrowband pulse in an absorbing medium as a function of carrier frequency.

before the peak of the incoming pulse enters the medium! Keep in mind that the exiting pulse is tiny and resides well within the original envelope of the pulse propagated forward at speed c, as indicated in Fig. 7.10. Thus, with or without the absorbing material in place, the signal is detectable just as early. Similar results can be obtained in amplifying media.

As the injected pulse becomes more sharply defined in time, the superluminal behavior does not persist. Fig. 7.13 shows the clearly subluminal transit time for the broadband pulse with the shorter duration τ_2. While Fig. 7.12 can be generated using the traditional narrowband context of group delay, Fig. 7.13 requires the new context presented in this section. It demonstrates that sharply defined waveforms (i.e. broadband) do not propagate superluminally. In addition, while a long smooth pulse can exhibit so-called superlumi-

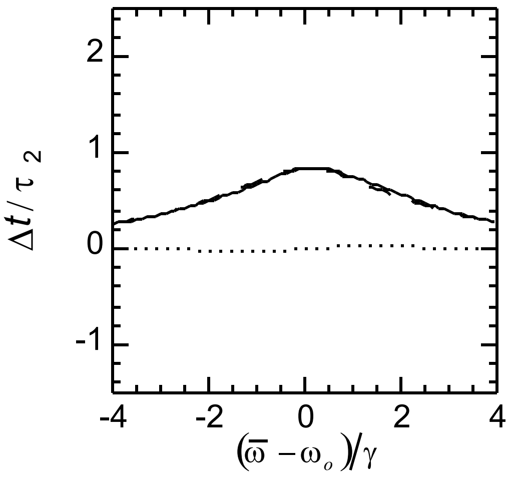


Figure 7.13 Pulse transit time for a broadband pulse in an absorbing medium.

nal behavior over short propagation distances, the behavior does not persist as the pulse spectrum is modified by the medium.

As we have mentioned, the group delay function indicates the average arrival of field energy to a point. Since this is only part of the whole energy story, there is no problem when it becomes superluminal. The overly rapid appearance of electromagnetic energy at one point and its simultaneous disappearance at another point merely indicates an exchange of energy between the electric field and the medium. In appendix 7.A we discuss the energy transport velocity (involving *all* energy—strictly luminal) and the velocity of locus of electromagnetic *field* energy.

Appendix 7.A Causality and Exchange of Energy with the Medium

In accordance with Poynting's theorem (2.50), the total energy density stored in an electromagnetic field and in a medium is given by

$$u(\mathbf{r},t) = u_{\text{field}}(\mathbf{r},t) + u_{\text{exchange}}(\mathbf{r},t) + u(\mathbf{r},-\infty) \tag{7.59}$$

This expression for the energy density includes all (relevant) forms of energy, including a non-zero integration constant $u(\mathbf{r},-\infty)$ corresponding to energy stored in the medium before the arrival of any pulse (important in the case of an amplifying medium). $u_{\text{field}}(\mathbf{r},t)$ and $u_{\text{exchange}}(\mathbf{r},t)$ are both zero before the arrival of the pulse (i.e. at $t=-\infty$). In addition, $u_{\text{field}}(\mathbf{r},t)$, given by (2.52), returns to zero after the pulse has passed (i.e. at $t=+\infty$).

The time-dependent accumulation of energy transferred into the medium from the field is given by

$$u_{\text{exchange}}(\mathbf{r},t) = \int_{-\infty}^{t} \mathbf{E}(\mathbf{r},t') \cdot \frac{\partial \mathbf{P}(\mathbf{r},t')}{\partial t'}\, dt' \tag{7.60}$$

where we ignore the possibility of any free current $\mathbf{J}_{\text{free}}$ in (2.53). As u_{exchange} increases, the energy in the medium increases. Conversely, as u_{exchange} decreases, the medium surrenders energy to the electromagnetic field. While it is possible for u_{exchange} to become negative, the combination $u_{\text{exchange}} + u(-\infty)$ (i.e. the net energy in the medium) can never go negative since a material cannot surrender more energy than it has to begin with.

We next consider the concept of the energy transport velocity. Poynting's theorem (2.50) has the form of a continuity equation which when integrated spatially over a small volume V yields

$$\oint_A \mathbf{S} \cdot d\mathbf{a} = -\frac{\partial}{\partial t} \int_V u \, dV \tag{7.61}$$

where the left-hand side has been transformed into an surface integral representing the power leaving the volume. Let the volume be small enough to take $\mathbf{S}$ to be uniform throughout V. The energy transport velocity (directed along $\mathbf{S}$) is then defined to be the *effective* speed at which the energy contained in the volume (i.e. the result of the volume integral) would need to travel in order to achieve the power transmitted through one side of the volume (e.g. the power transmitted through one end of a tiny cylinder aligned with $\mathbf{S}$). The energy transport velocity as traditionally written is then

$$\mathbf{v}_E \equiv \frac{\mathbf{S}}{u} \tag{7.62}$$

When the total energy density u is used in computing (7.62), the energy transport velocity has a *fictitious* nature; it is not the actual velocity of the total energy (since part is stationary), but rather the effective velocity necessary to achieve the same energy transport that the electromagnetic flux alone delivers. There is no behind-the-scenes flow of mechanical energy. Note that if only u_{field} is used in evaluating (7.62), the Cauchy-Schwartz inequality (i.e. $\alpha^2 + \beta^2 \geq 2\alpha\beta$) ensures an energy transport velocity v_E that is strictly bounded by the speed of light in vacuum c. The total energy density u at least as great as the field energy density u_{field}. Hence, this strict luminality is maintained.

Since the point-wise energy transport velocity defined by (7.62) is strictly luminal, it follows that the global energy transport velocity (the average speed of *all energy*) is also bounded by c. To obtain the global properties of energy transport, we begin with a weighted average of the energy transport velocity at each point in space. A suitable weighting parameter is the energy density at each position. The global energy transport velocity is then

$$\langle \mathbf{v}_E \rangle \equiv \frac{\int \mathbf{v}_E u \, d^3r}{\int u \, d^3r} = \frac{\int \mathbf{S} \, d^3r}{\int u \, d^3r} \tag{7.63}$$

where we have substituted from (7.62). The integral is taken over all relevant space (note $d^3r = dV$).

Integration by parts leads to

$$\langle \mathbf{v}_E \rangle = -\frac{\int \mathbf{r} \nabla \cdot \mathbf{S} \, d^3r}{\int u \, d^3r} = \frac{\int \mathbf{r} \frac{\partial u}{\partial t} \, d^3r}{\int u \, d^3r} \tag{7.64}$$

where we have assumed that the volume for the integration encloses all energy in the system and that the field near the edges of this volume is zero. Since we have included all energy, Poynting's theorem (2.50) can be written with no source terms (i.e. $\nabla \cdot \mathbf{S} + \partial u/\partial t = 0$). This means that the total energy in the system is conserved and is given by the integral in the denominator of (7.64). This allows the derivative to be brought out in front of the entire expression giving

$$\langle \mathbf{v}_E \rangle = \frac{\partial \langle \mathbf{r} \rangle}{\partial t} \tag{7.65}$$

where

$$\langle \mathbf{r} \rangle \equiv \frac{\int \mathbf{r} u \; d^3 r}{\int u \; d^3 r} \tag{7.66}$$

The latter expression represents the "center-of-mass" or centroid of the total energy in the system.

This precise relationship between the energy transport velocity and the centroid requires that *all* forms of energy be included in the energy density u. If, for example, only the field energy density u_{field} is used in defining the energy transport velocity, the steps leading to (7.66) would not be possible. Although (7.66) guarantees that the centroid of the *total* energy moves strictly luminally, there is no such limitation on the centroid of field energy alone. Explicitly we have

$$\left\langle \frac{\mathbf{S}}{u_{\text{field}}} \right\rangle \neq \frac{\partial}{\partial t} \frac{\int \mathbf{r} u_{\text{field}} d^3 r}{\int u_{\text{field}} d^3 r} \tag{7.67}$$

While, as was pointed out, the left-hand side of (7.67) is strictly luminal, the right-hand side can easily exceed c as the medium exchanges energy with the field. In an amplifying medium exhibiting superluminal behavior, the rapid appearance of a pulse downstream is merely an artifact of not recognizing the energy already present in the medium until it converts to the form of field energy. The traditional group velocity is connected to this method of accounting, which is why it can become superluminal. Note the similarity between (7.52), which is a time center-of-mass, and the right-hand side of (7.67), which is the spatial center of mass. Both expressions can be connected to group velocity. Group velocity tracks the presence of field energy alone without necessarily implying the actual motion of that energy.

It is enlightening to consider u_{exchange} within a frequency-domain context. We utilize the field represented in terms of an inverse Fourier transform (7.20). Similarly, the polarization $\mathbf{P}$ can be written as an inverse Fourier transform:

$$\mathbf{P}(\mathbf{r}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{P}(\mathbf{r}, \omega) e^{-i\omega t} d\omega \Rightarrow \frac{\partial \mathbf{P}(\mathbf{r}, t)}{\partial t} = \frac{-i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega \mathbf{P}(\mathbf{r}, \omega) e^{-i\omega t} d\omega \tag{7.68}$$

In an isotropic medium, the polarization for an individual plane wave can be written in terms of the linear susceptibility defined in (1.39):

$$\mathbf{P}(\mathbf{r}, \omega) = \epsilon_0 \chi(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega) \tag{7.69}$$

With (7.21), (7.68), and (7.69), the exchange energy density (7.60), can be written as

$$u_{\text{exchange}}(\mathbf{r}, t) = \int_{-\infty}^{t} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r}, \omega') e^{-i\omega' t'} d\omega' \right] \cdot \left[\frac{-i\epsilon_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega \chi(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega) e^{-i\omega t'} d\omega \right] dt' \tag{7.70}$$

After interchanging the order of integration, the expression becomes

$$u_{\text{exchange}}(\mathbf{r},t) = -i\epsilon_0 \int\limits_{-\infty}^{\infty} d\omega\omega\chi\left(\mathbf{r},\omega\right)\mathbf{E}\left(\mathbf{r},\omega\right)\cdot \int\limits_{-\infty}^{\infty} d\omega'\mathbf{E}\left(\mathbf{r},\omega'\right)\frac{1}{2\pi}\int\limits_{-\infty}^{t} e^{-i(\omega+\omega')t'}dt' \tag{7.71}$$

The final integral in (7.71) becomes the delta function when t goes to $+\infty$. In this case, the middle integral can also be performed. Therefore, after the point $\mathbf{r}$ experiences the entire pulse, the final amount of energy density exchanged between the field and the medium at that point is

$$u_{\text{exchange}}\left(\mathbf{r},+\infty\right) = -i\epsilon_0 \int\limits_{-\infty}^{\infty} \omega\chi\left(\mathbf{r},\omega\right)\mathbf{E}\left(\mathbf{r},\omega\right)\cdot\mathbf{E}\left(\mathbf{r},-\omega\right)\ d\omega \tag{7.72}$$

In this appendix, for convenience we consider the fields to be written using real notation. Then we can employ the symmetry (7.22) along with the symmetry

$$\mathbf{P}^*\left(\mathbf{r},\omega\right) = \mathbf{P}\left(\mathbf{r},-\omega\right) \tag{7.73}$$

and hence

$$\chi^*\left(\mathbf{r},\omega\right) = \chi\left(\mathbf{r},-\omega\right). \tag{7.74}$$

Then we obtain

$$u_{\text{exchange}}\left(\mathbf{r},+\infty\right) = \epsilon_0 \int\limits_{-\infty}^{\infty} \omega\text{Im}\chi\left(\mathbf{r},\omega\right)\mathbf{E}\left(\mathbf{r},\omega\right)\cdot\mathbf{E}^*\left(\mathbf{r},\omega\right)\ d\omega \tag{7.75}$$

This expression describes the net exchange of energy density after all action has finished. It involves the power spectrum of the pulse. We can modify this formula in an intuitive way so that it describes the exchange energy density for any time during the pulse. The principle of causality guides us in considering how the medium perceives the electric field for any time.

Since the medium is unable to anticipate the spectrum of the entire pulse before experiencing it, the material responds to the pulse according to the history of the field up to each instant. In particular, the material has to be prepared for the possibility of an abrupt cessation of the pulse at any moment, in which case all exchange of energy with the medium immediately ceases. In this extreme scenario, there is no possibility for the medium to recover from previously incorrect attenuation or amplification, so it must have gotten it right already.

If the pulse were in fact to abruptly terminate at a given instant, then the expression (7.75) would immediately apply since the pulse would be over; it would not be necessary to integrate the inverse Fourier transform (7.21) beyond the termination time t for which all contributions are zero. Causality requires that the medium be indifferent to whether a pulse actually terminates if it hasn't happened yet. Therefore, (7.75) applies at all times where the spectrum (7.21) is evaluated over that portion of the field previously experienced by the medium.

The following is then an exact representation for the exchange energy density defined in (7.60):

$$u_{\text{exchange}}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{\infty} \omega \text{Im}\chi\left(\mathbf{r}, \omega\right) \mathbf{E}_t\left(\mathbf{r}, \omega\right) \cdot \mathbf{E}_t^*\left(\mathbf{r}, \omega\right)\ d\omega \tag{7.76}$$

where

$$E_t\left(\mathbf{r}, \omega\right) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} E\left(\mathbf{r}, t'\right) e^{i\omega t'}\ dt' \tag{7.77}$$

This time dependence enters only through $\mathbf{E}_t\left(\mathbf{r}, \omega\right) \cdot \mathbf{E}_t^*\left(\mathbf{r}, \omega\right)$, known as the *instantaneous power spectrum.*

The expression (7.76) for the exchange energy reveals physical insights into the manner in which causal dielectric materials exchange energy with different parts of an electromagnetic pulse. Since the function $E_t\left(\omega\right)$ is the Fourier transform of the pulse truncated at the current time t and set to zero thereafter, it can include many frequency components that are not present in the pulse taken in its entirety. This explains why the medium can respond differently to the front of a pulse than to the back. Even though absorption or amplification resonances may lie outside of the spectral envelope of a pulse taken in its entirety, the instantaneous spectrum on a portion of the pulse can momentarily lap onto or off of resonances in the medium.

In view of (7.76) and (7.77) it is straightforward to predict when the electromagnetic energy of a pulse will exhibit superluminal or subluminal behavior. In section 7.7, we saw that this behavior is controlled by the group velocity function. However, with (7.76) and (7.77), it is not necessary to examine the group velocity directly, but only the imaginary part of the susceptibility $\chi\left(\mathbf{r}, \omega\right)$.

If the entire pulse passing through point $\mathbf{r}$ has a spectrum in the neighborhood of an amplifying resonance, but not on the resonance, superluminal behavior can result (Chiao effect). The instantaneous spectrum during the front portion of the pulse is generally wider and can therefore lap onto the nearby gain peak. The medium accordingly amplifies this perceived spectrum, and the front of the pulse grows. The energy is then returned to the medium from the latter portion of the pulse as the instantaneous spectrum narrows and withdraws from the gain peak. The effect is not only consistent with the principle of causality, it is a direct and general consequence of causality as demonstrated by (7.76) and (7.77).

As an illustration, consider the broadband waveform with $\tau_2 = 1/\gamma$ described in section 7.7. Consider an amplifying medium with index shown in Fig. 7.14 with the amplifying resonance (negative oscillator strength) set on the frequency $\omega_0 = \bar{\omega} + 2\gamma$, where $\bar{\omega}$ is the carrier frequency. Thus, the resonance structure is centered a modest distance above the carrier frequency, and there is only minor spectral overlap between the pulse and the resonance structure.

Superluminal behavior can occur in amplifying materials when the forward edge of a narrow-band pulse can receive extra amplification. Fig. 7.15(a) shows the broadband waveform experienced by the initial position $\mathbf{r}_0$ in the medium. Fig. 7.15(b) shows the real and imaginary parts of the refractive index in the neighborhood of the carrier frequency $\bar{\omega}$. Fig. 7.15(c) depicts the exchange energy density u_{exchange} as a function of time, where

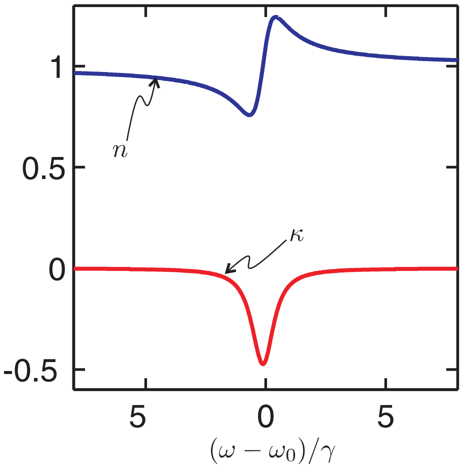


Figure 7.14 Real and imaginary parts of the refractive index for an amplifying medium.

rapid oscillations have been averaged out. The overshooting of the curve indicates excess amplification during the early portion of the pulse. The energy is then returned (in part) to the medium during the later portion of the pulse, a clear indication of superluminal behavior. Fig. 7.15(d) displays the instantaneous power spectrum (used in computing u_{exchange}) evaluated at various times during the pulse. The corresponding times are indicated with vertical lines in both Figs. 7.15(a) and 7.21(c). The format of each vertical line matches a corresponding spectral curve. The instantaneous spectrum exhibits wings, which lap onto the nearby resonance and vary in strength depending on when the integral (7.77) truncates the pulse. As the wings grow and access the neighboring resonance, the pulse extracts excess energy from the medium. As the wings diminish, the pulse surrenders that energy back to the medium, which gives the appearance of superluminal transit times.

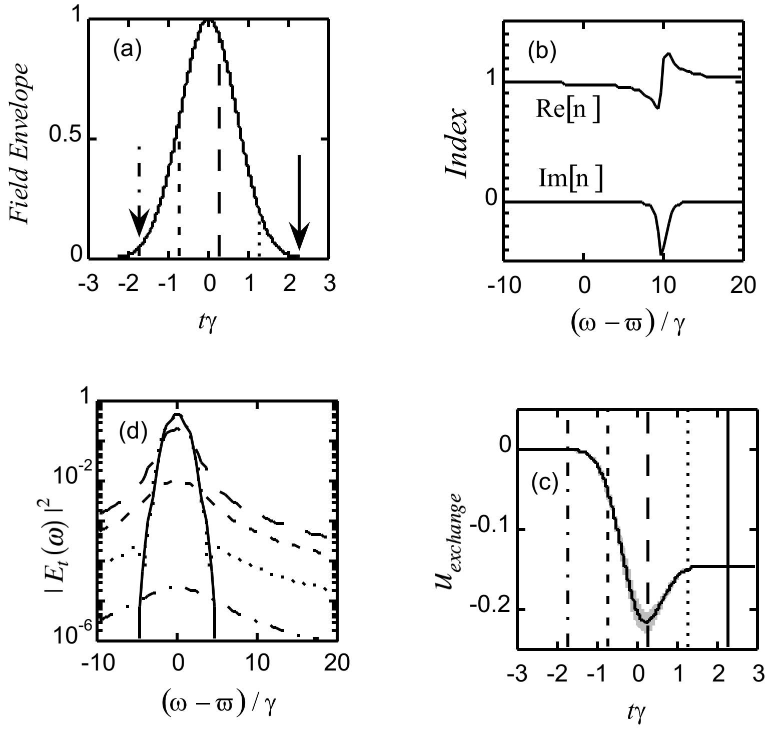


Figure 7.15 (a) Electric field envelope in units of E_0. Vertical lines indicate times for assessment of the instantaneous spectrum. (b) Refractive index associated with an *amplifying* resonance. (c) Exchange energy density in units of $\epsilon_0 E_0^2/2$. (d) Instantaneous spectra of the field pulse in units of E_0^2/γ^2. Spectra are assessed at the times indicated in (a) and (c).

Exercises

7.2 Intensity

P7.1 (a) Let $\hat{\mathbf{x}}E_1 e^{i(kz-\omega t)}$ and $\hat{\mathbf{x}}E_2 e^{i(-kz-\omega t)}$ be two counter-propagating plane waves where E_1 and E_2 are both real. Show that their sum can be written as

$$\hat{\mathbf{x}}E_{\text{tot}}\left(z\right)e^{i(\Phi(z)-\omega t)}$$

where

$$E_{\text{tot}}\left(z\right) = E_1\sqrt{\left(1-\frac{E_2}{E_1}\right)^2 + 4\frac{E_2}{E_1}\cos^2 kz}$$

and

$$\Phi\left(z\right) = \tan^{-1}\left[\frac{(1-E_2/E_1)}{(1+E_2/E_1)}\tan kz\right]$$

Outside the range $-\frac{\pi}{2} \le kz \le \frac{\pi}{2}$ the pattern repeats.

(b) Suppose that two counter-propagating laser fields have separate intensities, I_1 and $I_2 = I_1/100$. The ratio of the fields is then $E_2/E_1 = 1/10$. In the standing interference pattern that results, what is the ratio of the peak *intensity* to the minimum *intensity*? Are you surprised how high this is?

P7.2 Equation (7.11) implies that there is no *interference* between fields that are polarized along orthogonal dimensions. That is, the intensity of

$$\mathbf{E}(\mathbf{r},t) = \hat{\mathbf{x}}E_0 e^{i[(k\hat{\mathbf{z}})\cdot\mathbf{r}-\omega t]} + \hat{\mathbf{y}}E_0 e^{i[(k\hat{\mathbf{x}})\cdot\mathbf{r}-\omega t]}$$

according to (7.11) is uniform throughout space. Of course (7.11) does not apply since the $\mathbf{k}$-vectors are not parallel. Show that the time-average of $\mathbf{S}\left(\mathbf{r},t\right)$ according to (7.6) exhibits interference in the distribution of net energy flow.

7.3 Group vs. Phase Velocity: Sum of Two Plane Waves

P7.3 Show that (7.12) can be written as

$$\mathbf{E}(\mathbf{r},t) = 2\mathbf{E}_0 e^{i\left(\frac{\mathbf{k}_2+\mathbf{k}_1}{2}\cdot\mathbf{r}-\frac{\omega_2+\omega_1}{2}t\right)}\cos\left(\frac{\Delta\mathbf{k}}{2}\cdot\mathbf{r}-\frac{\Delta\omega}{2}t\right)$$

From this show that the speed at which the rapid-oscillation peaks move in Fig. 7.1 is

$$\frac{(v_{p1}+v_{p2})}{2}$$

P7.4 Confirm the right-hand side of (7.19).

7.4 Frequency Spectrum of Light

P7.5 The continuous field of a very narrowband continuous laser may be approximated as a pure plane wave: $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{i(k_0 z - \omega_0 t)}$. Suppose the wave encounters a shutter at the plane $z = 0$.

(a) Compute the power spectrum of the light before the shutter. HINT: The answer is proportional to the square of a delta function centered on ω_0 (see (0.43)).

(b) Compute the power spectrum after the shutter if it is opened during the interval $-\tau/2 \le t \le \tau/2$. Plot the result. Are you surprised that the shutter appears to create extra frequency components?

HINT: Write your answer in terms of the sinc function defined by $\text{sinc}\alpha \equiv \sin\alpha/\alpha$.

P7.6 (a) Determine the Full-Width-at-Half-Maximum of the intensity (i.e. the width of $I(\mathbf{r},t)$ represented by Δt_{FWHM}) and of the power spectrum (i.e. the width of $I(\mathbf{r},\omega)$ represented by $\Delta\omega_{\text{FWHM}}$) for the Gaussian pulse defined in (7.26).

HINT: Both answers are in terms of τ.

(b) Give an uncertainty principle for the product of Δt_{FWHM} and $\Delta\omega_{\text{FWHM}}$.

P7.7 Verify (7.27) for the Gaussian pulse defined by (7.24) and (7.26).

7.6 Quadratic Dispersion

P7.8 Suppose that the intensity of a Gaussian laser pulse has duration $\Delta t_{\text{FWHM}} = 25$ fs with carrier frequency ω_0 corresponding to $\lambda_{\text{vac}} = 800$ nm. The pulse goes through a lens of thickness $\ell = 1$ cm (laser quality glass type BK7) with index of refraction given approximately by

$$n(\omega) \cong 1.4948 + 0.016\frac{\omega}{\omega_0}$$

What is the full-width-at-half-maximum of the intensity for the emerging pulse?

HINT: For the input pulse we have

$$\tau = \frac{\Delta t_{\text{FWHM}}}{2\sqrt{\ln 2}}$$

(see P 7.6).

P7.9 If the pulse defined in (7.49) travels through the material for a very long distance z such that $\mathrm{T}(z) \to \tau\Phi(z)$ and $\tan^{-1}\Phi(z) \to \pi/2$, show that the instantaneous frequency of the pulse is

$$\omega_0 + \frac{t - 2z/v_g}{4\alpha z}$$

COMMENT: As the wave travels, the earlier part of the pulse oscillates more slowly than the later part. This is called chirp, and it means that the red frequencies get ahead of the blue ones since they experience a lower index.

7.7 Generalized Context for Group Delay

P7.10 When the spectrum is narrow compared to features in a resonance (such as in Fig. 7.11), the reshaping delay (7.58) tends to zero and can be ignored. Show that when the spectrum is narrow the net group delay (7.56) reduces to

$$\lim_{\tau\to\infty} \Delta t_G (\mathbf{r}) = \left. \frac{\partial \mathrm{Re}\mathbf{k}}{\partial \omega} \cdot \Delta\mathbf{r} \right|_{\bar{\omega}}$$

P7.11 When the spectrum is very broad the reshaping delay (7.58) also tends to zero and can be ignored. Show that when the spectrum is extremely broad, the net group delay reduces to

$$\lim_{\tau\to 0} \Delta t_G (\mathbf{r}) = \frac{\Delta r}{c}$$

assuming **k** and $\Delta\mathbf{r}$ are parallel. This means that a sharply defined signal cannot travel faster than c.

HINT: The real index of refraction n goes to unity far from resonance, and the imaginary part κ goes to zero.

P7.12 Work through the derivation of (7.55).

HINT: This somewhat lengthy derivation can be found in Optics Express **9**, 506-518 (2001).

Chapter 8

Coherence Theory

8.1 Introduction

Most students of physics become familiar with a Michelson interferometer (shown in Fig. 8.1) early in their course work. This preliminary understanding is usually gained in terms of a single-frequency plane wave that travels through the instrument. A Michelson interferometer divides the initial beam into two identical beams and then delays one beam with respect to the other before bringing them back together. Depending on the relative path difference d (roundtrip by our convention) between the two arms of the system, the light can interfere constructively or destructively in the direction of the detector. One way to view the relative path difference is in terms of the relative time delay $\tau \equiv d/c$. The intensity seen at the detector as a function of path difference is computed to be

$$\begin{aligned}
I_{\text{det}}(\tau) &= \frac{c\epsilon_0}{2}\left[\mathbf{E}_0 e^{i(kz-\omega t)} + \mathbf{E}_0 e^{i(kz-\omega(t-\tau))}\right] \cdot \left[\mathbf{E}_0 e^{i(kz-\omega t)} + \mathbf{E}_0 e^{i(kz-\omega(t-\tau))}\right]^* \\
&= \frac{c\epsilon_0}{2}\left[2\mathbf{E}_0 \cdot \mathbf{E}_0^* + 2\mathbf{E}_0 \cdot \mathbf{E}_0^* \cos(\omega\tau)\right] \\
&= 2I_0\left[1+\cos(\omega\tau)\right]
\end{aligned} \tag{8.1}$$

where $I_0 \equiv \frac{c\epsilon_0}{2}\mathbf{E}_0 \cdot \mathbf{E}_0^*$ is the intensity from one beam alone (when the other arm of the interferometer is blocked). This formula is familiar and it describes how the intensity at the detector oscillates between zero and four times the intensity of one beam alone. Notice that the intensity of one beam alone will be one fourth of the intensity originating from the source since it meets the beam splitter twice (assuming a 50:50 beam splitter).

In this chapter, we consider what happens when light containing a continuous band of frequencies is sent through the interferometer. In section 8.2, we derive an appropriate replacement for (8.1), which describes the intensity arriving at the detector when broadband light is sent through the interferometer. We will find that oscillations in the intensity at the detector become less pronounced as the mirror in one arm of the interferometer is scanned away from the position where the two paths are equal. Remarkably, this decrease in *fringe visibility* depends only upon the frequency content of the light without regard to whether the frequency components are organized into a short pulse or left as a longer pattern in time. In section 8.3, the concept of temporal coherence is explained in the context of what is observed in a Michelson interferometer. Section 8.4 gives an interpretation of the results

Albert Abraham Michelson

(1852–1931, United States)

Michelson (pronounced “Michael sun”) was born in Poland, but he grew up in the rough mining towns of California. He joined the navy, and later returned to teach at the naval academy. Michelson was fascinated by the problem of determining the speed of light, and developed several experiments to measure it more carefully. He is probably most famous for his experiment conducted with Edward Morley to detect the motion of the earth through the ether. He won the Nobel prize in 1907 for his contributions to optics.

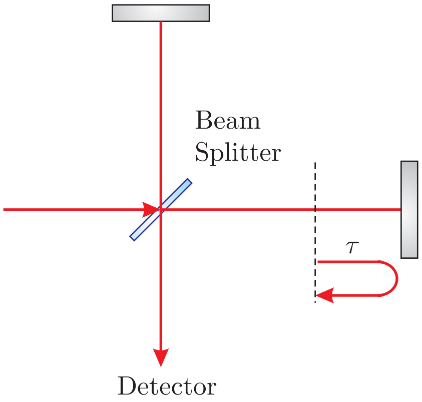


Figure 8.1 Michelson interferometer.

in terms of the fringe visibility and the *coherence length.*

In section 8.5, we discuss a practical application known as Fourier spectroscopy. This powerful technique makes it possible to deduce the spectral content of light using a Michelson interferometer. In section 8.6, we examine a Young’s two-slit setup and show how it is similar to a Michelson interferometer. Finally, the concept of spatial coherence is introduced in section 8.A in the context of a Young’s two-slit setup.

8.2 Michelson Interferometer

Consider a waveform $\mathbf{E}(t)$ that has traveled through the first arm of a Michelson interferometer to arrive at the detector in Fig. 8.1. Specifically, $\mathbf{E}(t)$ is the value of the field at the detector when the second arm of the interferometer is blocked. The waveform $\mathbf{E}(t)$ in general may be composed of many frequency components according to the inverse Fourier transform (7.20). For convenience we will think of $\mathbf{E}\,(t)$ as a pulse containing a finite amount of energy. (We will comment on continuous light sources in the next section.) The beam that travels through the second arm of the interferometer is associated with the same waveform, albeit with a delay τ according to the path difference between the two arms. Thus, $\mathbf{E}\left(t-\tau\right)$ indicates the field at the detector from the second arm when the first arm of the interferometer is blocked. Again, τ represents the *round-trip delay* of the adjustable path relative to the position where the two paths have equal lengths.

The total field at the detector is composed of the two waveforms:

$$\mathbf{E}_{\text{det}}(t,\tau)=\mathbf{E}\left(t\right)+\mathbf{E}\left(t-\tau\right) \tag{8.2}$$

With (7.28) we compute the intensity at the detector:

$$\begin{aligned}
I_{\text{det}}\left(t,\tau\right) &= \frac{c\epsilon_0}{2}\mathbf{E}_{\text{det}}(t,\tau)\cdot\mathbf{E}^*_{\text{det}}(t,\tau)\\
&= \frac{c\epsilon_0}{2}\left[\mathbf{E}(t)\cdot\mathbf{E}^*(t)+\mathbf{E}(t)\cdot\mathbf{E}^*(t-\tau)+\mathbf{E}(t-\tau)\cdot\mathbf{E}^*(t)+\mathbf{E}(t-\tau)\cdot\mathbf{E}^*(t-\tau)\right]\\
&= I(t)+I(t-\tau)+\frac{c\epsilon_0}{2}\left[\mathbf{E}(t)\cdot\mathbf{E}^*(t-\tau)+\mathbf{E}(t-\tau)\cdot\mathbf{E}^*(t)\right]\\
&= I(t)+I(t-\tau)+c\epsilon_0\text{Re}\left\{\mathbf{E}(t)\cdot\mathbf{E}^*(t-\tau)\right\}
\end{aligned} \tag{8.3}$$

The function $I(t)$ stands for the intensity of one of the beams arriving at the detector while the opposite path of the interferometer is blocked. Notice that we have retained the dependence on t in $I_{\text{det}}\left(t,\tau\right)$ in addition to the dependence on the path delay τ. This allows us to accommodate pulses of light that have a time-varying envelope. The rapid oscillations of the light are automatically averaged away in $I(t)$, but not the slowly varying form of the pulse.

The total energy (per area) accumulated at the detector is found by integrating the intensity over time. In other words, we let the detector integrate the energy of the entire pulse before taking a reading. For short laser pulses (sub-nanosecond), the detector automatically integrates the entire energy (per area) of the pulse since the detector cannot keep up with the detailed temporal variations of the pulse envelope. The integration of (8.3) over time yields

$$\int\limits_{-\infty}^{\infty} I_{\text{det}}\left(t,\tau\right)dt=\int\limits_{-\infty}^{\infty} I(t)dt+\int\limits_{-\infty}^{\infty} I\left(t-\tau\right)dt+c\epsilon_0\text{Re}\int\limits_{-\infty}^{\infty}\mathbf{E}\left(t\right)\cdot\mathbf{E}^*\left(t-\tau\right)dt \tag{8.4}$$

The final integral remains unchanged if we take a Fourier transform followed by an inverse Fourier transform:

$$\int\limits_{-\infty}^{\infty}\mathbf{E}(t)\cdot\mathbf{E}^*\left(t-\tau\right)dt=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}d\omega e^{-i\omega\tau}\left[\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}d\tau e^{i\omega\tau}\int\limits_{-\infty}^{\infty}\mathbf{E}\left(t\right)\cdot\mathbf{E}^*\left(t-\tau\right)dt\right] \tag{8.5}$$

The reason for this procedure is so that we can take advantage of the autocorrelation theorem (see P 0.30). We can use this theorem to replace the expression in brackets in (8.5):

$$\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}d\tau e^{i\omega\tau}\int\limits_{-\infty}^{\infty}\mathbf{E}\left(t\right)\cdot\mathbf{E}^*\left(t-\tau\right)dt=\sqrt{2\pi}\mathbf{E}\left(\omega\right)\cdot\mathbf{E}^*\left(\omega\right)=\sqrt{2\pi}\frac{2I\left(\omega\right)}{c\epsilon_0} \tag{8.6}$$

We can apply Parseval's theorem (see (7.27)) to the first two integrals on the right-hand side of (8.4):

$$\int_{-\infty}^{\infty} I(t)dt = \int_{-\infty}^{\infty} I\left(t-\tau\right)dt = \int_{-\infty}^{\infty} I\left(\omega\right)d\omega \tag{8.7}$$

Notice that the middle integral is insensitive to the delay τ since the integral is performed over all time (i.e. a change of variables $t' = t - \tau$ converts the middle integral into the first). With the aid of (8.6) and (8.7), the accumulated energy (8.4) at the detector becomes

$$\begin{aligned}\int_{-\infty}^{\infty} I_{\text{det}}\left(t,\tau\right)dt &= 2\int_{-\infty}^{\infty} I\left(\omega\right)d\omega + 2\text{Re}\int_{-\infty}^{\infty} I(\omega)e^{-i\omega\tau}d\omega \\ &= \left[2\int_{-\infty}^{\infty} I\left(\omega\right)d\omega\right]\left[1+\frac{\text{Re}\int_{-\infty}^{\infty} I\left(\omega\right)e^{-i\omega\tau}d\omega}{\int_{-\infty}^{\infty} I\left(\omega\right)d\omega}\right]\end{aligned} \tag{8.8}$$

It is convenient to rewrite this in terms of the *Degree of Coherence* function $\gamma\left(\tau\right)$:

$$\int_{-\infty}^{\infty} I_{\text{det}}\left(t,\tau\right)dt = \left[2\int_{-\infty}^{\infty} I(t)dt\right]\left[1+\text{Re}\gamma\left(\tau\right)\right] \tag{8.9}$$

where

$$\gamma\left(\tau\right) \equiv \frac{\int_{-\infty}^{\infty} I\left(\omega\right)e^{-i\omega\tau}d\omega}{\int_{-\infty}^{\infty} I\left(\omega\right)d\omega} \tag{8.10}$$

Notice that in writing (8.9) we have again applied Parseval's theorem (8.7) to part of the equation. In summary, (8.9) describes the accumulated energy (per area) arriving to the detector after the Michelson interferometer. The dependence on the path delay τ is entirely contained in the function $\gamma\left(\tau\right)$.

8.3 Temporal Coherence

We could have derived (8.9) using another strategy, which may seem more intuitive than the approach in the previous section. Equation (8.1) gives the intensity at the detector when a single plane wave of frequency ω goes through the interferometer. Now suppose that a waveform composed of many frequencies is sent through the interferometer. The intensity associated with each frequency acts independently, obeying (8.1) individually.

The total energy (per area) accumulated at the detector is then a linear superposition of the spectral intensities of all frequencies present:

$$\int_{-\infty}^{\infty} I_{\text{det}}\left(\omega,\tau\right)d\omega = \int_{-\infty}^{\infty} 2I\left(\omega\right)\left[1+\cos\left(\omega\tau\right)\right]d\omega \tag{8.11}$$

While this procedure may seem obvious, the fact that we can do it is remarkable! Remember that it is usually the fields that we must add together before finding the intensity of the resulting superposition. The formula (8.11) with its superposition of intensities relies on the fact that the different frequencies inside the interferometer when *time-averaged* (over all time) do not interfere. Certainly, the fields at different frequencies do interfere (or beat in time). However, they constructively interfere as often as they destructively interfere, and over time it is as though the individual frequency components transmit independently. Again, in writing (8.11) we considered the light to be pulsed rather than continuous so that the integrals converge.

We can manipulate (8.11) as follows:

$$\int\limits_{-\infty}^{\infty} I_{\text{det}}(\omega,\tau)\,d\omega = \left[2\int\limits_{-\infty}^{\infty} I(\omega)\,d\omega\right]\left[1+\frac{\int\limits_{-\infty}^{\infty} I(\omega)\cos(\omega\tau)\,d\omega}{\int\limits_{-\infty}^{\infty} I(\omega)\,d\omega}\right] \tag{8.12}$$

This is the same as (8.8) since we can replace $\cos(\omega\tau)$ with $\text{Re}\left\{e^{-i\omega\tau}\right\}$, and we can apply Parseval's theorem (8.7) to the other integrals. Thus, the above arguments lead to (8.9) and (8.10), in complete agreement with the previous section.

Finally, let us consider the case of a continuous light source for which the integrals in (8.9) diverge. This is the case for starlight or for a *continuous wave* (CW) laser source. The integral $\int_{-\infty}^{\infty} I(t)dt$ diverges since a source that is on forever (or at least for a very long time) emits infinite (or very much) energy. However, note that the integrals on both sides of (8.9) diverge in the same way. We can renormalize (8.9) in this case by replacing the integrals on each side with the average value of the intensity:

$$I_{\text{ave}} \equiv \langle I(t)\rangle_t = \frac{1}{T}\int\limits_{-T/2}^{T/2} I(t)dt \quad \text{(continuous source)} \tag{8.13}$$

The duration T must be large enough to average over any fluctuations that are present in the light source. The average in (8.13) should not be used on a pulsed light source since the result would depend on the duration T of the temporal window.

In the continuous wave (CW) case (e.g. starlight or a CW laser), the signal at the detector (8.9) becomes

$$\langle I_{\text{det}}(t,\tau)\rangle_t = 2\langle I(t)\rangle_t\left[1+\text{Re}\gamma(\tau)\right] \quad \text{(continuous source)} \tag{8.14}$$

Although technically the integrals involved in computing $\gamma(\tau)$ (8.10) also diverge in the case of CW light, the numerator and the denominator diverge in the same way. Therefore, we may renormalize $I(\omega)$ in any way we like to deal with this problem, and this does not affect the final result. Regardless of how large $I(\omega)$ is, and regardless of the units on the measurement (volts or whatever), we can simply plug the instrument reading directly into (8.10). The units in the numerator and denominator cancel so that $\gamma(\tau)$ always remains dimensionless.

A very remarkable aspect of the above result is that the behavior of the light in the Michelson interferometer does not depend on the phase of $\mathbf{E}(\omega)$. It depends only on the

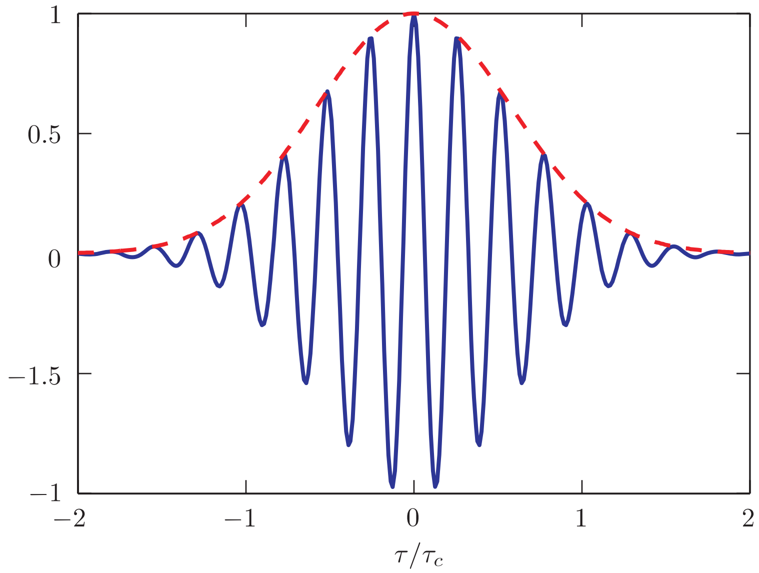


Figure 8.2 Re[$\gamma(\tau)$] (solid) and $|\gamma(\tau)|$ (dashed) for a light pulse having a Gaussian spectrum (7.26).

amount of light associated with each frequency component through $I(\omega) \equiv \frac{\epsilon_0 c}{2}\mathbf{E}(\omega)\cdot\mathbf{E}^*(\omega)$. When the light at one frequency undergoes constructive interference for a given path difference τ, the light at another frequency might undergo destructive interference. The net effect is given in the degree of coherence function $\gamma(\tau)$, which contains the essential information describing interference. Fig. 8.2 depicts the degree of coherence function as one arm of the interferometer is adjusted through various delays τ. In summary, narrowband light is temporally more coherent than broadband light because there is less "interference" between different frequencies.

8.4 Fringe Visibility and Coherence Length

The degree of coherence function $\gamma(\tau)$ is responsible for oscillations in intensity at the detector as the mirror in one of the arms is moved. The real part Re$\gamma(\tau)$ is analogous to cos($\omega\tau$) in (8.1). For large delays τ, the oscillations tend to die off as different frequencies individually interfere, some constructively, others destructively. For large path differences, the intensity at the detector tends to remain steady as the mirror is moved further. We define the coherence time to be the amount of delay necessary to cause $\gamma(\tau)$ to quit oscillating (i.e. its amplitude approaches zero). A useful (although arbitrary) definition for the coherence time is

$$\tau_c \equiv \int_{-\infty}^{\infty} |\gamma(\tau)|^2 \, d\tau = 2\int_{0}^{\infty} |\gamma(\tau)|^2 \, d\tau \tag{8.15}$$

The coherence length is the distance that light travels in this time:

$$\ell_c \equiv c\tau_c \tag{8.16}$$

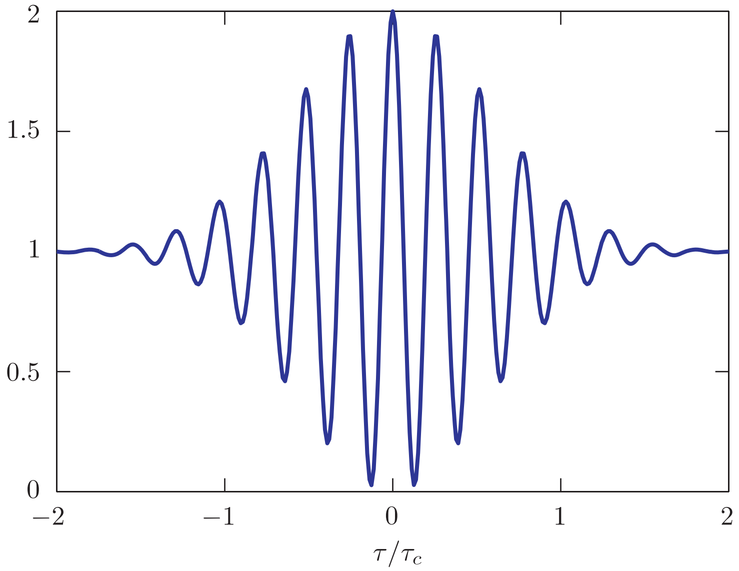


Figure 8.3 The output of a Michelson interferometer for a Gaussian spectrum (8.21)

Another useful concept is *fringe visibility.* The fringe visibility is defined in the following way:

$$V(\tau) \equiv \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \qquad \text{(continuous)} \tag{8.17}$$

or

$$V(\tau) \equiv \frac{\mathcal{E}_{\max} - \mathcal{E}_{\min}}{\mathcal{E}_{\max} + \mathcal{E}_{\min}} \qquad \text{(pulsed)} \tag{8.18}$$

where $\mathcal{E}_{\max} \equiv \max \int_{-\infty}^{\infty} I_{\text{det}}(t,\tau)\, dt$ refers to the accumulated energy (per area) at the detector when the mirror is positioned such that the amount of throughput to the detector is a local maximum (i.e. the left-hand side of (8.9)). $\mathcal{E}_{\min}$ refers to the accumulated energy at the detector when the mirror is positioned such that the amount of throughput to the detector is a local minimum. As the mirror moves a large distance from the equal-path-length position, the oscillations become less pronounced because the values of $\mathcal{E}_{\min}$ and $\mathcal{E}_{\max}$ tend to take on the same value, and the fringe visibility goes to zero. The fringe visibility goes to zero when $\gamma(\tau)$ goes to zero. It is left as an exercise to show that the fringe visibility can be written as

$$V(\tau) = |\gamma(\tau)| \tag{8.19}$$

In the case of a Gaussian spectral distribution (7.26)

$$I(\omega) = I(\omega_0)\, e^{-\left(\frac{\omega-\omega_0}{\Delta\omega}\right)^2} \tag{8.20}$$

the result of (8.10) is

$$\gamma(\tau) = e^{-i\tau\omega_0 - \frac{(\Delta\omega)^2\tau^2}{4}} \tag{8.21}$$

Figure 8.2 plots the magnitude and real part of (8.21). From (8.15) the coherence time is

$$\tau_{\mathrm{c}} = \frac{\sqrt{2\pi}}{\Delta\omega} \tag{8.22}$$

Figure 8.3 shows $1 + \mathrm{Re}\gamma\left(\tau\right)$, which is proportional to the energy (per area) arriving at the detector. As expected, the fringes die off for a delay interval of τ_{c}.

8.5 Fourier Spectroscopy

As we have seen in the previous discussion, the signal output from a Michelson interferometer for a pulsed input is given by

$$\mathrm{Sig}\left(\tau\right) \propto \int\limits_{-\infty}^{\infty} I_{\mathrm{det}}\left(t,\tau\right) dt = \left[2\int\limits_{-\infty}^{\infty} I\left(t\right) dt\right]\left[1 + \mathrm{Re}\gamma\left(\tau\right)\right] \tag{8.23}$$

where

$$\gamma(\tau) \equiv \frac{\int\limits_{-\infty}^{\infty} I(\omega)e^{-i\omega\tau}d\omega}{\int\limits_{-\infty}^{\infty} I(\omega)d\omega} \tag{8.24}$$

Typically, the signal comes in the form of a voltage or a current from a sensor. However, the signal can be normalized to the signal level occurring when τ is large (i.e. fringe visibility goes to zero: $\gamma\left(\tau\right) = 0$). In this case, the normalized signal must approach

$$\lim_{\tau\to\infty} \eta\mathrm{Sig}\left(\tau\right) = 2\mathcal{E}_0 \tag{8.25}$$

where η is the appropriate normalization constant that changes the proportionality (8.23) into an equation, and

$$\mathcal{E}_0 \equiv \int_{-\infty}^{\infty} I(t)dt = \int_{-\infty}^{\infty} I(\omega)d\omega \tag{8.26}$$

denotes the total energy (per area) that would arrive at the detector from one arm of the interferometer (i.e. if the other arm were blocked).

Given our measurement of $\mathrm{Sig}(\tau)$, we would like to find $I(\omega)$, or the spectrum of the light. Unfortunately, $I(\omega)$ is buried within the integrals (8.23). However, since the denominator of $\gamma(\tau)$ is constant (equal to $\mathcal{E}_0$) and since the numerator of $\gamma(\tau)$ looks like an inverse Fourier transform of $I(\omega)$, we are able to extract the desired spectrum after some manipulation. This procedure for extracting $I(\omega)$ from an interferometric measurement is known as *Fourier spectroscopy.*

We now describe the procedure for obtaining $I(\omega)$. We can write the properly normalized signal (8.23) as

$$\eta\mathrm{Sig}\left(\tau\right) = 2\mathcal{E}_0 + 2\mathrm{Re}\int\limits_{-\infty}^{\infty} I(\omega)e^{-i\omega\tau}d\omega \tag{8.27}$$

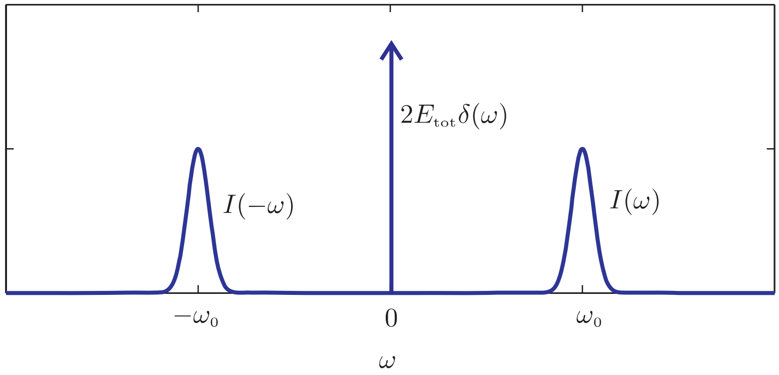


Figure 8.4 Depiction of $F\{\text{Sig}(\tau)\}/\sqrt{2\pi}$.

Next, we take the Fourier transform of this equation:

$$\mathcal{F}\left\{\eta \text{Sig}\left(\tau\right)\right\} = \mathcal{F}\left\{2\mathcal{E}_0\right\} + \mathcal{F}\left\{2\text{Re}\int\limits_{-\infty}^{\infty} I\left(\omega\right) e^{-i\omega\tau} d\omega\right\} \tag{8.28}$$

The left-hand side is known since it is the measured data, and a computer can be employed to take the Fourier transform of it. The first term on the right-hand side is the Fourier transform of a constant:

$$\mathcal{F}\left\{2\mathcal{E}_0\right\} = 2\mathcal{E}_0 \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{i\omega\tau} d\tau = 2\mathcal{E}_0 \sqrt{2\pi}\delta\left(\omega\right) \tag{8.29}$$

Notice that (8.29) is zero everywhere except where $\omega = 0$, where a spike occurs. This represents the DC component of $\mathcal{F}\left\{\eta \text{Sig}\left(\tau\right)\right\}$.

The second term of (8.28) can be written as

$$\begin{aligned}
\mathcal{F}\left\{2\text{Re}\int\limits_{-\infty}^{\infty} I\left(\omega\right) e^{-i\omega\tau} d\omega\right\} &= \mathcal{F}\left\{\int\limits_{-\infty}^{\infty} I\left(\omega\right) e^{-i\omega\tau} d\omega + \int\limits_{-\infty}^{\infty} I\left(\omega\right) e^{i\omega\tau} d\omega\right\} \\
&= \int\limits_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty} I(\omega')e^{-i\omega'\tau}d\omega'\right) e^{i\omega\tau}d\tau + \int\limits_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty} I(\omega')e^{i\omega'\tau}d\omega'\right) e^{i\omega\tau}d\tau \\
&= \sqrt{2\pi}\left[\int\limits_{-\infty}^{\infty} I(\omega')\left(\frac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{-i(\omega'-\omega)\tau}d\tau\right)d\omega' + \int\limits_{-\infty}^{\infty} I(\omega')\left(\frac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{-i(\omega'+\omega)\tau}d\tau\right)d\omega'\right] \\
&= \sqrt{2\pi}\left[\int\limits_{-\infty}^{\infty} I(\omega')\delta\left(\omega'-\omega\right)d\omega' + \int\limits_{-\infty}^{\infty} I(\omega')\delta\left(\omega'+\omega\right)d\omega'\right] \\
&= \sqrt{2\pi}\left[I\left(\omega\right) + I\left(-\omega\right)\right]
\end{aligned} \tag{8.30}$$

With (8.29) and (8.30) we can write (8.28) as

$$\frac{\mathcal{F}\left\{\eta \text{Sig}\left(\tau\right)\right\}}{\sqrt{2\pi}} = 2\mathcal{E}_0\delta\left(\omega\right) + I\left(\omega\right) + I\left(-\omega\right) \tag{8.31}$$

Thomas Young

(1773–1829, English)

Young was a physician by trade, but studied widely in other fields. His double slit experiment gave convincing evidence of the wave nature of light. He also did extensive research into color vision. On the side, he translated hieroglyphics and studied many other languages.

The Fourier transform of the measured signal is seen to contain three terms, one of which is the power spectrum that we are after, namely $I(\omega)$. Fortunately, when graphed as a function of ω (shown in Fig. 8.4), the three terms on the right-hand side typically do not overlap. As a reminder, the measured signal as a function of τ looks something like that in Fig. 8.3. The oscillation frequency of the fringes lies in the neighborhood of ω_0. To obtain $I(\omega)$ the procedure is clear: Record $\mathrm{Sig}(\tau)$; if desired, normalize by its value at large τ; take its Fourier transform; extract the curve at positive frequencies.

8.6 Young's Two-Slit Setup and Spatial Coherence

In close analogy with the Michelson interferometer, which is able to investigate temporal coherence, the Young's two-slit experiment can be used to investigate *spatial coherence* of quasi-monochromatic light. Thomas Young, who lived nearly a century before Michelson, used his two-slit setup for the first conclusive demonstration that light is a wave. The Young's two-slit setup and the Michelson interferometer have in common that two beams of light travel different paths and then interfere. In the Michelson interferometer, one path is delayed with respect to the other so that temporal effects can be studied. In the Young's two-slit setup, two laterally separate points of the same wave are compared as they are sent through two slits. Depending on the coherence of the wave at the two points, the fringe pattern observed can exhibit good or poor visibility.

Just as the Michelson interferometer is sensitive to the *spectral content* of light, the Young's two-slit setup is sensitive to the *spatial extent* of the light source illuminating the two slits. For example, if light from a distant star (restricted by a filter to a narrow spectral range) is used to illuminate a double-slit setup, the resulting interference pattern appearing on a subsequent screen contains information regarding the angular width of the star. Michelson was the first to use this type of setup to measure the angular width of stars.

Light emerging from a single ideal point source has wave fronts that are spatially uniform in a lateral sense (see Fig. 8.5). Such wave fronts are said to be *spatially coherent*, even

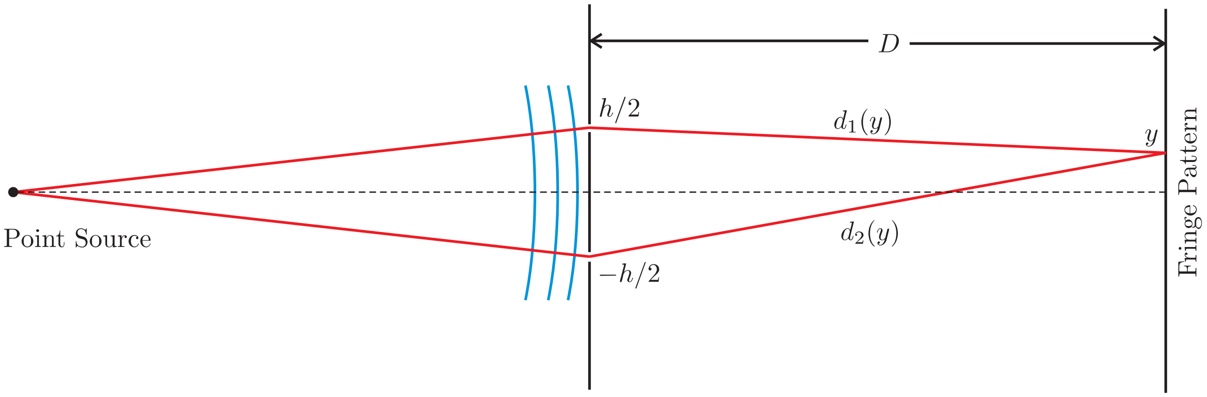


Figure 8.5 A point source produces coherent (locked phases) light. When this light which traverses two slits and arrives at a screen it produces a fringe pattern.

if the temporal coherence is not perfect (i.e. if a range of frequencies is present). When spatially coherent light illuminates a Young's two-slit setup, fringes of maximum visibility are seen at a distant screen, meaning the fringes vary between a maximum intensity and zero. If a larger source of light (with randomly varying phase across its extent) is used to illuminate the Young's two-slit setup (see Fig. 8.6), the wave fronts at the two slits are less correlated, and the visibility of the fringes on the distant screen diminishes because fringes fluctuate rapidly in time and partially "wash out."

We now consider the details of the Young's two-slit setup. When both slits of a Young's two-slit setup are illuminated with spatially coherent light, the resulting pattern on a far-away screen is given by

$$I = 2I_0 \left[1 + \cos\left[k\left(d_2 - d_1\right) + \phi_2 - \phi_1\right]\right] = 2I_0 \left[1 + \cos\left(khy/D + \Delta\phi\right)\right] \tag{8.32}$$

where ϕ_1 and ϕ_2 are the phases of the wave front at the two slits, respectively. Notice the close similarity with a Michelson interferometer (see (8.1)). Here the controlling variable is h (the separation of the slits) rather than τ (the delay introduced by moving a mirror in the Michelson interferometer). To obtain the final expression in (8.32) we have made the approximations

$$d_1\left(y\right) = \sqrt{\left(y - h/2\right)^2 + D^2} = D\sqrt{1 + \frac{\left(y - h/2\right)^2}{D^2}} \cong D\left[1 + \frac{\left(y - h/2\right)^2}{2D^2} + \cdots\right] \tag{8.33}$$

and

$$d_2\left(y\right) = \sqrt{\left(y + h/2\right)^2 + D^2} = D\sqrt{1 + \frac{\left(y + h/2\right)^2}{D^2}} \cong D\left[1 + \frac{\left(y + h/2\right)^2}{2D^2} + \cdots\right] \tag{8.34}$$

These approximations are valid as long as $D \gg y$ and $D \gg h$.

We now consider how to modify (8.32) so that it applies to the case when the two slits are illuminated by a host of point sources distributed over a finite lateral extent. This situation is depicted in Fig. 8.6 and it leads to partial *spatial coherence* when the phase of

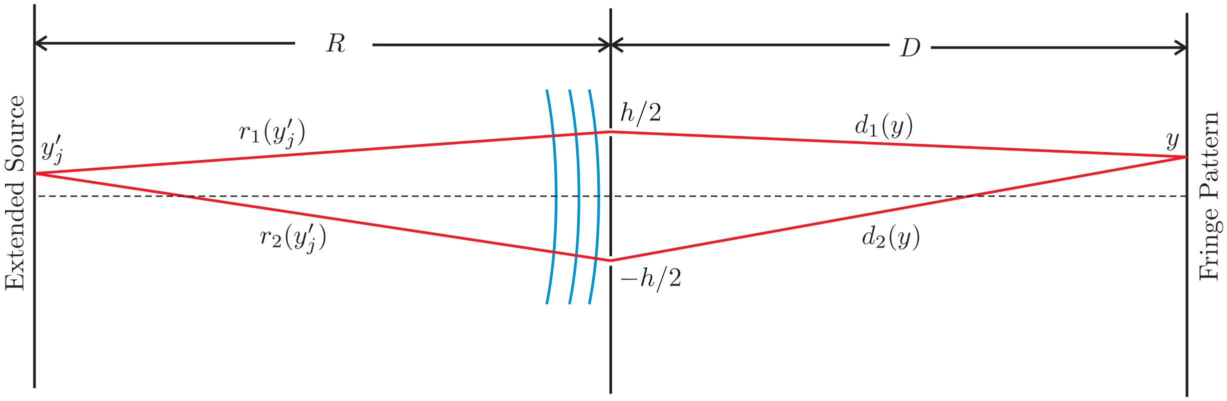


Figure 8.6 Light from an extended source is only partially coherent. Fringes are still possible, but they exhibit less contrast.

each emitter is random. Again, *spatial coherence* is a term used to describe whether the phase of the wave fronts at one slit are correlated with the phase of the wave fronts at the other slit. We will find that a larger source gives less coherent wave fronts at the slits.

To simplify our analysis, let us consider the many point sources to be arranged in one dimension (in the plane of the figure). We restrict the distribution of point sources to vary only in the y' dimension. This ensures that the light has uniform phase along either slit (in and out of the plane of Fig. 8.6). We assume that the light is *quasi-monochromatic* so that its frequency is approximately ω with a phase that fluctuates randomly over time intervals much longer than the period of oscillation $2\pi/\omega$. This necessarily implies that there will be some frequency bandwidth, however small.

The light emerging from the j^{th} point at y'_j travels by means of two very narrow slits to a point y on a screen. Let $E_1(y'_j)$ and $E_2(y'_j)$ be the fields on the screen at y, each originating from the point y'_j and traveling respectively through the two slits. We suppress the vectorial nature of $E_1(y'_j)$ and $E_2(y'_j)$, and we ignore possible complications due to field polarization. The total field contribution at the screen from the j^{th} point is obtained by adding $E_1(y'_j)$ and $E_2(y'_j)$. Let us make the assumption that $E_1(y'_j)$ and $E_2(y'_j)$ have the same amplitude $|E(y'_j)|$. Thus, the two fields differ only in their phases according to the respective distances traveled to the screen. This allows us to write the two fields as

$$E_1(y'_j) = \left|E(y'_j)\right| e^{i\left\{k\left[r_1(y'_j)+d_1(y)\right]-\omega t+\phi(y'_j)\right\}} \tag{8.35}$$

and

$$E_2(y'_j) = \left|E(y'_j)\right| e^{i\left\{k\left[r_2(y'_j)+d_2(y)\right]-\omega t+\phi(y'_j)\right\}} \tag{8.36}$$

Notice that we have explicitly included an arbitrary phase $\phi(y'_j)$, which is different for each point source.

We now set about finding the cumulative field at y arising from the many points indexed by the subscript j. We therefore sum over the index j. Again, for simplicity we have assumed that the point sources are distributed along one dimension, in the y'-direction. The upcoming results can be generalized to a two-dimensional source where the point sources are distributed also in and out of the plane of Fig. 8.6. However, in this case, the slits should be replaced with two pinholes.

The net field on the screen at point y is

$$E_{\text{net}}(h) = \sum_j \left[E_1(y_j') + E_2(y_j')\right] \tag{8.37}$$

This net field depends not only on h, but also on y, R, D, and k as well as on the phase $\phi(y_j')$ at each point. Nevertheless, in the end we will mainly emphasize the dependence on the slit separation h. The intensity of this field is

$$\begin{aligned} I_{\text{net}}(h) &= \frac{\epsilon_0 c}{2}\left|E_{\text{net}}(h)\right|^2 \\ &= \frac{\epsilon_0 c}{2}\left[\sum_j E_1(y_j') + E_2(y_j')\right]\left[\sum_m E_1(y_m') + E_2(y_m')\right]^* \\ &= \frac{\epsilon_0 c}{2}\sum_{j,m}\left[E_1(y_j')E_1^*(y_m') + E_2(y_j')E_2^*(y_m') + 2\text{Re}E_1(y_j')E_2^*(y_m')\right] \end{aligned} \tag{8.38}$$

When inserting the field expressions (8.35) and (8.36) into this expression for the intensity at the screen, we get

$$\begin{aligned} I_{\text{net}}(h) = \frac{\epsilon_0 c}{2}\sum_{j,m}\Big[&\left|E(y_j')\right|\left|E(y_m')\right| e^{ik\left[r_1(y_j')-r_1(y_m')\right]} e^{i\left[\phi(y_j')-\phi(y_m')\right]} \\ &+\left|E(y_j')\right|\left|E(y_m')\right| e^{ik\left[r_2(y_j')-r_2(y_m')\right]} e^{i\left[\phi(y_j')-\phi(y_m')\right]} \\ &+2\text{Re}\left|E(y_j')\right|\left|E(y_m')\right| e^{ik\left[r_1(y_j')-r_2(y_m')\right]} e^{ik\left[d_1(y)-d_2(y)\right]} e^{i\left[\phi(y_j')-\phi(y_m')\right]}\Big] \end{aligned} \tag{8.39}$$

At this juncture we make a critical assumption that the phase of the emission $\phi(y_j')$ varies in time independently at every point on the source. This assumption is appropriate for the emission from thermal sources such as starlight, a glowing filament (filtered to a narrow frequency range), or spontaneous emission from an excited gas or plasma. The assumption of random phase, however, is inappropriate for coherent sources such as laser light. We comment on this in Appendix 8.B.

A wonderful simplification happens to (8.39) when $\phi(y_j') - \phi(y_m')$ varies randomly in time for $j \neq m$ (i.e. when there is no correlation between the two phases). Keep in mind that to the extent that the phases vary in time, the frequency spectrum of the light broadens in competition with our quasi-monochromatic assumption. If we average the intensity over an extended time, then $e^{i\left[\phi(y_j')-\phi(y_m')\right]}$ averages to zero unless we have $j = m$ in which case the factor reduces to e^0 which is always one. Thus, we have

$$\left\langle e^{i\left[\phi(y_j')-\phi(y_m')\right]}\right\rangle_t = \delta_{j,m} \equiv \begin{cases} 1 \text{ if } j = m, \\ 0 \text{ if } j \neq m. \end{cases} \quad \text{(random phase assumption)} \tag{8.40}$$

The function $\delta_{j,m}$ is known as the Kronecker delta function.

The time-averaged intensity under the random-phase assumption (8.40) becomes

$$\left\langle I_{\text{net}}(h)\right\rangle_t = \sum_j I(y_j') + \sum_j I(y_j') + 2\text{Re}\sum_j I(y_j')e^{ik\left[r_1(y_j')-r_2(y_j')\right]}e^{ik\left[d_1(y)-d_2(y)\right]} \tag{8.41}$$

We may use (8.33) to simplify $d_1(y) - d_2(y) \cong hy/D$, and similarly, we may simplify $r_1(y_j') - r_2(y_j') \cong y_j'h/R$ with the approximations

$$r_1(y_j') = \sqrt{\left(y_j' - h/2\right)^2 + R^2} \cong R\left[1 + \frac{\left(y_j' - h/2\right)^2}{2R^2} + \cdots\right] \tag{8.42}$$

and

$$r_2(y_j') = \sqrt{\left(y_j' + h/2\right)^2 + R^2} \cong R\left[1 + \frac{\left(y_j' + h/2\right)^2}{2R^2} + \cdots\right] \tag{8.43}$$

With these simplifications, (8.41) becomes

$$\left\langle I_{\text{net}}\left(h\right)\right\rangle_t = 2\sum_j I\left(y_j'\right) + 2\text{Re}e^{-i\frac{khy}{D}}\sum_j I\left(y_j'\right)e^{-i\frac{khy_j'}{R}} \quad \text{(random phase assumption)} \tag{8.44}$$

The only thing left to do is to put this formula into a slightly more familiar form:

$$\left\langle I_{\text{net}}\left(h\right)\right\rangle_t = \left[2\sum_j I\left(y_j'\right)\right]\left[1 + \text{Re}\gamma\left(h\right)\right] \quad \text{(random phase assumption)} \tag{8.45}$$

where

$$\gamma\left(h\right) \equiv \frac{e^{-i\frac{khy}{D}}\sum\limits_j I\left(y_j'\right)e^{-i\frac{khy_j'}{R}}}{\sum\limits_j I\left(y_j'\right)} \tag{8.46}$$

Students should notice the close similarity to the Michelson interferometer, (8.9) and (8.10). As before, $\gamma(h)$ is known as the *degree of coherence*, in this case spatial coherence. It controls the fringe pattern seen at the screen.

The factor $\exp\left(-ikhy/D\right)$ defines the positions of the periodic fringes on the screen. The remainder of (8.46) controls the depth of the fringes as the slit separation h is varied. When the slit separation h increases, the amplitude of $\gamma\left(h\right)$ tends to diminish until the intensity at the screen becomes uniform. When the two slits have very small separation (such that $e^{-i\frac{khy'}{R}} \cong 1$ wherever $I(y')$ is significant) then we have $\left|\gamma\left(h\right)\right| = 1$ and very good fringe visibility results. As the slit separation h increases, the fringe visibility

$$V\left(h\right) = \left|\gamma\left(h\right)\right| \tag{8.47}$$

diminishes, eventually approaching zero (see (8.19)). In analogy to the temporal case (see (8.15)), we can define a slit separation sufficiently large to make the fringes at the screen disappear:

$$h_{\text{c}} \equiv 2\int\limits_0^\infty \left|\gamma\left(h\right)\right|^2 dh \tag{8.48}$$

We can generalize (8.46) so that it applies to the case of a continuous distribution of light as opposed to a collection of discrete point sources. In Appendix 8.A we show how summations in (8.45) and (8.46) become integrals over the source intensity distribution, and we write

$$\langle I_{\text{net}}(h)\rangle_t = 2\langle I_{\text{oneslit}}\rangle_t \left[1 + \operatorname{Re}\gamma(h)\right] \quad \text{(random phase assumption)} \tag{8.49}$$

where

$$\gamma(h) \equiv \frac{e^{-i\frac{khy}{D}} \int\limits_{-\infty}^{\infty} I(y')e^{-i\frac{khy'}{R}}dy'}{\int\limits_{-\infty}^{\infty} I(y')dy'} \tag{8.50}$$

Note that $I(y')$ has units of intensity per length in this expression.

Appendix 8.A Spatial Coherence with a Continuous Source

In this appendix we examine the coherence of light from a continuous spatial distribution (as opposed to a collection of discrete point sources) and justify (8.50) and (8.47) under the assumption of randomly varying phase at the source. We begin by replacing the summations in (8.39) with integrals over a continuous emission source. As we do this, we must consider the field contributions to be in units of field per length of the extended source. We make the following replacements:

$$\begin{aligned}
\sum_j E_1(y'_j) &\to \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} E_1(y')dy' \\
\sum_m E_1(y'_m) &\to \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} E_1(y'')dy'' \\
\sum_j E_2(y'_j) &\to \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} E_2(y')dy' \\
\sum_m E_2(y'_m) &\to \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} E_2(y'')dy''
\end{aligned} \tag{8.51}$$

We include the factor $1/\sqrt{2\pi}$ here as part of the definition of the field distributions for later convenience.

With the above replacements, (8.39) becomes

$$\begin{aligned}I_{\text{net}}(h) = \frac{\epsilon_0 c}{2}\Bigg[&\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|E(y')\right| e^{ikr_1(y')}e^{i\phi(y')}dy'\int_{-\infty}^{\infty}\left|E(y'')\right| e^{-ikr_1(y'')}e^{-i\phi(y'')}dy''\\ &+\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|E(y')\right| e^{ikr_2(y')}e^{i\phi(y')}dy'\int_{-\infty}^{\infty}\left|E(y'')\right| e^{-ikr_2(y'')}e^{-i\phi(y'')}dy''\\ &+2\text{Re}\frac{e^{ik[d_1(y)-d_2(y)]}}{2\pi}\int_{-\infty}^{\infty}\left|E(y')\right| e^{ikr_1(y')}e^{i\phi(y')}dy'\int_{-\infty}^{\infty}\left|E(y'')\right| e^{-ikr_2(y'')}e^{-i\phi(y'')}dy''\Bigg]\end{aligned} \tag{8.52}$$

The next step is to make the average over random phases. Rather than deal with a time average of randomly varying phases, we will instead work with a linear superposition of all conceivable phase factors. That is, we will write the phase as $\phi(y'_j) \rightarrow Ky'$, where K is a parameter with units of inverse length, which we allow to take on all possible real values with uniform likelihood. The way we modify (8.40) for the continuous case is then

$$\left\langle e^{i\left[\phi(y'_j)-\phi(y'_m)\right]}\right\rangle_t = \delta_{j,m} \rightarrow \int_{-\infty}^{\infty} e^{iK(y'-y'')}dK = 2\pi\delta(y''-y') \tag{8.53}$$

Instead of taking the time average, we integrate both sides of (8.52) over all possible values of the phase parameter K, whereupon the delta function in (8.53) naturally arises on the right-hand side of the equation.

When (8.52) is integrated over K, the result is

$$\begin{aligned}\int_{-\infty}^{\infty} I_{\text{net}}(h)\,dK = \frac{\epsilon_0 c}{2}\Bigg[&\int_{-\infty}^{\infty}|E(y')|\,e^{ikr_1(y')}dy'\int_{-\infty}^{\infty}|E(y'')|\,e^{-ikr_1(y'')}\delta(y''-y')\,dy''\\ &+\int_{-\infty}^{\infty}|E(y')|\,e^{ikr_2(y')}dy'\int_{-\infty}^{\infty}|E(y'')|\,e^{-ikr_2(y'')}\delta(y''-y')\,dy''\\ &+2\text{Re}e^{ik[d_1(y)-d_2(y)]}\int_{-\infty}^{\infty}|E(y')|\,e^{ikr_1(y')}dy'\int_{-\infty}^{\infty}|E(y'')|\,e^{-ikr_2(y'')}\delta(y''-y')\,dy''\Bigg]\\ &\text{(random phase assumption)}\end{aligned} \tag{8.54}$$

It may seem strange at first that the left-hand side of (8.54) has units of intensity per unit length. This is somewhat abstract. However, these units result from the natural way of dealing with the random phases when the source is continuous. As K varies, the phase distribution at the source varies. The integral in (8.54) averages all of these possibilities.

The delta functions in (8.54) allow us to perform another stage of integration for each term on the right-hand side. We can also make substitutions from (8.33), (8.34), (8.42) and

(8.43). The result is

$$\int_{-\infty}^{\infty} I_{\text{net}}\left(h\right)dK = 2\int_{-\infty}^{\infty} I(y')dy' + 2\text{Re}e^{-i\frac{khy}{D}}\int_{-\infty}^{\infty} I(y')e^{-i\frac{khy'}{R}}dy' \text{ (random phase assumption)} \tag{8.55}$$

where

$$I(y') \equiv \frac{1}{2}\epsilon_0 c\left|E(y')\right|^2 \tag{8.56}$$

Notice that $I(y')$ in the present context has units of intensity per length squared since $E(y')$ has units of field per length. As they should, the units on the two sides of (8.55) match, both having units of intensity per length. (Recall that K has units of per length and $I_{\text{net}}\left(h\right)$ has usual units of intensity.) We can renormalize these strange units on each side of the equation. We can redefine the left-hand side $\int_{-\infty}^{\infty} I_{\text{net}}\left(h\right)dK$ to be the intensity at the screen and the integral on the right-hand side $\int_{-\infty}^{\infty} I(y')dy'$ to be the intensity at the screen when only one slit is open. Then (8.55) reduces to (8.49) and (8.50).

Appendix 8.B The van Cittert-Zernike Theorem

In this appendix we avoid making the assumption of randomly varying phase. This would be the case when the source of light is, for example, a laser. By substituting (8.35) and (8.36) into (8.52) we have

$$\begin{aligned} I_{\text{net}}(h) = \frac{\epsilon_0 c}{2\sqrt{2\pi}} \Bigg[& \left|\int_{-\infty}^{\infty}\left[\left|E(y')\right|e^{i\phi(y')+i\frac{ky'^2}{2R}}\right]e^{-i\frac{khy'}{2R}}dy'\right|^2 + \left|\int_{-\infty}^{\infty}\left[\left|E(y')\right|e^{i\phi(y')+i\frac{ky'^2}{2R}}\right]e^{i\frac{khy'}{2R}}dy'\right|^2 \\ & + 2\text{Re}\frac{e^{i\frac{khy}{D}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\left|E\left(y'\right)\right|e^{i\phi(y')+i\frac{ky'^2}{2R}}\right]e^{-i\frac{khy'}{2R}}dy'\left\{\int_{-\infty}^{\infty}\left[\left|E\left(y''\right)\right|e^{i\phi(y'')+i\frac{ky''^2}{2R}}\right]e^{i\frac{khy''}{2R}}dy''\right\}^*\Bigg] \end{aligned} \tag{8.57}$$

The three terms on the right-hand side of (8.57) can be understood as follows. The first term is the intensity on the screen when the lower slit is covered. The second term is the intensity on the screen when the upper slit is covered. The last term is the interference term, which modifies the sum of the individual intensities when both slits are uncovered.

Notice the occurrence of Fourier transforms (over position) on the quantities inside of the square brackets. Later, when we study diffraction theory, we will recognize these transforms. The Fourier transforms here determine the strength of fields impinging on the individual slits. We have essentially worked out diffraction theory for this specific case. The appearance of the strength of the field illuminating each of the slits explains the major difference between the coherent source and the random-phase source. With the random-phase source, the slits are always illuminated with the same strength regardless of the separation. However, with a coherent source, "beaming" can occur such that the strength (and phase) of the field at each slit depends on its exact position.

A wonderful simplification occurs when the phase of the emitted light has the following distribution:

$$\phi(y') = -\frac{ky'^2}{2R} \quad \text{(converging spherical wave)} \tag{8.58}$$

Equation (8.58) is not as arbitrary as it may first appear. The particular phase is an approximation to a concave spherical wave front converging to the center between the two slits. This type of wave front is created when a plane wave passes through a lens. With the special phase (8.58), the intensity (8.57) reduces to

$$\begin{aligned} I_{\text{net}}(h) = \frac{\epsilon_0 c}{2}\Bigg[& \left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|E(y')\right| e^{-i\frac{khy'}{2R}}dy'\right|^2 + \left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|E\left(y'\right)\right| e^{i\frac{khy'}{2R}}dy'\right|^2 \\ & +2\text{Re}\frac{e^{i\frac{khy}{D}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|E(y')\right| e^{-i\frac{khy'}{2R}}dy'\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|E(y')\right| e^{i\frac{khy''}{2R}}dy''\right\}^*\Bigg] \end{aligned} \tag{8.59}$$

(converging spherical wave)

There is a close resemblance between the expression

$$\left|E_{\text{slit one}}(h/2)\right| \equiv \left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|E(y')\right| e^{-i\frac{khy'}{2R}}dy'\right| \tag{8.60}$$

and the magnitude of the degree of coherence $V = |\gamma(h)|$ from (8.50). Here E_{slitone} denotes the field impinging on the screen that goes through the upper slit positioned at a distance $h/2$ from center. The field strength when the single slit is positioned at h compared to that when it is positioned at zero is

$$\left|\frac{E_{\text{slit one}}(h)}{E_{\text{slit one}}(0)}\right| = \left|\frac{\int_{-\infty}^{\infty}\left|E(y')\right| e^{-i\frac{khy'}{R}}dy'}{\int_{-\infty}^{\infty}\left|E(y')\right| dy'}\right| \quad \text{(converging spherical wave assumption)} \tag{8.61}$$

This looks very much like $|\gamma(h)|$ of (8.50) except that the magnitude of the field appears in (8.61), whereas the intensity appears in (8.50).

If we replace the field in (8.61) with one that is proportional to the intensity (i.e. $|E_{\text{new}}(y')| \propto I(y') \propto |E_{\text{old}}(y')|^2$), then the expression becomes the same as (8.50). This may seem rather contrived, but at least it is cute, and it is known as the van Cittert-Zernike theorem. It says that the spatial coherence of an extended source with randomly varying phase corresponds to the field distribution created by replacing the extended source with a converging spherical wave whose *field amplitude* distribution is the same as the original intensity distribution.

Exercises

8.3 Temporal Coherence

P8.1 Show that $\text{Re}\gamma(\tau)$ defined in (8.10) reduces to $\cos(\omega_0\tau)$ in the case of a plane wave $E(t) = E_0 e^{i(k_0 z - \omega_0 t)}$ being sent through a Michelson interferometer. In other words, the output intensity from the interferometer reduces to

$$I = 2I_0\left[1 + \cos(\omega_0\tau)\right]$$

as you already expect.

HINT: Don't be afraid of delta functions. After integration, the left-over delta functions cancel.

P8.2 Light emerging from a dense hot gas has a collisionally broadened power spectrum described by the Lorentzian function

$$I(\omega) = \frac{I(\omega_0)}{1 + \left(\frac{\omega - \omega_0}{\Delta\omega_{\text{FWHM}}/2}\right)^2}$$

The light is sent into a Michelson interferometer. Make a graph of the average power arriving to the detector as a function of τ.

HINT: See (0.53).

P8.3 (a) Regardless of how the phase of $E(\omega)$ is organized, the oscillation of the energy arriving to the detector as a function of τ is the same. The spectral phase of the light in P 8.2 is randomly organized. Describe qualitatively how the light probably looks as a function of time.

(b) Now suppose that the phase of the light is somehow neatly organized such that

$$E(\omega) = \frac{iE(\omega_0)\, e^{i\frac{\omega}{c}z}}{i + \frac{\omega - \omega_0}{\Delta\omega_{\text{FWHM}}/2}}$$

Perform the inverse Fourier transform on the field and find how the *intensity* of the light looks a function of time.

HINT:

$$\int_{-\infty}^{\infty} \frac{e^{-iax}}{x + \beta} dx = \begin{cases} -2i\pi e^{ia\beta} & \text{if } a>0 \\ 0 & \text{if } a<0 \end{cases} \qquad (\text{Im}\beta > 0)$$

The constants $I(\omega_0)$, and $\Delta\omega_{\text{FWHM}}$ will appear in the answer.

8.4 Fringe Visibility and Coherence Length

P8.4 (a) Verify (8.19). HINT: Write $\gamma = |\gamma|\, e^{i\phi}$ and assume that the oscillations in γ that give rise to fringes are due entirely to changes in ϕ and that $|\gamma|$ is a slowly varying function in comparison to the oscillations.

(b) What is the coherence time τ_c of the light in P 8.2?

P8.5 (a) Show that the fringe visibility of the Gaussian distribution (8.20) (i.e. the magnitude of γ in (8.21)) goes from 1 to $e^{-\pi/2} = 0.21$ as the round-trip path in one arm of the instrument is extended by a coherence length.

(b) Find the FWHM bandwidth in *wavelength* $\Delta\lambda_{\text{FWHM}}$ in terms of the coherence length ℓ_c and the center wavelength λ_0 associated with (8.20).

HINT: Derive $\Delta\omega_{\text{FWHM}} = 2\sqrt{\ln 2}\Delta\omega$. To convert to a wavelength difference, use $\lambda = \frac{2\pi c}{\omega} \Rightarrow \Delta\lambda \cong -\frac{2\pi c}{\omega^2}\Delta\omega$. You can ignore the minus sign; it simply means that wavelength decreases as frequency increases.

8.5 Fourier Spectroscopy

L8.6 (a) Use a scanning Michelson interferometer to measure the wavelength of ultrashort laser pulses produced by a mode-locked Ti:sapphire oscillator.

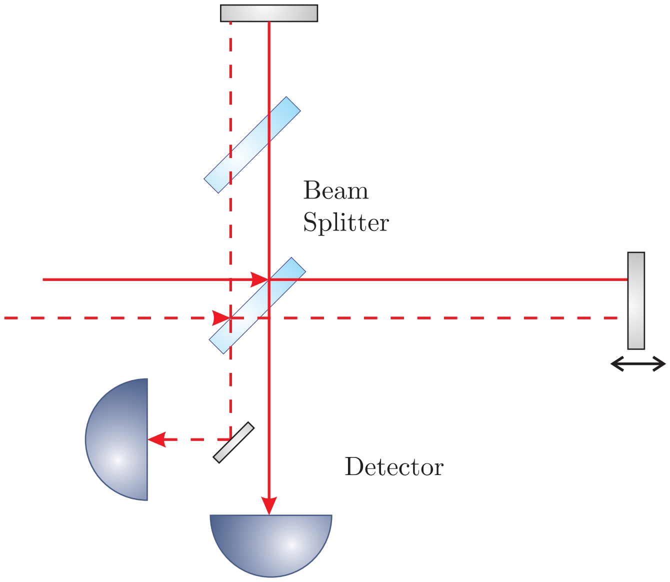


Figure 8.7

(b) Measure the coherence length of the source by observing the distance over which the visibility diminishes. From your measurement, what is the bandwidth $\Delta\lambda_{\text{FWHM}}$ of the source, assuming the Gaussian profile in the previous problem? See P 8.5.

(c) Use a computer to perform a fast Fourier transform (FFT) of the signal output. For the positive frequencies, plot the laser spectrum as a function of λ and compare with the results of (a) and (b).

(d) How do the results change if the ultrashort pulses are first stretched in time by traversing a thick piece of glass?

8.6 Young's Two-Slit Setup and Spatial Coherence

P8.7 (a) A point source with wavelength $\lambda = 500$ nm illuminates two parallel slits separated by $h = 1.0$ mm. If the screen is $D = 2$ m away, what is the separation between the diffraction peaks on the screen? Make a sketch.

(b) A thin piece of glass with thickness $d = 0.01$ mm and index $n = 1.5$ is placed in front of one of the slits. By how many fringes does the pattern at the screen move?

HINT: This effectively introduces a relative phase $\Delta\phi$ in (8.32). Compare the phase of the light when traversing the glass versus traversing an empty region of the same thickness.

L8.8 (a) Carefully measure the separation of a double slit in the lab ($h \sim 1$ mm separation) by shining a HeNe laser ($\lambda = 633$ nm) through it and measuring the diffraction peak separations on a distant wall (say, 2 m from the slits).

HINT: For better accuracy, measure across several fringes and divide.

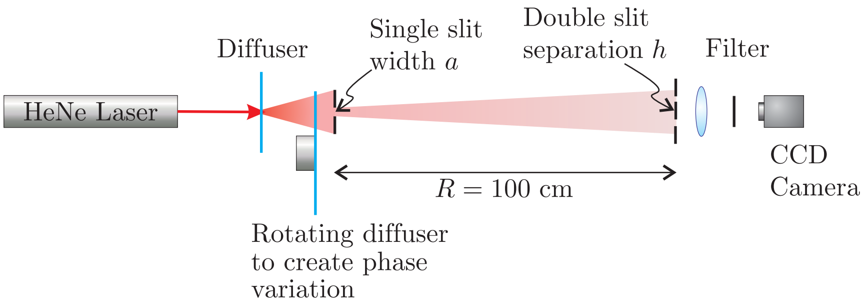

Diffuser
Single slit width a
Double slit separation h
Filter
HeNe Laser
CCD Camera
$R = 100$ cm
Rotating diffuser to create phase variation

Figure 8.8

(b) Create an extended light source with a HeNe laser using a time-varying diffuser followed by an adjustable single slit. (The diffuser must rotate rapidly to create random time variation of the phase at each point as would occur automatically for a natural source such as a star.) Place the double slit at a distance of $R \approx 100$ cm after the first slit. (Take note of the exact value of R, as you will need it for the next problem.) Use a lens to image the diffraction pattern that would have appeared on a far-away screen into a video camera. Observe the visibility of the fringes. Adjust the width of the source with the single slit until the visibility of the fringes disappears. After making the source wide enough to cause the fringe pattern to degrade, measure the single slit width a by shining a HeNe laser through it and observing the diffraction pattern on the distant wall.

HINT: A single slit of width a produces an intensity pattern described by Eq. (11.45) with $N = 1$ and $\Delta x = a$.

NOTE: It would have been nicer to vary the separation of the two slits to determine the width of a fixed source. However, because it is hard to make an adjustable double slit, we varied the size of the source until the spatial coherence of the light matched the slit separation.

P8.9 (a) Compute h_c for a uniform intensity distribution of width a using (8.48).

(b) Use this formula to check that your measurements in L 8.8 agree with spatial coherence theory.

HINT: In your experiment h_c is the double slit separation. Use your measured R and h to calculate what the width of the single slit (i.e. a) should have been when the fringes disappeared and compare this calculation to your direct measurement of a.

Solution: (This is only a partial solution)

$$\gamma\left(h\right) = \frac{\int\limits_{-a/2}^{a/2} I_0 \exp\left[-ikh\left(\frac{y'}{R}+\frac{y}{D}\right)\right] dy'}{\int\limits_{-a/2}^{a/2} I_0 dy'} = \frac{e^{-ikh\frac{y}{D}} \int\limits_{-a/2}^{a/2} e^{-ikh\frac{y'}{R}} dy'}{a} = \frac{e^{-ikh\frac{y}{D}} \left[\frac{e^{-ikh\frac{y'}{R}}}{-i\frac{kh}{R}}\right]_{-a/2}^{a/2}}{a}$$

$$= e^{-ikh\frac{y}{D}} \left[\frac{e^{-ikh\frac{a/2}{R}} - e^{-ikh\frac{-a/2}{R}}}{-2ikh\frac{a/2}{R}}\right] = e^{-ikh\frac{y}{D}} \operatorname{sinc}\frac{kha}{2R}$$

Note that

$$\int\limits_0^\infty \frac{\sin^2 \alpha x}{\left(\alpha x\right)^2} dx = \frac{\pi}{2\alpha}$$

Review, Chapters 6–8

True and False Questions

R24 T or F: It is always possible *to completely eliminate* reflections with a single-layer antireflection coating as long as the right thickness is chosen for a given real index.

R25 T or F: For a given incident angle and value of n, there is only one single-layer coating thickness d that will minimize reflections.

R26 T or F: When coating each surface of a lens with a single-layer antireflection coating, the thickness of the coating on the exit surface will need to be different from the thickness of the coating on the entry surface.

R27 T or F: In our notation (widely used), $I(t)$ is the Fourier transform of $I(\omega)$.

R28 T or F: The integral of $I(t)$ over all t equals the integral of $I(\omega)$ over all ω.

R29 T or F: The phase velocity of light (the speed of an individual frequency component of the field) never exceeds the speed of light c.

R30 T or F: The group velocity of light in a homogeneous material can exceed c if absorption or amplification takes place.

R31 T or F: The group velocity of light never exceeds the phase velocity.

R32 T or F: A Michelson interferometer can be used to measure the spectral intensity of light $I(\omega)$.

R33 T or F: A Michelson interferometer can be used to measure the duration of a short laser pulse and thereby characterize its chirp.

R34 T or F: A Michelson interferometer can be used to measure the wavelength of light.

R35 T or F: A Michelson interferometer can be used to measure the phase of $E(\omega)$.

R36 T or F: The Fourier transform (or inverse Fourier transform if you prefer) of $I(\omega)$ is proportional to the degree of temporal coherence.

R37 T or F: A Michelson interferometer is ideal for measuring the *spatial* coherence of light.

R38 T or F: The Young's two-slit setup is ideal for measuring the *temporal* coherence of light.

R39 T or F: Vertically polarized light illuminates a Young's double-slit setup and fringes are seen on a distant screen with good visibility. A half wave plate is placed in front of one of the slits so that the polarization for that slit becomes horizontally polarized. **Here's the statement:** The fringes at the screen will shift position but maintain their good visibility.

Problems

R40 A thin glass plate with index $n = 1.5$ is oriented at Brewster's angle so that p-polarized light with wavelength $\lambda_{\mathrm{vac}} = 500$ nm goes through with 100% transmittance.

(a) What is the minimum thickness that will make the reflection of s-polarized light be maximum?

(b) What is the transmittance T_s^{tot} for this thickness assuming s-polarized light? HINT:

$$r_s = -\frac{\sin(\theta_{\mathrm{i}} - \theta_{\mathrm{t}})}{\sin(\theta_i + \theta_t)}, \quad r_p = -\frac{\tan(\theta_{\mathrm{i}} - \theta_{\mathrm{t}})}{\tan(\theta_{\mathrm{i}} + \theta_{\mathrm{t}})}, \quad t_s = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin(\theta_{\mathrm{i}} + \theta_{\mathrm{t}})}$$

$$T_s^{\mathrm{tot}} = \frac{T_s^{\max}}{1 + F_s \sin^2\left(\frac{\Phi}{2}\right)} \quad (\theta_{\mathrm{m}} \text{ real})$$

$$T_s^{\max} \equiv \frac{n_{\mathrm{t}}\cos\theta_{\mathrm{t}} \left|t_s^{\mathrm{i}\to\mathrm{m}}\right|^2 \left|t_s^{\mathrm{m}\to\mathrm{t}}\right|^2}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} \left(1 - \left|r_s^{\mathrm{m}\to\mathrm{i}}\right| \left|r_s^{\mathrm{m}\to\mathrm{t}}\right|\right)^2}$$

$$F_s \equiv \frac{4\left|r_s^{m\to i}\right| \left|r_s^{m\to t}\right|}{\left(1 - \left|r_s^{m\to i}\right| \left|r_s^{m\to t}\right|\right)^2}$$

$$\Phi = \delta + \delta_{r_s}, \quad \delta \equiv 2k_{\mathrm{m}} d\cos\theta_{\mathrm{m}}, \quad \delta_{r_s} \equiv \delta_{r_s^{\mathrm{m}\to\mathrm{i}}} + \delta_{r_s^{\mathrm{m}\to\mathrm{t}}}$$

$$r_s^{\mathrm{m}\to\mathrm{i}} = \left|r_s^{\mathrm{m}\to\mathrm{i}}\right| e^{i\delta_{r_s^{\mathrm{m}\to\mathrm{i}}}}, \quad r_s^{\mathrm{m}\to\mathrm{t}} = \left|r_s^{\mathrm{m}\to\mathrm{t}}\right| e^{i\delta_{r_s^{\mathrm{m}\to\mathrm{t}}}}$$

R41 Consider a Fabry-Perot interferometer. Note: $R_1 = R_2 = R$.

(a) Show that the free spectral range for a Fabry-Perot interferometer is

$$\Delta\lambda_{\mathrm{FSR}} = \frac{\lambda^2}{2nd\cos\theta}$$

(b) Show that the fringe width $\Delta\lambda_{\mathrm{FWHM}}$ is

$$\frac{\lambda^2}{\pi\sqrt{F}nd\cos\theta}$$

where $F \equiv \frac{4R}{(1-R)^2}$.

(c) Derive the reflecting finesse $f = \Delta\lambda_{\mathrm{FSR}}/\Delta\lambda_{\mathrm{FWHM}}$.

R42 For a Fabry-Perot etalon, let $R = 0.90$, $\lambda_{\text{vac}} = 500$ nm, $n = 1$, and $d = 5.0$ mm.

(a) Suppose that a maximum transmittance occurs at the angle $\theta = 0$. What is the nearest angle where the transmittance will be half of the maximum transmittance? You may assume that $\cos\theta \cong 1 - \theta^2/2$.

(b) You desire to use a Fabry-Perot etalon to view the light from a large diffuse source rather than a point source. Draw a diagram depicting where lenses should be placed, indicating relevant distances. Explain briefly how it works.

R43 You need to make an antireflective coating for a glass lens designed to work at normal incidence.

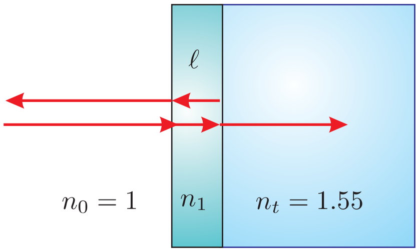


Figure 8.9

The matrix equation relating the incident field to the reflected and transmitted fields (at normal incidence) is

$$\begin{bmatrix} 1 \\ n_0 \end{bmatrix} + \begin{bmatrix} 1 \\ -n_0 \end{bmatrix} \frac{E_0^{\text{reflected}}}{E_0^{\text{incident}}} = \begin{bmatrix} \cos k_1\ell & \frac{-i}{n_1}\sin k_1\ell \\ -in_1 \sin k_1\ell & \cos k_1\ell \end{bmatrix} \begin{bmatrix} 1 \\ n_{\text{t}} \end{bmatrix} \frac{E_{\text{t}}^{\text{transmitted}}}{E_0^{\text{incident}}}$$

(a) What is the minimum thickness the coating should have?

HINT: It is less work if you can figure this out without referring to the above equation. You may assume $n_1 < n_{\text{t}}$.

(b) Find the index of refraction n_1 that will make the reflectivity be zero.

R44 (a) What is the spectral content (i.e., $I(\omega)$) of a square laser pulse

$$E(t) = \begin{cases} E_0 e^{-i\omega_0 t} & , \quad |t| \le \tau/2 \\ 0 & , \quad |t| > \tau/2 \end{cases}$$

Make a sketch of $I(\omega)$, indicating the location of the first zeros.

(b) What is the temporal shape (i.e., $I(t)$) of a light pulse with frequency content

$$E(\omega) = \begin{cases} E_0 & , \quad |\omega - \omega_0| \le \Delta\omega/2 \\ 0 & , \quad |\omega - \omega_0| > \Delta\omega/2 \end{cases}$$

where in this case E_0 has units of E-field per frequency. Make a sketch of $I(t)$, indicating the location of the first zeros.

(c) If $E(\omega)$ is known (any arbitrary function, not the same as above), and the light goes through a material of thickness ℓ and index of refraction $n(\omega)$, how would you find the form of the pulse $E(t)$ after passing through the material? Please set up the integral.

R45 (a) Prove Parseval's theorem:

$$\int_{-\infty}^{\infty} |E(\omega)|^2 \, d\omega = \int_{-\infty}^{\infty} |E(t)|^2 \, dt.$$

HINT:

$$\delta\left(t' - t\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega(t'-t)} d\omega$$

(b) Explain the physical relevance of Parseval's theorem to light pulses. Suppose that you have a detector that measures the total energy in a pulse of light, say 1 mJ directed onto an area of 1 mm^2. Next you measure the spectrum of light and find it to have a width of $\Delta\lambda = 50$ nm, centered at $\lambda_0 = 800$ nm. Assume that the light has a Gaussian frequency profile

$$I(\omega) = I(\omega_0) e^{-\left(\frac{\omega - \omega_0}{\delta\omega}\right)^2}$$

Use as an approximate value $\delta\omega \cong \frac{2\pi c}{\lambda^2} \Delta\lambda$. Find a value and correct units for $I(\omega_0)$.

HINT:

$$\int_{-\infty}^{\infty} e^{-Ax^2 + Bx + C} dx = \sqrt{\frac{\pi}{A}} e^{B^2/4A + C} \qquad \mathrm{Re}\{A\} > 0$$

R46 Continuous light entering a Michelson interferometer has a spectrum described by

$$I(\omega) = \begin{cases} I_0 & , \quad |\omega - \omega_0| \le \Delta\omega/2 \\ 0 & , \quad |\omega - \omega_0| > \Delta\omega/2 \end{cases}$$

The Michelson interferometer uses a 50:50 beam splitter. The emerging light has intensity $\langle I_{\mathrm{det}}(t, \tau)\rangle_t = 2 \langle I(t)\rangle_t [1 + \mathrm{Re}\gamma(\tau)]$, where degree of coherence is

$$\gamma(\tau) = \int_{-\infty}^{\infty} I(\omega) e^{-i\omega\tau} d\omega \Bigg/ \int_{-\infty}^{\infty} I(\omega) d\omega$$

Find the fringe visibility $V \equiv (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$ as a function of τ (i.e. the round-trip delay due to moving one of the mirrors).

R47 Light emerging from a point travels by means of two very narrow slits to a point y on a screen. The intensity at the screen arising from a *point* source at position y' is found to be

$$I_{\mathrm{screen}}\left(y', h\right) = 2I(y') \left\{1 + \cos\left[kh\left(\frac{y}{D} + \frac{y'}{R}\right)\right]\right\}$$

where an approximation has restricted us to small angles.

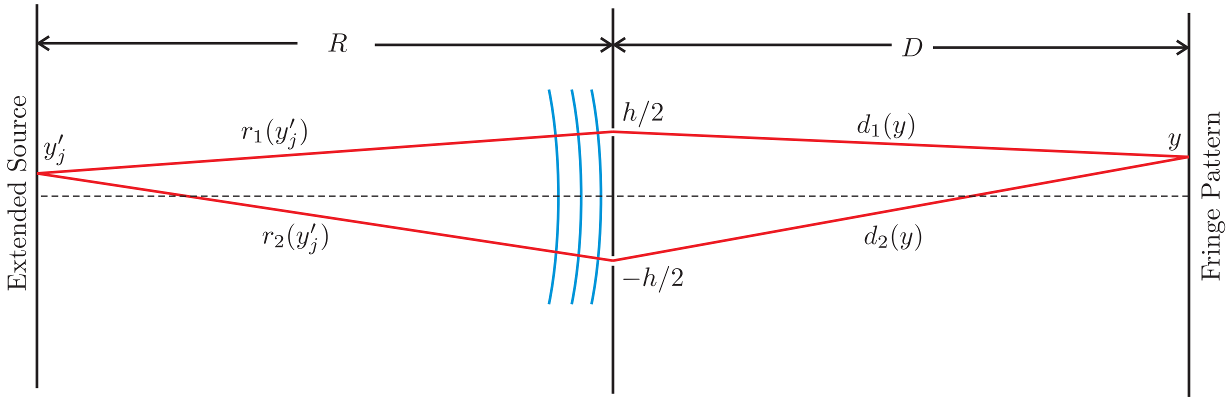


Figure 8.10

(a) Now, suppose that $I(y')$ characterizes emission from a wider source with randomly varying phase across its width. Write down an expression (in integral form) for the resulting intensity at the screen:

$$I_{\text{screen}}\left(h\right) \equiv \int_{-\infty}^{\infty} I_{\text{screen}}\left(y', h\right) dy'$$

(b) Assume that the source has an emission distribution with the form $I(y') = (I_0/\Delta y')\, e^{-y'^2/\Delta y'^2}$. What is the function $\gamma(h)$ where the intensity is written $I_{\text{screen}}(h) = 2\sqrt{\pi} I_0 \left[1 + \text{Re}\gamma(h)\right]$?

HINT:

$$\int_{-\infty}^{\infty} e^{-Ax^2+Bx+C} dx = \sqrt{\frac{\pi}{A}} e^{B^2/4A+C} \qquad \text{Re}\left\{A\right\} > 0.$$

(c) As h varies, the intensity at a point on the screen y oscillates. As h grows wider, the amplitude of oscillations decreases. How wide must the slit separation h become (in terms of R, k, and $\Delta y'$) to reduce the visibility to

$$V \equiv \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{1}{3}$$

Selected Answers

R40: (a) 100 nm. (b) 0.55.

R42: (a) 0.074°.

R43: (b) 1.24.

R45: (b) $3.8 \times 10^{-16} \text{J}/\left(\text{cm}^2 \cdot \text{s}^{-1}\right)$.

Chapter 9

Light as Rays

9.1 Introduction

So far in our study of optics, we have described light in terms of waves, which satisfy Maxwell's equations. However, as is well known to students, in many situations light can be thought of as rays directed along the flow of energy. A ray picture is useful when one is interested in the macroscopic distribution of light energy, but rays fail to reveal how intensity varies when light is concentrated in small regions of space. Moreover, simple ray theory suggests that a lens can focus light down to a point. However, if a beam of light were concentrated onto a true point, the intensity would be infinite! In this scenario ray theory can clearly not be used to predict the intensity profile in a focus. In this case, it is necessary to consider waves and diffraction phenomena. Nevertheless, ray theory is useful for predicting where a focus occurs. It is also useful for describing imaging properties of optical systems (e.g. lenses and mirrors).

Beginning in section 9.4 we study the details of ray theory and the imaging properties of optical systems. First, however, we examine the justification for ray theory starting from Maxwell's equations. Section 9.2 gives a derivation of the *eikonal* equation, which governs the direction of rays in a medium with an index of refraction that varies with position. The word "eikonal" comes from the Greek "$\epsilon\iota\kappa\omega\nu$s" from which the modern word "icon" derives. The eikonal equation therefore has a descriptive title since it controls the formation of images. Although we will not use the eikonal equation extensively, we will show how it embodies the underlying justification for ray theory. As will be apparent in its derivation, the eikonal equation relies on an approximation that the features of interest in the light distribution are large relative to the wavelength of the light.

The eikonal equation describes the flow of energy in an optical medium. This applies even to complicated situations such as desert mirages where air is heated near the ground and has a different index than the air further from the ground. Rays of light from the sky that initially are directed toward the ground can be bent such that they travel parallel to the ground owing to the inhomogeneous refractive index. If the index of refraction as a function of position is known, the eikonal equation can be used to determine the propagation of such rays. This also applies to practical problems such as the propagation of rays through lenses (where the index also varies with position).

In section 9.3, we deduce Fermat's principle from the eikonal equation. Of course Fermat asserted his principle more than a century before Maxwell assembled his equations, but it is nice to give justification retroactively to Fermat's principle using the modern perspective. In short, Fermat asserted that light travels from point A to point B following a path that takes the minimum time.

In section 9.4, we begin our study of *paraxial ray theory*, which is used to analyze the propagation of rays through optical systems composed of lenses and/or curved mirrors. The paraxial approximation restricts rays to travel nearly parallel to the axis of such a system. We consider the effects of three different optical elements acting on paraxial rays. The first element is simply the unobstructed propagation of a ray through a distance d in a uniform medium; if the ray is not exactly parallel to the optical axis, then it moves further away from (or closer to) the optical axis as it travels. The second element is a curved spherical mirror, which reflects a ray and changes its angle. The third element, which is similar, is a spherical interface between two materials with differing refractive indices. We demonstrate that the effects of each of these elements on a ray of light can be represented as a 2×2 matrix. These three basic elements can be combined to construct more complex imaging systems (such as a lens or a series of lenses and curved mirrors). The overall effect of a complex system on a ray can be computed by multiplying together the matrices associated with each of the basic elements.

We discuss the condition for image formation in section 9.6 and make contact with the familiar formula

$$\frac{1}{f} = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} \tag{9.1}$$

which describes the location of images produced by curved mirrors or *thin lenses*. In section 9.7 we introduce the concept of *principal planes*, which exist for multi-element optical systems. If the distance d_o is measured from one principal plane while d_i is measured from a second principal plane, then the thin lens formula (9.1) can be applied even to complicated systems with an appropriate effective focal length f_eff.

Finally, in section 9.8 we use paraxial ray theory to study the stability of laser cavities. The ray formalism can be used to predict whether a ray, after many round trips in the cavity, remains near the optical axis (trapped and therefore stable) or if it drifts endlessly away from the axis of the cavity on successive round trips. In appendix 9.9 we address deviations from the paraxial ray theory known as aberrations. We also comment on ray-tracing techniques, used for designing optical systems that minimize such aberrations.

9.2 The Eikonal Equation

We begin with the wave equation (2.20) for a medium with a real index of refraction:

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) - \frac{n^2(\mathbf{r})}{c^2} \frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2} = 0 \tag{9.2}$$

Although in chapter 2 we considered solutions to the wave equation in a homogeneous material, the wave equation is also perfectly valid when the index of refraction varies throughout space. Here we allow the medium (i.e. the density) to vary with position. Hence the index $n(\mathbf{r})$ is an arbitrary function of $\mathbf{r}$. In this case, the usual plane-wave solutions no longer satisfy the wave equation.

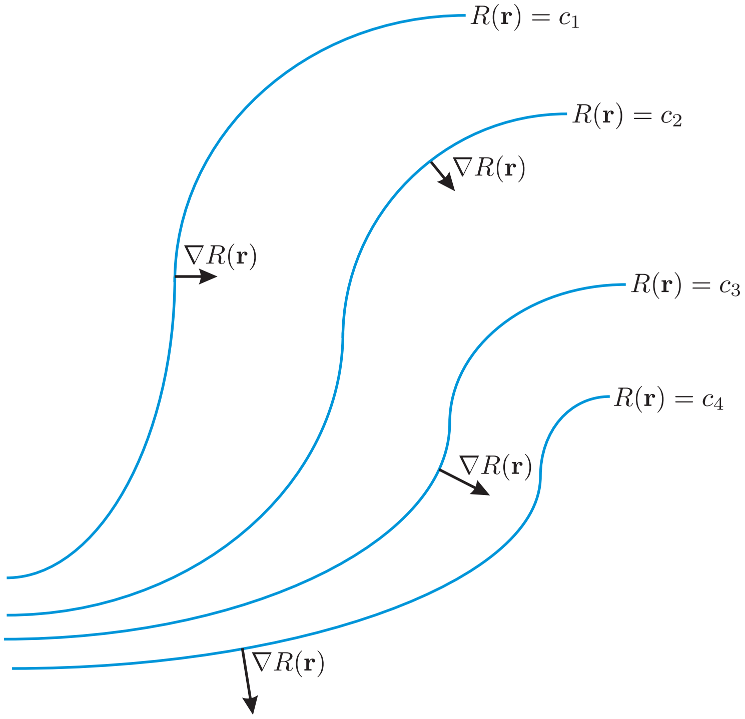


Figure 9.1 Wave fronts distributed throughout space in the presence of a spatially inhomogeneous refractive index.

We consider the light to have a single frequency ω. As a trial solution for (9.2), we take

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0(\mathbf{r})\, e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} \tag{9.3}$$

where

$$k_{\text{vac}} = \frac{\omega}{c} = \frac{2\pi}{\lambda_{\text{vac}}} \tag{9.4}$$

Here $R(\mathbf{r})$ is a real scalar function (which depends on position) having the dimension of length. By assuming that $R(\mathbf{r})$ is real, we do not account for absorption or amplification in the medium. Even though the trial solution (9.3) looks somewhat like a plane wave, the function $R(\mathbf{r})$ accommodates wave fronts that can be curved or distorted as depicted in Fig. 9.1. At any given instant t, the phase of the curved surfaces described by $R(\mathbf{r}) =$ constant can be interpreted as wave fronts of the solution. The wave fronts travel in the direction for which $R(\mathbf{r})$ varies the fastest. This direction is given by $\nabla R(\mathbf{r})$, which lies in the direction perpendicular to surfaces of constant phase.

Note that if the index is spatially independent (i.e. $n(\mathbf{r}) \to n$), then (9.3) reduces to the usual plane-wave solution of the wave equation. In this case, we have $R(\mathbf{r}) = \mathbf{k}\cdot\mathbf{r}/k_{\text{vac}}$ and the field amplitude becomes constant (i.e. $\mathbf{E}_0(\mathbf{r}) \to \mathbf{E}_0$).

The substitution of the trial solution (9.3) into the wave equation (9.2) gives

$$\nabla^2\left[\mathbf{E}_0(\mathbf{r})\, e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}\right] + \frac{n^2(\mathbf{r})\,\omega^2}{c^2}\mathbf{E}_0(\mathbf{r})\, e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} = 0 \tag{9.5}$$

We divide each term by $e^{-i\omega t}$ and utilize (9.4) to rewrite the wave equation as

$$\frac{1}{k_{\text{vac}}^2}\nabla^2\left[\mathbf{E}_0(\mathbf{r})\,e^{ik_{\text{vac}}R(\mathbf{r})}\right]+n^2(\mathbf{r})\,\mathbf{E}_0(\mathbf{r})\,e^{ik_{\text{vac}}R(\mathbf{r})}=0 \tag{9.6}$$

Our next task is to evaluate the spatial derivative, which is worked out in the following example.

Example 9.1

Compute the Laplacian needed in (9.6).

Solution: The gradient of the x component of the field is

$$\nabla\left[E_{0x}(\mathbf{r})\,e^{ik_{\text{vac}}R(\mathbf{r})}\right]=\left[\nabla E_{0x}(\mathbf{r})\right]e^{ik_{\text{vac}}R(\mathbf{r})}+ik_{\text{vac}}E_{0x}(\mathbf{r})\left[\nabla R(\mathbf{r})\right]e^{ik_{\text{vac}}R(\mathbf{r})}$$

The Laplacian of the x component is

$$\begin{aligned}\nabla\cdot\nabla\left[E_{0x}(\mathbf{r})\,e^{ik_{\text{vac}}R(\mathbf{r})}\right]=&\left\{\nabla^2E_{0x}(\mathbf{r})-k_{\text{vac}}^2E_{0x}(\mathbf{r})\left[\nabla R(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]\right.\\&\left.+ik_{\text{vac}}E_{0x}(\mathbf{r})\left[\nabla^2R(\mathbf{r})\right]+2ik_{\text{vac}}\left[\nabla E_{0x}(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]\right\}e^{ik_{\text{vac}}R(\mathbf{r})}\end{aligned}$$

Upon combining the result for each vector component of $\mathbf{E}_0(\mathbf{r})$, the required spatial derivative can be written as

$$\begin{aligned}\nabla^2\left[\mathbf{E}_0(\mathbf{r})\,e^{ik_{\text{vac}}R(\mathbf{r})}\right]=&\left(\nabla^2\mathbf{E}_0(\mathbf{r})-k_{\text{vac}}^2\mathbf{E}_0(\mathbf{r})\left[\nabla R(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]+ik_{\text{vac}}\mathbf{E}_0(\mathbf{r})\left[\nabla^2R(\mathbf{r})\right]\right.\\&+2ik_{\text{vac}}\left\{\hat{\mathbf{x}}\left[\nabla E_{0x}(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]+\hat{\mathbf{y}}\left[\nabla E_{0y}(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]\right.\\&\left.\left.+\ \hat{\mathbf{z}}\left[\nabla E_{0z}(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]\right\}\right)e^{ik_{\text{vac}}R(\mathbf{r})}\end{aligned}$$

Using the result from Example 9.1 with some additional rearranging, (9.6) becomes

$$\begin{aligned}\left[\nabla R(\mathbf{r})\cdot\nabla R(\mathbf{r})-n^2(\mathbf{r})\right]\mathbf{E}_0(\mathbf{r})=&\frac{\nabla^2\mathbf{E}_0(\mathbf{r})}{k_{\text{vac}}^2}+\frac{i}{k_{\text{vac}}}\nabla^2R(\mathbf{r})+\frac{2i}{k_{\text{vac}}}\hat{\mathbf{x}}\nabla E_{0x}(\mathbf{r})\cdot\nabla R(\mathbf{r})\\&+\frac{2i}{k_{\text{vac}}}\left[\hat{\mathbf{y}}\left[\nabla E_{0y}(\mathbf{r})\right]\cdot\nabla R(\mathbf{r})+\hat{\mathbf{z}}\nabla E_{0z}(\mathbf{r})\cdot\nabla R(\mathbf{r})\right]\end{aligned} \tag{9.7}$$

At this point we are ready to make an important approximation. We take the limit of a very short wavelength (i.e. $1/k_{\text{vac}}=\lambda_{\text{vac}}/2\pi\to 0$). This means that we lose the effects of diffraction. We also lose surface reflections at abrupt index changes unless specifically considered. This approximation works best in situations where only macroscopic features are of concern. Under the assumption of an infinitesimal wavelength, the entire right-hand side of (9.7) vanishes (thank goodness) and the wave equation imposes

$$\left[\nabla R(\mathbf{r})\right]\cdot\left[\nabla R(\mathbf{r})\right]=n^2(\mathbf{r}), \tag{9.8}$$

Written another way, this equation is

$$\nabla R(\mathbf{r})=n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r}) \tag{9.9}$$

This latter form is called the *eikonal equation* where $\hat{\mathbf{s}}$ is a unit vector pointing in the direction $\nabla R(\mathbf{r})$, the direction normal to wave front surfaces.

Pierre de Fermat

(1601–1665, French)

Fermat was a distinguished mathematician. He loved to publish results, but was often quite secretive about the methods used to obtain his results. Fermat was the first to state that the path taken by a beam of light is the one that can be traveled in the least amount of time.

Under the assumption of an infinitely short wavelength, the Poynting vector is directed along $\hat{\mathbf{s}}$ as demonstrated in P 9.2. In other words, the direction of $\hat{\mathbf{s}}$ specifies the direction of energy flow. The unit vector $\hat{\mathbf{s}}$ at each location in space points perpendicular to the wave fronts and indicates the direction that the waves travel as seen in Fig. 9.1. We refer to a collection of vectors $\hat{\mathbf{s}}$ distributed throughout space as *rays*.

In retrospect, we might have jumped straight to (9.9) without going through the above derivation. After all, we know that each part of a wave front advances in the direction of its gradient $\nabla R(\mathbf{r})$ (i.e. in the direction that $R(\mathbf{r})$ varies most rapidly). We also know that each part of a wave front defined by $R(\mathbf{r}) =$ constant travels at speed $c/n(\mathbf{r})$. The slower a given part of the wave front advances, the more rapidly $R(\mathbf{r})$ changes with position $\mathbf{r}$ and the closer the contours of constant phase. It follows that $\nabla R(\mathbf{r})$ must be proportional to $n(\mathbf{r})$ since $\nabla R(\mathbf{r})$ denotes the rate of change in $R(\mathbf{r})$.

9.3 Fermat's Principle

The eikonal equation (9.9) governs the path that rays follow as they traverse a region of space, where the index varies as a function of position. An analysis of the eikonal equation renders Fermat's principle as we now show. We begin by taking the curl of (9.9) to obtain

$$\nabla \times [n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r})] = \nabla \times [\nabla R(\mathbf{r})] = 0 \tag{9.10}$$

(The curl of a gradient is identically zero for any function $R(\mathbf{r})$.) Integration of (9.10) over an open surface of area A results in

$$\int\limits_A \nabla \times [n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r})]\,da = 0 \tag{9.11}$$

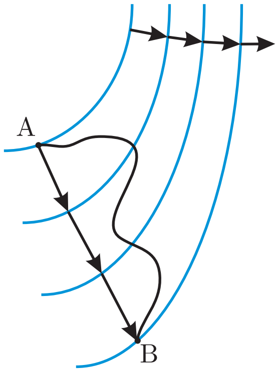


Figure 9.2 A ray of light leaving point A arriving at B.

We next apply Stokes' theorem (0.27) to the integral and convert it to a path integral around the perimeter of the area. Then we get

$$\oint_C n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r})\cdot d\ell = 0 \tag{9.12}$$

The integration of $n\hat{\mathbf{s}}\cdot d\ell$ around a closed loop is always zero. Keep in mind that the proper value for $\hat{\mathbf{s}}(\mathbf{r})$ must be used, and this is determined by the eikonal equation (9.9). Equation (9.12) implies Fermat's principle, but to see this fact requires some subtle arguments.

Equation (9.12) implies the following:

$$\int_A^B n\hat{\mathbf{s}}\cdot d\ell \quad \text{is independent of path from A to B.} \tag{9.13}$$

Now consider points A and B that lie along a path that is always parallel to $\hat{\mathbf{s}}$ (i.e. perpendicular to the wave fronts as depicted in Fig. 9.2). When integrating along the path parallel to $\hat{\mathbf{s}}$, the cosine in the dot product in (9.13) is always one. If we choose some other path that connects A and B, the cosine associated with the dot product is often less than one. Since in both cases the result of the integral must be the same, the other factors inside the integral must render a larger value to compensate for the cosine term's occasional dip below unity when the path is not parallel to $\hat{\mathbf{s}}$. Thus, if we artificially remove the dot product from the integral (i.e. exclude the cosine factor), the result of the integral is smallest when the path is taken along the direction of $\hat{\mathbf{s}}$.

With the dot product removed from (9.13), the result of the integration agrees with the true result only for the path taken along $\hat{\mathbf{s}}$ (i.e. only for the path that corresponds to the one that light rays actually follow). In mathematical form, this argument can be expressed as

$$\int_A^B n\hat{\mathbf{s}}\cdot d\ell = \min\left\{\int_A^B n d\ell\right\} \tag{9.14}$$

The integral on the right gives the *optical path length* (OPL) between A and B:

$$OPL|_A^B \equiv \int_A^B n d\ell \tag{9.15}$$

where the n in general can be different for each of the incremental distances $d\ell$. The conclusion is that the true path that light follows between two points (i.e. the one that follows along $\hat{\mathbf{s}}$) is the one with smallest optical path length.

Fermat's principle is usually stated in terms of the time it takes light to travel between points. The travel time Δt depends not only on the path taken by the light but also on the velocity of the light $v(\mathbf{r})$, which varies spatially with the refractive index:

$$\Delta t|_A^B = \int_A^B \frac{d\ell}{v(\mathbf{r})} = \int_A^B \frac{d\ell}{c/n(\mathbf{r})} = \frac{OPL|_A^B}{c} \tag{9.16}$$

Fermat's principle is then described as follows: Consider a source of light at some point A in space. Rays may emanate from point A in many different directions. Now consider another point B in space where the light from the first point is to be observed. Under ordinary circumstances, only one of the many rays leaving point A will pass through the point B. Fermat's principle states that the ray crossing the second point takes the path that requires the least time to travel between the two points. It should be noted that Fermat's principle, as we have written it, does not work for non-isotropic media such as crystals where n depends on the direction of a ray as well as on its location (see P 9.4).

To find the correct path for the light ray that leaves point A and crosses point B, we need only minimize the optical path length between the two points. Minimizing the optical path length is equivalent to minimizing the time of travel since it differs from the time of travel only by the constant c. The optical path length is not the actual distance that the light travels; it is proportional to the number of wavelengths that fit into that distance (see (2.26)). Thus, as the wavelength shortens due to a higher index of refraction, the optical path length increases. The correct ray traveling from A to B does not necessarily follow a straight line but can follow a complicated curve according to how the index varies.

Example 9.2

Use Fermat's principle to derive Snell's law.

Solution: Consider the many rays of light that leave point A seen in Fig. 9.3. Only one of the rays passes through point B. Within each medium we expect the light to travel in a straight line since the index is uniform. However, at the boundary we must allow for bending since the index changes.

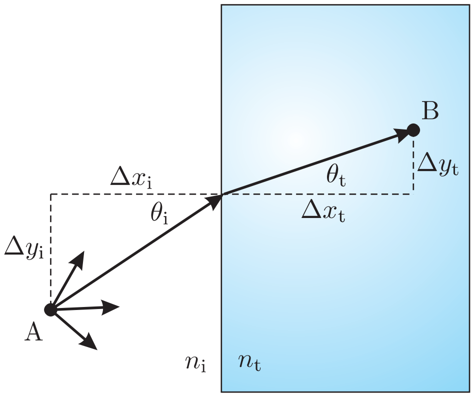


Figure 9.3 Rays of light leaving point A; not all of them will traverse point B.

The optical path length between points A and B (in terms of the unknown coordinate of the point where the ray penetrates the interface) is

$$OPL = n_{\mathrm{i}}\sqrt{\Delta x_{\mathrm{i}}^2 + \Delta y_{\mathrm{i}}^2} + n_{\mathrm{t}}\sqrt{\Delta x_{\mathrm{t}}^2 + \Delta y_{\mathrm{t}}^2} \tag{9.17}$$

We need to minimize this optical path length to find the correct one according to Fermat's principle.

Since points A and B are fixed, we may regard Δx_{i} and Δx_{t} as constants. The distances Δy_{i} and Δy_{t} are not constants although the combination

$$\Delta y_{\mathrm{tot}} = \Delta y_{\mathrm{i}} + \Delta y_{\mathrm{t}} \tag{9.18}$$

is constant. Thus, we may rewrite (9.17) as

$$OPL\left(\Delta y_{\mathrm{i}}\right) = n_{\mathrm{i}}\sqrt{\Delta x_{\mathrm{i}}^2 + \Delta y_{\mathrm{i}}^2} + n_{\mathrm{t}}\sqrt{\Delta x_{\mathrm{t}}^2 + \left(\Delta y_{\mathrm{tot}} - \Delta y_{\mathrm{i}}\right)^2} \tag{9.19}$$

where everything in the right-hand side of the expression is constant except for Δy_{i}.

We now minimize the optical path length by taking the derivative and setting it equal to zero:

$$\frac{dOPL}{d\Delta y_{\mathrm{i}}} = n_i \frac{\Delta y_{\mathrm{i}}}{\sqrt{\Delta x_{\mathrm{i}}^2 + \Delta y_{\mathrm{i}}^2}} + n_{\mathrm{t}} \frac{-\left(\Delta y_{\mathrm{tot}} - \Delta y_{\mathrm{i}}\right)}{\sqrt{\Delta x_{\mathrm{t}}^2 + \left(\Delta y_{\mathrm{tot}} - \Delta y_{\mathrm{i}}\right)^2}} = 0 \tag{9.20}$$

Notice that

$$\sin\theta_{\mathrm{i}} = \frac{\Delta y_{\mathrm{i}}}{\sqrt{\Delta x_{\mathrm{i}}^2 + \Delta y_{\mathrm{i}}^2}} \quad \text{and} \quad \sin\theta_{\mathrm{t}} = \frac{\Delta y_{\mathrm{t}}}{\sqrt{\Delta x_{\mathrm{t}}^2 + \Delta y_{\mathrm{t}}^2}} \tag{9.21}$$

When these are substituted into (9.20) we obtain

$$n_{\mathrm{i}}\sin\theta_{\mathrm{i}} = n_{\mathrm{t}}\sin\theta_{\mathrm{t}} \tag{9.22}$$

which is the familiar Snell's law.

An imaging situation occurs when many paths from point A to point B have the same optical path length. An example of this occurs when a lens causes an image to form. In this case all rays leaving point A (on an object) and traveling through the system to point B (on the image) experience equal optical path lengths. This situation is depicted in Fig. 9.4. Note that while the rays traveling through the center of the lens have a shorter geometric path length, they travel through more material so that the optical path length is the same for all rays.

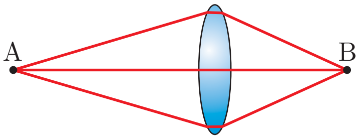


Figure 9.4 Rays of light leaving point A with the same optical path length to B.

Example 9.3

Use Fermat's principle to derive the equation of curvature for a reflective surface that causes all rays leaving one point to image to another. Do the calculation in two dimensions rather than in three. This configuration is used in laser heads to direct flash lamp energy into the amplifying material. One "point" represents the end of a long cylindrical laser rod and the other represents the end of a long flash lamp.

Solution: We adopt the convention that the origin is half way between the points, which are separated by a distance $2a$, as shown in Fig. 9.5.

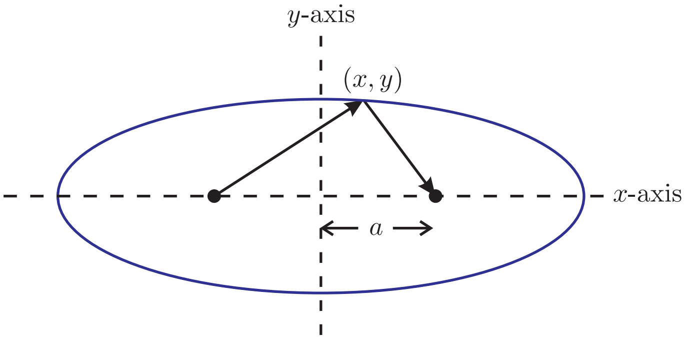


Figure 9.5

If the points are to image to each other, Fermat's principle requires that the total path length be a constant, say b. By inspection of the figure, we obtain an equation describing the curvature of the reflective surface

$$b = \sqrt{(x+a)^2 + y^2} + \sqrt{(x-a)^2 + y^2} \tag{9.23}$$

To get (9.23) into a more recognizable form, we isolate the first square root

$$\sqrt{(x+a)^2 + y^2} = b - \sqrt{(x-a)^2 + y^2},$$

square both sides of the equation

$$(x+a)^2 + y^2 = b^2 + (x-a)^2 + y^2 - 2b\sqrt{(x-a)^2 + y^2},$$

and then carry out the square of two of the binomial terms

$$\left(x^2 + a^2 + 2ax\right) + y^2 = b^2 + \left(x^2 + a^2 - 2ax\right) + y^2 - 2b\sqrt{(x-a)^2 + y^2}.$$

Some nice cancelation occurs, and we gather the remaining non-square-rooted terms on the left

$$4ax - b^2 = -2b\sqrt{(x-a)^2 + y^2}.$$

We square both sides of the equation and carry out the square of the remaining binomial term to obtain

$$16a^2x^2 - 4ab^2x + b^4 = 4b^2\left(x^2 - 2ax + a^2 + y^2\right),$$

and then cancel and regroup terms to arrive at

$$\left(16a^2 - 4b^2\right)x^2 - 4b^2y = 4a^2b^2 - b^4.$$

Finally, we divide both sides of the equation by the term on the right to obtain the (hopefully) familiar form of an ellipse

$$\frac{x^2}{\left(\frac{b^2}{4}\right)} + \frac{y^2}{\left(\frac{b^2}{4} - a^2\right)} = 1$$

9.4 Paraxial Rays and ABCD Matrices

In the remainder of this chapter we develop a formalism for describing the effects of mirrors and lenses on rays of light. Keep in mind that when describing light as a collection of rays rather than as waves, the results can only describe features that are macroscopic compared to a wavelength. The rays of light at each location in space describe approximately the direction of travel of the wave fronts at that location. Since the wavelength of visible light is extraordinarily small compared to the macroscopic features that we perceive in our day-to-day world, the ray approximation is often a very good one. This is the reason that ray optics was developed long before light was understood as a wave.

We consider ray theory within the *paraxial approximation*, meaning that we restrict our attention to rays that are near and almost parallel to an *optical axis* of a system, say the z-axis. It is within this approximation that the familiar imaging properties of lenses occur. An image occurs when all rays from a *point* on an *object* converge to a corresponding *point* on what is referred to as the *image*. To the extent that the paraxial approximation is violated, the clarity of an image can suffer, and we say that there are *aberrations* present. Very often in the field of optical engineering, one is primarily concerned with minimizing aberrations in cases where the paraxial approximation is not strictly followed. This is done so that, for example, a camera can take pictures of subjects that occupy a fairly wide angular field of view, where rays violate the paraxial approximation. Optical systems are typically engineered using the science of *ray tracing*, which is described briefly in section 9.9.

As we develop paraxial ray theory, we should remember that rays impinging on devices such as lenses or curved mirrors should strike the optical component at near normal incidence. To quantify this statement, the paraxial approximation is valid to the extent that we have

$$\sin\theta \cong \theta \tag{9.24}$$

and similarly

$$\tan\theta \cong \theta \tag{9.25}$$

Here, the angle θ (in radians) represents the angle that a particular ray makes with respect to the optical axis. There is an important mathematical reason for this approximation.

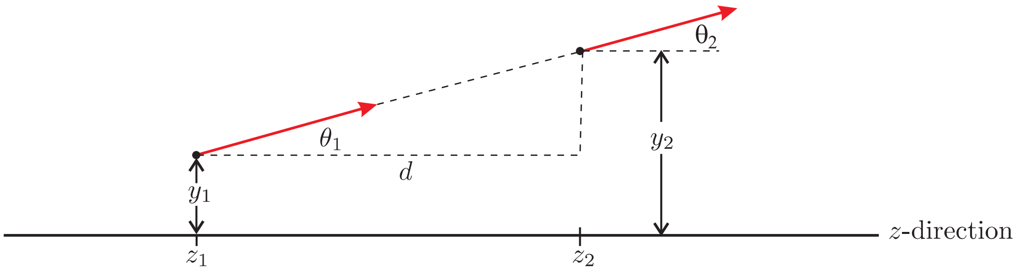


Figure 9.6 The behavior of a ray as light traverses a distance d.

The sine is a nonlinear function, but at small angles it is approximately linear and can be represented by its argument. It is this linearity that is crucial to the process of forming images. The linearity also greatly simplifies the formulation since it reduces the problem to linear algebra. Conveniently, we will be able to keep track of imaging effects with a 2×2 matrix formalism.

Consider a ray confined to the y–z plane where the optical axis is in the z-direction. Let us specify a ray at position z_1 by two coordinates: the displacement from the axis y_1 and the orientation angle θ_1 (see Fig. 9.6). The ray continues along a straight path as it travels through a uniform medium. This makes it possible to predict the coordinates of the same ray at other positions, say at z_2. The connection is straightforward. First, since the ray continues in the same direction, we have

$$\theta_2 = \theta_1 \tag{9.26}$$

By referring to Fig. 9.6 we can write y_2 in terms of y_1 and θ_1:

$$y_2 = y_1 + d \tan \theta_1 \tag{9.27}$$

where $d \equiv z_2 - z_1$. Equation (9.27) is nonlinear in θ_1. However, in the paraxial approximation (9.25) it becomes linear, which after all is the point of the approximation. In this approximation the expression for y_2 becomes

$$y_2 = y_1 + d\theta_1 \tag{9.28}$$

Equations (9.26) and (9.28) describe a linear transformation which in matrix notation can be consolidated into the form

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \quad \text{(propagation through a distance } d\text{)} \tag{9.29}$$

Here, the vectors in this equation specify the essential information about the ray before and after traversing the distance d, and the matrix describes the effect of traversing the distance. This type of matrix is called an ABCD matrix.

Suppose that the distance d is subdivided into two distances, a and b, such that $d = a+b$. If we consider individually the effects of propagation through a and through b, we have

$$\begin{aligned} \begin{bmatrix} y_{\text{mid}} \\ \theta_{\text{mid}} \end{bmatrix} &= \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \\ \begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} &= \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{\text{mid}} \\ \theta_{\text{mid}} \end{bmatrix} \end{aligned} \tag{9.30}$$

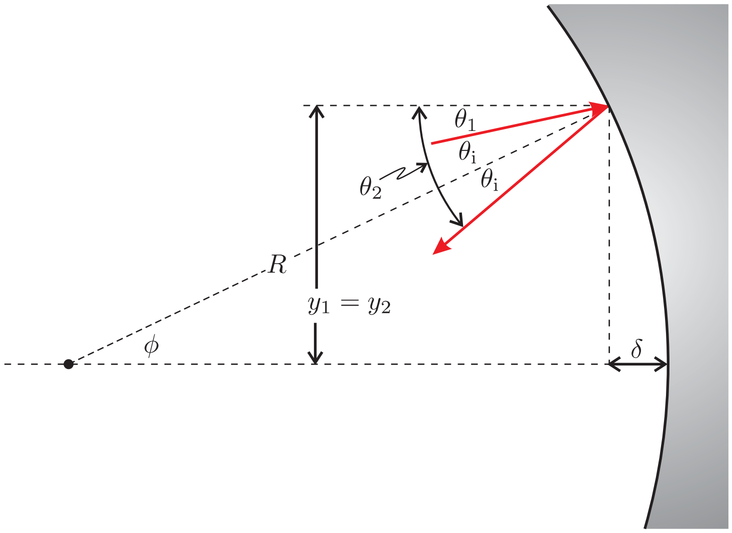


Figure 9.7 A ray depicted in the act of reflection from a curved surface.

where the subscript "mid" refers to the ray in the middle position after traversing the distance a. If we combine the equations, we get

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \tag{9.31}$$

which is in complete agreement with (9.29) since the ABCD matrix for the entire displacement is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & a+b \\ 0 & 1 \end{bmatrix} \tag{9.32}$$

9.5 Reflection and Refraction at Curved Surfaces

We next consider the effect of reflection from a spherical surface as depicted in Fig. 9.7. We consider only the act of reflection without considering propagation before or after the reflection takes place. Thus, the incident and reflected rays in the figure are symbolic only of the direction of propagation before and after reflection; they do not indicate any amount of travel. Upon reflection we have

$$y_2 = y_1 \tag{9.33}$$

since the ray has no chance to go anywhere.

We adopt the widely used convention that, upon reflection, the positive z-direction is reoriented so that we consider the rays still to travel in the positive z sense. Notice that in Fig. 9.7, the reflected ray approaches the z-axis. In this case θ_2 is a negative angle (as opposed to θ_1 which is drawn as a positive angle) and is equal to

$$\theta_2 = -(\theta_1 + 2\theta_i) \tag{9.34}$$

where θ_i is the angle of incidence with respect to the normal to the spherical mirror surface. By the law of reflection, the reflected ray also occurs at an angle θ_i referenced to the surface normal. The surface normal points towards the center of curvature, which we assume is on the z-axis a distance R away. By convention, the radius of curvature R is a positive number if the mirror surface is *concave* and a negative number if the mirror surface is *convex.*

We must eliminate θ_i from (9.34) in favor of θ_1 and y_1. By inspection of Fig. 9.7 we can write

$$\frac{y_1}{R} = \sin\phi \cong \phi \tag{9.35}$$

where we have applied the paraxial approximation (9.24). (Note that the angles in the figure are exaggerated.) We also have

$$\phi = \theta_1 + \theta_\mathrm{i} \tag{9.36}$$

and when this is combined with (9.35), we get

$$\theta_\mathrm{i} = \frac{y_1}{R} - \theta_1 \tag{9.37}$$

With this we are able to put (9.34) into a useful linear form:

$$\theta_2 = -\frac{2}{R}y_1 + \theta_1 \tag{9.38}$$

Equations (9.33) and (9.38) describe a linear transformation that can be concisely formulated as

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2/R & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \quad \text{(concave mirror)} \tag{9.39}$$

The ABCD matrix in this transformation describes the act of reflection from a concave mirror with radius of curvature R. The radius R is negative when the mirror is convex.

The final basic element that we shall consider is a spherical interface between two materials with indices n_i and n_t (see Fig. 9.8). This has an effect similar to that of the curved mirror, which changes the direction of a ray without altering its distance y_1 from the optical axis. Please note that here the radius of curvature is considered to be positive for a convex surface (opposite convention from that of the mirror). Again, we are interested only in the act of transmission without any travel before or after the interface. As before, (9.33) applies (i.e. $y_2 = y_1$).

To connect θ_1 and θ_2 we must use Snell's law which in the paraxial approximations is

$$n_\mathrm{i}\theta_\mathrm{i} = n_\mathrm{t}\theta_\mathrm{t} \tag{9.40}$$

As seen in the Fig. 9.8, we have

$$\theta_\mathrm{i} = \theta_1 + \phi \tag{9.41}$$

and

$$\theta_\mathrm{t} = \theta_2 + \phi \tag{9.42}$$

As before, (9.35) applies (i.e. $\phi \cong y_1/R$). When this is used in (9.41) and (9.42), Snell's law (9.40) becomes

$$\theta_2 = \left(\frac{n_\mathrm{i}}{n_\mathrm{t}} - 1\right)\frac{y_1}{R} + \frac{n_\mathrm{i}}{n_\mathrm{t}}\theta_1 \tag{9.43}$$

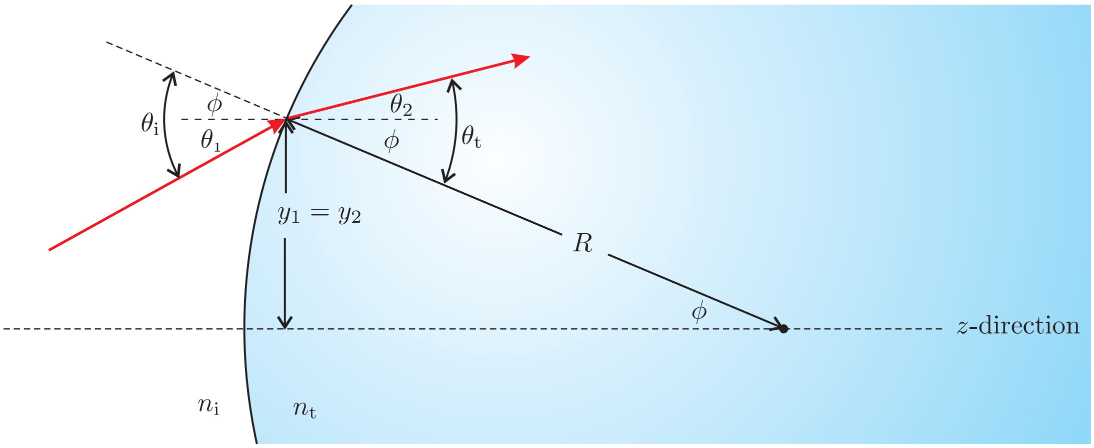


Figure 9.8 A ray depicted in the act of transmission at a curved material interface.

The compact matrix form of (9.33) and (9.43) turns out to be

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \left(n_i/n_t - 1\right)/R & n_i/n_t \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \text{ (from } n_i \text{ to } n_t\text{; interface radius } R\text{)} \tag{9.44}$$

In summary, we have developed three basic ABCD matrices seen in (9.29), (9.39), and (9.44). All other ABCD matrices that we will use are composites of these three. For example, one can construct the ABCD matrix for a lens by using two matrices like those in (9.44) to represent the entering and exiting surfaces of the lens. A distance matrix (9.29) can be inserted to account for the thickness of the lens. It is left as an exercise to derive the ABCD matrix for such a thick lens (see P 9.6).

The three ABCD matrices discussed can be used for many different composite systems. As another example, consider a ray that propagates through a distance a, followed by a reflection from a mirror of radius R, and then propagates through a distance b. This example is depicted in Fig. 9.9. The vector depicting the final ray in terms of the initial one is computed as follows:

$$\begin{aligned} \begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} &= \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2/R & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \\ &= \begin{bmatrix} 1 - 2b/R & a + b - 2ab/R \\ -2/R & 1 - 2a/R \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \end{aligned} \tag{9.45}$$

The ordering of the matrices is important. The first effect that the light experiences is the matrix to the right, in the position that first operates on the vector representing the initial ray.

We have continually worked within the y–z plane as indicated in Figs. 9.6–9.9. This may have given the impression that it is necessary to work within that plane, or a plane containing the z-axis. However, within the paraxial approximation, our ABCD matrices are still valid for rays contained in planes that do not include the optical axis (as long as the rays are nearly parallel to the optical axis.

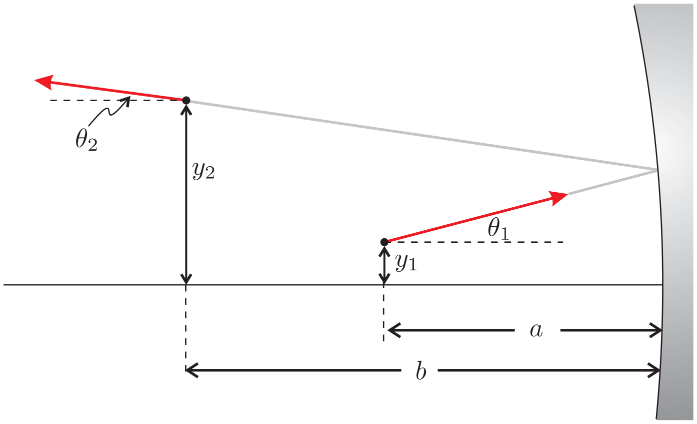


Figure 9.9 A ray that travels through a distance a, reflects from a mirror, and then travels through a distance b.

Imagine a ray contained within a plane that is parallel to the y–z plane but for which $x > 0$. One might be concerned that when the ray meets, for example, a spherically concave mirror, the radius of curvature *in the perspective of the* y–z *dimension* might be different for $x > 0$ than for $x = 0$ (at the center of the mirror). This concern is actually quite legitimate and is the source of what is known as spherical aberration. Nevertheless, in the paraxial approximation the intersection with the curved mirror of all planes that are parallel to the optical axis always give the same curvature.

To see why this is so, consider the curvature of the mirror in Fig. 9.7. As we move away from the mirror center (in either the x or y-dimension or some combination thereof), the mirror surface deviates to the left by the amount

$$\delta = R - R\cos\phi \tag{9.46}$$

In the paraxial approximation, we have $\cos\phi \cong 1 - \phi^2/2$. And since in this approximation we may also write $\phi \cong \sqrt{x^2 + y^2}\Big/ R$, (9.46) becomes

$$\delta \cong \frac{x^2 + y^2}{2R} \tag{9.47}$$

In the paraxial approximation, we see that the curve of the mirror is parabolic, and therefore separable between the x and y dimensions. That is, the curvature in the x-dimension (i.e. $\partial\delta/\partial x = x/R$) is independent of y, and the curvature in the y-dimension (i.e. $\partial\delta/\partial y = y/R$) is independent of x. A similar argument can be made for a spherical interface between two media within the paraxial approximation.

This allows us to deal conveniently with rays that have positioning and directional components in both the x and y dimensions. Each dimension can be treated separately without influencing the other. Most importantly, the identical matrices, (9.29), (9.39), and (9.44), are used for either dimension. Figs. 9.6–9.9 therefore represent projections of the actual rays onto the y–z plane. To complete the story, one would also need corresponding figures representing the projection of the rays onto the x–z plane.

9.6 Image Formation by Mirrors and Lenses

Consider the example shown in Fig. 9.9 where a ray travels through a distance a, reflects from a curved mirror, and then travels through a distance b. From (9.45) we know that the ABCD matrix for the overall process is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 - 2b/R & a + b - 2ab/R \\ -2/R & 1 - 2a/R \end{bmatrix} \tag{9.48}$$

As is well known, it is possible to form an image with a concave mirror. Suppose that the initial ray is one of many which leaves a point on an object positioned at $d_\text{o} = a$ before the mirror. In order for an image to occur at $d_\text{i} = b$, it is essential that all rays leaving the original point on the object converge to a single point on the image. That is, we want rays leaving the point y_1 on the object (which may take on a range of angles θ_1) all to converge to a single point y_2 at the image. In the following equation we need y_2 to be independent of θ_1:

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} Ay_1 + B\theta_1 \\ Cy_1 + D\theta_1 \end{bmatrix} \tag{9.49}$$

The condition for image formation is therefore

$$B = 0 \quad \text{(condition for image formation)} \tag{9.50}$$

When this condition is applied to (9.48), we obtain

$$d_\text{o} + d_\text{i} - \frac{2d_\text{o}d_\text{i}}{R} = 0 \Rightarrow \frac{2}{R} = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} \tag{9.51}$$

which is the familiar imaging formula for a mirror, in agreement with (9.1). When the object is infinitely far away (i.e. $d_\text{o} \to \infty$), the image appears at $d_\text{i} \to R/2$. This distance is called the *focal length* and is denoted by

$$f = \frac{R}{2} \quad \text{(focal length of a mirror)} \tag{9.52}$$

Please note that d_o and d_i can each be either positive (*real* as depicted in Fig. 9.9) or negative (*virtual* or behind the mirror).

The magnification of the image is found by comparing the size of y_2 to y_1. From (9.48)–(9.51), the magnification is found to be

$$M \equiv \frac{y_2}{y_1} = A = 1 - \frac{2d_\text{i}}{R} = -\frac{d_\text{i}}{d_\text{o}} \tag{9.53}$$

The negative sign indicates that for positive distances d_o and d_i the image is inverted.

Another common and very useful example is that of a thin lens, where we ignore the thickness between the two surfaces of the lens. Using the ABCD matrix in (9.44) twice, we find the overall matrix for the thin lens is

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ \frac{1}{R_2}(n-1) & n \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{R_1}\left(\frac{1}{n} - 1\right) & \frac{1}{n} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -(n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) & 1 \end{bmatrix} \quad \text{(Thin Lens)} \end{aligned} \tag{9.54}$$

Galileo Galilei

(1564–1642, Italian)

While Galileo did not invent the telescope, he was one of the few people of his time who knew how to build one. He also constructed a compound microscope. He attempted to measure the speed of light by having his assistant position himself on a distant hill and measuring the time it took for his assistant to uncover a lantern in response to a light signal. He was, of course, unable to determine the speed of light. His conclusion was that light is "really fast" if not instantaneous.

where we have taken the index outside of the lens to be unity while that of the lens material to be n. R_1 is the radius of curvature for the first surface which is positive if convex, and R_2 is the radius of curvature for the second surface which is also positive if convex *from the perspective of the rays which encounter it.*

Notice the close similarity between (9.54) and the matrix in (9.39). The ABCD matrix for either a thin lens or a mirror can be written as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \tag{9.55}$$

where in the case of the thin lens the focal length is given by the lens maker's formula

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \quad \text{(focal length of thin lens)} \tag{9.56}$$

All of the arguments about image formation given above for the curved mirror work equally well for the thin lens. The only difference is that the focal length (9.56) is used in place of (9.52). That is, if we consider a ray traveling though a distance d_o impinging on a thin lens whose matrix is given by (9.55), and then afterwards traveling a distance d_i, the overall ABCD matrix is exactly like that in (9.48):

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 - d_\text{i}/f & d_\text{o} + d_\text{i} - d_\text{o}d_\text{i}/f \\ -1/f & 1 - d_\text{o}/f \end{bmatrix} \tag{9.57}$$

When we use the imaging condition (9.50), the imaging formula (9.1) emerges naturally.

9.7 Image Formation by Complex Optical Systems

A complicated series of optical elements (e.g. a sequence of lenses and spaces) can be combined to form a composite imaging system. The matrices for each of the elements are multiplied together (the first element that rays encounter appearing on the right) to form

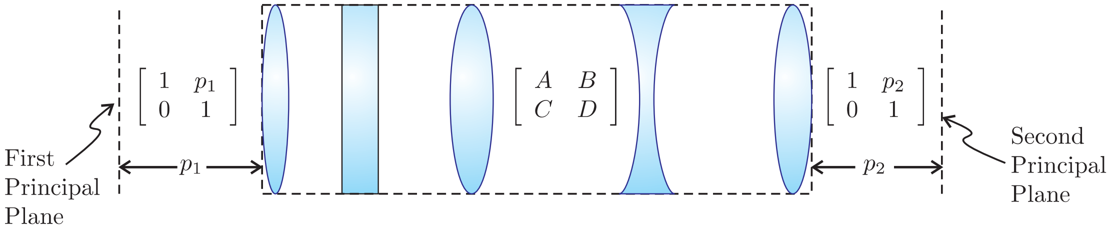

$\begin{bmatrix} 1 & p_1 \\ 0 & 1 \end{bmatrix}$
$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$
$\begin{bmatrix} 1 & p_2 \\ 0 & 1 \end{bmatrix}$
First Principal Plane
p_1
p_2
Second Principal Plane

Figure 9.10 A multi-element system represented as an ABCD matrix for which principal planes always exist.

the overall composite ABCD matrix. We can study the imaging properties of a composite ABCD matrix by combining the matrix with the matrices for the distances from an object to the system and from the system to the image formed:

$$\begin{aligned}\begin{bmatrix} 1 & d_\text{i} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 & d_\text{o} \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} A + d_\text{i}C & d_\text{o}A + B + d_\text{o}d_\text{i}C + d_\text{i}D \\ C & d_\text{o}C + D \end{bmatrix} \\ &= \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix}\end{aligned} \tag{9.58}$$

Imaging occurs according to (9.50) when $B' = 0$, or

$$d_\text{o}A + B + d_\text{o}d_\text{i}C + d_\text{i}D = 0, \quad \text{(general condition for image formation)} \tag{9.59}$$

with magnification

$$M = A + d_\text{i}C \tag{9.60}$$

There is a convenient way to simplify this analysis.

For every ABCD matrix representing a (potentially) complicated optical system, there exist two principal planes located (in our convention) a distance p_1 before entering the system and a distance p_2 after exiting the system. When the matrices corresponding to the (appropriately chosen) distances to those planes are appended to the original ABCD matrix of the system, the overall matrix simplifies to one that looks like the matrix for a simple thin lens (9.55). With knowledge of the positions of the principal planes, one can treat the complicated imaging system in the same way that one treats a simple thin lens. The only difference is that d_o is the distance from the object to the first principal plane and d_i is the distance from the second principal plane to the image. (In the case of an actual thin lens, both principal planes are at $p_1 = p_2 = 0$. For a composite system, p_1 and p_2 can be either positive or negative.)

Next we demonstrate that p_1 and p_2 can always be selected such that we can write

$$\begin{aligned}\begin{bmatrix} 1 & p_2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 & p_1 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} A + p_2C & p_1A + B + p_1p_2C + p_2D \\ C & p_1C + D \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -1/f_\text{eff} & 1 \end{bmatrix}\end{aligned} \tag{9.61}$$

The final matrix is that of a simple thin lens, and it takes the place of the composite system including the distances to the principal planes. Our task is to find the values of p_1 and

p_2 that make this matrix replacement work. We must also prove that this replacement is always possible for physically realistic values for A, B, C, and D.

We can straightaway make the definition

$$f_{\text{eff}} \equiv -1/C \tag{9.62}$$

We can also solve for p_1 and p_2 by setting the diagonal elements of the matrix to 1. Explicitly, we get

$$p_1 C + D = 1 \Rightarrow p_1 = \frac{1-D}{C} \tag{9.63}$$

and

$$A + p_2 C = 1 \Rightarrow p_2 = \frac{1-A}{C} \tag{9.64}$$

It remains to be shown that the upper right element in (9.61) (i.e. $p_1 A + B + p_1 p_2 C + p_2 D$) automatically goes to zero for our choices of p_1 and p_2. This may seem unlikely at first, but we can invoke an important symmetry in the matrix to show that it does in fact vanish for our choices of p_1 and p_2.

When (9.63) and (9.64) are substituted into the upper right matrix element of (9.61) we get

$$\begin{aligned} p_1 A + B + p_1 p_2 C + p_2 D &= \frac{1-D}{C} A + B + \frac{1-D}{C}\frac{1-A}{C} C + \frac{1-A}{C} D \\ &= \frac{1}{C}\left[1 - AD + BC\right] \\ &= \frac{1}{C}\left(1 - \begin{vmatrix} A & B \\ C & D \end{vmatrix}\right) \end{aligned} \tag{9.65}$$

This equation shows that the upper right element of (9.61) vanishes when the determinant of the original ABCD matrix equals one. Fortunately, this is always the case *as long as we begin and end in the same index of refraction.* Therefore, we have

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = 1 \tag{9.66}$$

Notice that the determinants of the matrices in (9.29), (9.39), and (9.55) are all one, and so ABCD matrices constructed of these will also have determinants equal to one. The determinant of (9.44) is not one. This is because it begins and ends in different indices, but when this matrix is used in succession to form a lens or even a strange conglomerate of successive material interfaces, the resulting matrix will have a determinant equal to one as long as the beginning and ending indices are the same. Table 9.1 is a summary of ABCD matrices of common optical elements. All of the matrices obey (9.66).

9.8 Stability of Laser Cavities

As a final example of the usefulness of paraxial ray theory, we apply the ABCD matrix formulation to a laser cavity. The basic elements of a laser cavity include an amplifying medium and mirrors to provide feedback. Presumably, at least one of the end mirrors is

$\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$	(Distance within any material, excluding interfaces)
$\begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix}$	(Window, starting and stopping in air)
$\begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}$	(Thin lens or a mirror with $f = R/2$)
$\begin{bmatrix} 1 + \frac{d}{R_1}\left(\frac{1}{n} - 1\right) & \frac{d}{n} \\ (1-n)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{d}{R_1 R_2}\left(2 - \frac{1}{n} - n\right) & 1 - \frac{d}{R_2}\left(\frac{1}{n} - 1\right) \end{bmatrix}$	(Thick lens)

Table 9.1 Summary of ABCD matrices for common optical elements.

partially transmitting so that energy is continuously extracted from the cavity. Here, we dispense with the amplifying medium and concentrate our attention on the optics providing the feedback. As might be expected, the mirrors must be carefully aligned or successive reflections might cause rays to "walk" continuously away from the optical axis, so that they eventually leave the cavity out the side. If a simple cavity is formed with two flat mirrors that are perfectly aligned parallel to each other, one might suppose that the mirrors would provide ideal feedback. However, all rays except for those that are perfectly aligned to the mirror surface normals eventually wander out of the side of the cavity as illustrated in Fig. 9.11a. Such a cavity is said to be *unstable.* We would like to do a better job of trapping the light in the cavity.

To improve the situation, a cavity can be constructed with concave end mirrors to help confine the beams within the cavity. Even so, one must choose carefully the curvature of the mirrors and their separation L. If this is not done correctly, the curved mirrors can "overcompensate" for the tendency of the rays to wander out of the cavity and thus aggravate the problem. Such an unstable scenario is depicted in Fig. 9.11b.

Figure 9.11c depicts a cavity made with curved mirrors where the separation L is chosen appropriately to make the cavity stable. Although a ray, as it makes successive bounces, can strike the end mirrors at a variety of points, the curvature of the mirrors keeps the "trajectories" contained within a narrow region so that they cannot escape out the sides of the cavity.

There are many ways to make a stable laser cavity. For example, a stable cavity can be made using a lens between two flat end mirrors as shown in Fig. 9.11d. Any combination of lenses (perhaps more than one) and curved mirrors can be used to create stable cavity configurations. *Ring cavities* can also be made to be stable where in no place do the rays retro-reflect from a mirror but circulate through a series of elements like cars going around a racetrack.

We now find the conditions that have to be met in order for a cavity to be stable. The

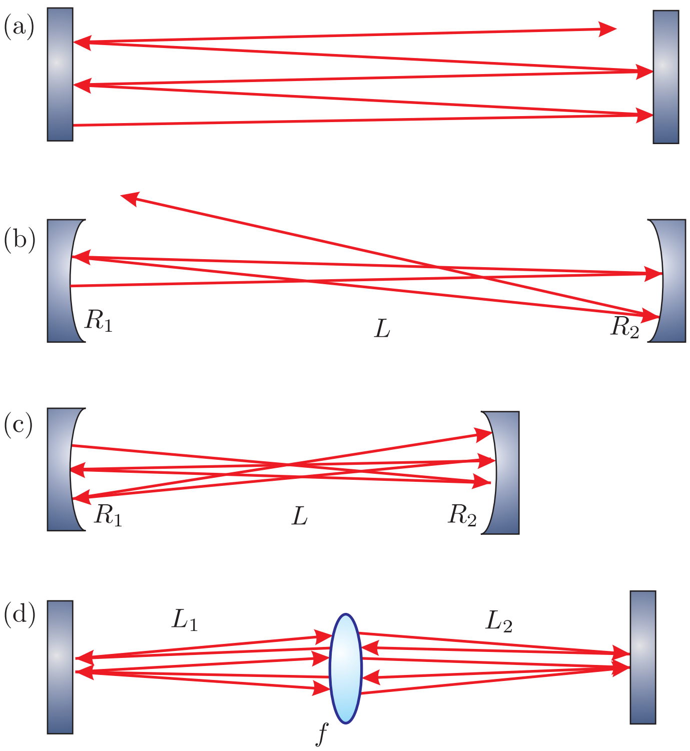


Figure 9.11 (a) A ray bouncing between two parallel flat mirrors. (b) A ray bouncing between two curved mirrors in an unstable configuration. (c) A ray bouncing between two curved mirrors in a stable configuration. (d) Stable cavity utilizing a lens and two flat end mirrors.

ABCD matrix for a round trip in the cavity is useful for this analysis. For example, the round-trip ABCD matrix for the cavity shown in Fig. 9.11c is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2/R_2 & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2/R_1 & 1 \end{bmatrix} \tag{9.67}$$

where we have begun the round trip just after a reflection from the first mirror. The round-trip ABCD matrix for the cavity shown in Fig. 9.11d is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 2L_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \begin{bmatrix} 1 & 2L_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \tag{9.68}$$

where we have begun the round trip just after a transmission through the lens moving to the right. It is somewhat arbitrary where the round trip begins.

To determine whether a given configuration of a cavity will be stable, we need to know what a ray does after making many round trips in the cavity. To find the effect of propagation through many round trips, we multiply the round-trip ABCD matrix together N times, where N is the number of round trips that we wish to consider. We can then examine what happens to an arbitrary ray after making N round trips in the cavity as follows:

$$\begin{bmatrix} y_{N+1} \\ \theta_{N+1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^N \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \tag{9.69}$$

At this point students might be concerned that taking an ABCD matrix to the N^{th} power can be a lot of work. (It is already a significant amount of work just to compute the ABCD matrix for a single round trip.) In addition, we are interested in letting N be very large, perhaps even infinity. Students can relax because we have a neat trick to accomplish this daunting task.

We use Sylvester's theorem from appendix 0.4, which states that if

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = 1 \tag{9.70}$$

then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^N = \frac{1}{\sin\theta} \begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix} \tag{9.71}$$

where

$$\cos\theta = \frac{1}{2}(A+D). \tag{9.72}$$

As we have already discussed, (9.70) is satisfied if the refractive index is the same before and after, which is guaranteed for any round trip. We therefore can employ Sylvester's theorem for any N that we might choose, including very large integers.

We would like the elements of (9.71) to remain finite as N becomes very large. If this is the case, then we know that a ray remains trapped within the cavity and stays reasonably close to the optical axis. Since N only appears within the argument of a sine function, which is always bounded between -1 and 1 for real arguments, it might seem that the elements of (9.71) always remain finite as N approaches infinity. However, it turns out that θ can become imaginary depending on the outcome of (9.72), in which case the sine becomes a hyperbolic sine, which can "blow up" as N becomes large. In the end, the condition for cavity stability is that a real θ must exist for (9.72), or in other words we need

$$-1 < \frac{1}{2}(A + D) < 1 \qquad \text{(condition for a stable cavity)} \tag{9.73}$$

It is left as an exercise to apply this condition to (9.67) and (9.68) to find the necessary relationships between the various element curvatures and spacing in order to achieve cavity stability.

9.9 Aberrations and Ray Tracing

The paraxial approximation places serious limitations on the performance of optical systems (see (9.24) and (9.25)). To stay within the approximation, all rays traveling in the system should travel very close to the optic axis with very shallow angles with respect to the optical axis. To the extent that this is not the case, the collection of rays associated with a single point on an object may not converge to a single point on the associated image. The resulting distortion or "blurring" of the image is known as *aberration*.

Common experience with photographic and video equipment suggests that it is possible to image scenes that have a relatively wide angular extent (many tens of degrees), in apparent serious violation of the paraxial approximation. The paraxial approximation is indeed violated in these devices, so they must be designed using more complicated analysis techniques than those we have learned in this chapter. The most common approach is to use a computationally intensive procedure called *ray tracing* in which $\sin\theta$ and $\tan\theta$ are rendered exactly. The nonlinearity of these functions precludes the possibility of obtaining analytic solutions describing the imaging performance of such optical systems.

The typical procedure is to start with a collection of rays from a test point such as shown in Fig. 9.12. Each ray is individually traced through the system using the exact representation of geometric surfaces as well as the exact representation of Snell's law. On close analysis, the rays typically do not converge to a distinct imaging point. Rather, the rays can be "blurred" out over a range of points where the image is supposed to occur. Depending on the angular distribution of the rays as well as on the elements in the setup, the spread of rays around the image point can be large or small. The engineer who designs the

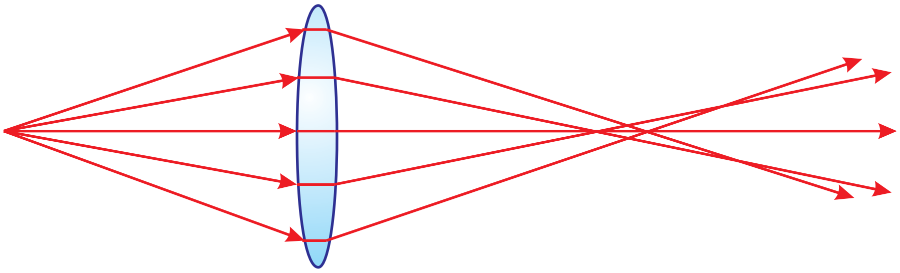


Figure 9.12 Ray tracing through a simple lens.

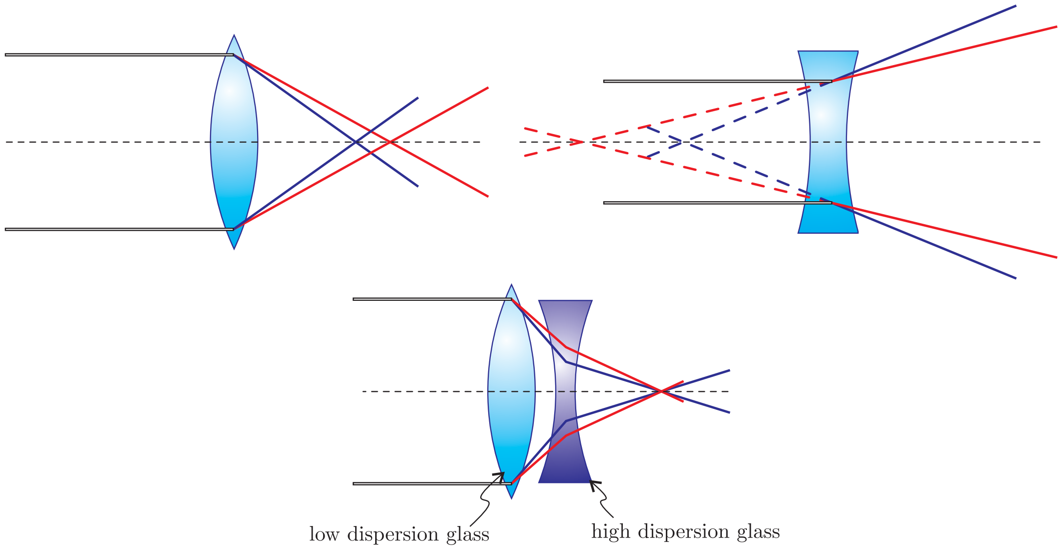


Figure 9.13 Chromatic abberation causes lenses to have different focal lengths for different wavelengths. It can be corrected using an achromatic doublet lens.

system must determine whether the amount of aberration is acceptable, given the various constraints of the device.

To minimize aberrations below typical tolerance levels, several lenses can be used together. If properly chosen, the lenses (some positive, some negative) separated by specific distances, can result in remarkably low aberration levels over certain ranges of operation for the device. Ray tracing is best done with commercial software designed for this purpose (e.g. Zemax or other professional products). Such software packages are able to develop and optimize designs for specific applications. A nice feature is that the user can specify that the design should employ only standard optical components available from known optics companies. In any case, it is typical to specify that all lenses in the system should have spherical surfaces since these are much less expensive to manufacture. We mention briefly a few types of aberrations that you may encounter. Multiple aberrations can often be observed in a single lens.

Chromatic abberation arises from the fact that the index of refraction for glass varies with the wavelength of light. Since the focal length of a lens depends on the index of refraction (see, for example, Eq. (9.56)), the focal length of a lens varies with the wavelength of light. Chromatic abberation can be compensated for by using a pair of lenses made from two types of glass as shown in Fig. 9.13 (the pair is usually cemented together to form a "doublet" lens). The lens with the shortest focal length is made of the glass whose index has the lesser dependence on wavelength. By properly choosing the prescription of the two lenses, you can exactly compensate for chromatic abberation at two wavelengths and do a good job for a wide range of others. Achromatic doublets can also be designed to minimize spherical abberation (see below), so they are often a good choice when you need a high quality lens.

Monochromatic abberations arise from the shape of the lens rather than the variation of n with wavelength. Before the advent computers facilitated the widespread use of ray tracing, these abberations had to be analyzed primarily with analytic techniques. The

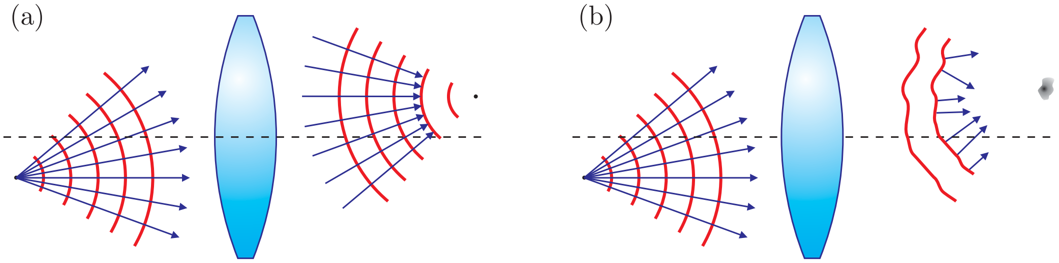


Figure 9.14 (a) Paraxial theory predicts that the light imaged from a point source will converge to a point (i.e. have spherical wave fronts coming to the image point). (b) The image of a point source made by a real lens is an extended and blurred patch of light and the converging wavefronts are only quasi-spherical.

analytic results derived previously in this chapter were based on first order approximations (e.g. $\sin\theta \approx \theta$). This analysis predicts that a lens can image a point source to an exact image point, which predicts spherically converging wavefronts at the image point as shown in Fig. 9.14(a). You can increase the accuracy of the theory for non-paraxial rays by retaining second-order correction terms in the analysis. With these second-order terms included, the wave fronts converging towards an image point are mostly spherical, but have second-order abberation terms added in (shown conceptually in Fig. 9.14(b)). There are five abberation terms in this second-order analysis, and these represent a convenient basis for discussing abberation.

The first abberation term is known as *spherical abberation.* This type of abberation results from the fact that rays traveling through a spherical lens at large radii experience a different focal length than those traveling near the axis. For a converging lens, this causes wide-radius rays to focus before the near-axis rays as shown in Fig. 9.15. This problem can be helped by orienting lenses so that the face with the least curvature is pointed towards the side where the light rays have the largest angle. This procedure splits the bending of rays more evenly between the front and back surface of the lens. As mentioned above, you can also cement two lenses made from different types of glass together so that spherical abberations from one lens are corrected by the other.

The abberation term referred to as *astigmatism* occurs when an off-axis object point

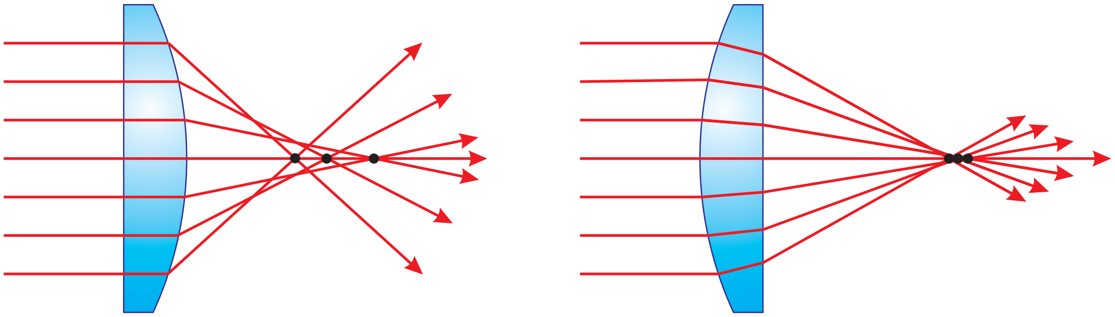


Figure 9.15 Spherical abberation in a plano-convex lens.

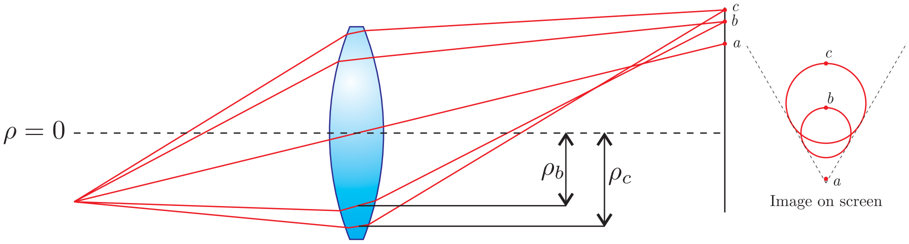


Figure 9.16 Illustration of coma. Rays traveling through the center of the lens are imaged to point a as predicted by paraxial theory. Rays that travel through the lens at radius ρ_b in the plane of the figure are imaged to point b. Rays that travel through the lens at radius ρ_b, but outside the plane of the figure are imaged to other points on the circle (in the image plane) containing point b. Rays at that travel through the lens at other radii on the lens (e.g. ρ_{c}) also form circles in the image plane with radius proportional to ρ^2 with the center offset from point a a distance proportional to ρ^2. When light from each of these circles combines on the screen it produces an imaged point with a "comet tail."

is imaged to an off-axis image point. In this case a spherical lens has a different focal length in the horizontal and vertical dimensions. For a focusing lens this causes the two dimensions to focus at different distances, producing a vertical line at one image plane and a horizontal line at another. A lens can also be inherently astigmatic even when viewed on axis if it is football shaped rather than spherical. In this case, the astigmatic abberation can be corrected by inserting a cylindrical lens at the correct orientation (this is a common correction needed in eyeglasses).

A third abberation term is referred to as *coma*. This is observed when off-axis points are imaged and produces a comet shaped tail with its head at the point predicted by paraxial theory. (The term "coma" refers to the atmosphere of a comet, which is how the abberation got its name.) This abberation is distinct from astigmatism, which is also observed for off-axis points, since coma is observed even when all of the rays are in one plane (see Fig. 9.16). You have probably seen coma if you've ever played with a magnifying glass in the sun—just tilt the lens slightly and you see a comet-like image rather than a point.

The *curvature of the field* abberation term arises from the fact that spherical lenses image spherical surfaces to another spherical surface, rather than imaging a plane to a plane. This is not so bad for your eyeball, which has a curved screen, but for things like cameras and movie projectors we would like to image to a flat screen. When a flat screen is used and the curvature of the field abberation is present, the image will be focus well near the center, but become progressively out of focus as you move to the edge of the screen (i.e. the flat screen is further from the curved image surface as you move from the center).

The final abberation term is referred to as *distortion*. This abberation occurs when the magnification of a lens depends on the distance from the center of the screen. If magnification decreases as the distance from the center increases, then "barrel" distortion is observed. When magnification increases with distance, "pincushion" distortion is observed (see Fig. 9.17).

All lenses will exhibit some combination of the abberations listed above (i.e. chromatic

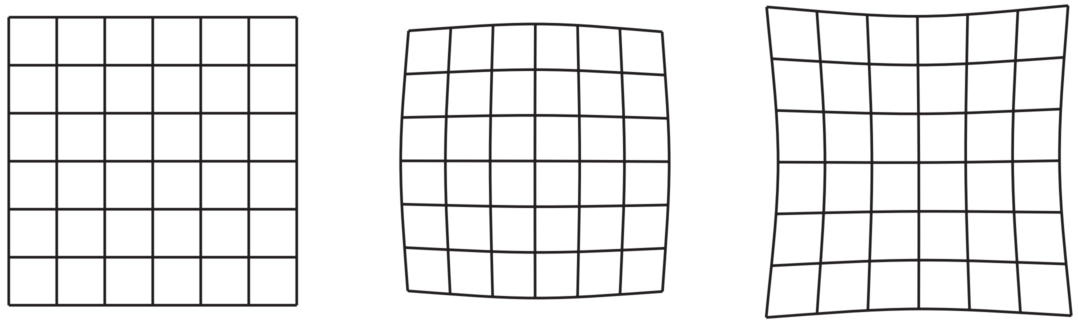


Figure 9.17 Distortion occurs when magnification is not constant across an extended image

abberation plus the five second-order abberation terms). In addition to the five named monochromatic abberations, there are many other higher order abberations that also have to be considered. Abberations can be corrected to a high degree with multiple-element systems (designed using ray-tracing techniques) composed of lenses and irises to eliminate off-axis light. For example, a camera lens with a focal length of 50 mm, one of the simplest lenses in photography, is typically composed of about six individual elements. However, optical systems never completely eliminate all abberation, so designing a system always involves some degree of compromise in choosing which abberations to minimize and which ones you can live with.

Exercises

9.2 The Eikonal Equation

P9.1 (a) Suppose that a region of air above the desert on a hot day has an index of refraction that varies with height y according to $n\left(y\right) = n_0\sqrt{1+y^2/h^2}$. Show that $R\left(x,y\right) = n_0 x - n_0 y^2/2h$ is a solution of the eikonal equation (9.9).

(b) Give an expression for $\hat{\mathbf{s}}$ as a function of y.

(c) Compute $\hat{\mathbf{s}}$ for $y = h$, $y = h/2$, and $y = h/4$. Represent these vectors graphically and place them sequentially point-to-tail to depict how the light bends as it travels.

P9.2 Prove that under the approximation of very short wavelength, the Poynting vector is directed along $\nabla R\left(\mathbf{r}\right)$ or $\hat{\mathbf{s}}$.

Solution: (partial)

From Faraday's law (1.37) we have

$$\begin{aligned}
\mathbf{B}(\mathbf{r},t) &= \frac{i}{\omega}\nabla\times\left[\mathbf{E}_0\left(\mathbf{r}\right)e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}\right] \\
&= \frac{i}{\omega}\left[\nabla\times\mathbf{E}_0\left(\mathbf{r}\right)\right]e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} - \frac{k_{\text{vac}}}{\omega}\left[\nabla R\left(\mathbf{r}\right)\right]e^{i[k_{\text{vac}}SR(\mathbf{r})-\omega t]}\times\mathbf{E}_0\left(\mathbf{r}\right)e^{ik_{\text{vac}}R(\mathbf{r})}, \\
&= \frac{i\lambda_{\text{vac}}}{2\pi c}\left[\nabla\times\mathbf{E}_0\left(\mathbf{r}\right)\right]e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} - \frac{1}{c}\left[\nabla R\left(\mathbf{r}\right)\right]e^{i[k_{\text{vac}}SR(\mathbf{r})-\omega t]}\times\mathbf{E}_0\left(\mathbf{r}\right)e^{ik_{\text{vac}}R(\mathbf{r})}
\end{aligned}$$

In the limit of very short wavelength, this becomes

$$\mathbf{B}(\mathbf{r},t) \to -\frac{1}{c}\left[\nabla R\left(\mathbf{r}\right)\right]\times\mathbf{E}_0\left(\mathbf{r}\right)e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}.$$

From Gauss'slaw (1.35) and from (2.15) we have

$$\begin{aligned}
&\nabla\cdot\left\{\left[1+\chi\left(\mathbf{r}\right)\right]\mathbf{E}(\mathbf{r},t)\right\} = \nabla\cdot\left\{\left[1+\chi\left(\mathbf{r}\right)\right]\mathbf{E}_0\left(\mathbf{r}\right)e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}\right\} = 0 \\
&\Rightarrow \left\{\nabla\cdot\left[\left[1+\chi\left(\mathbf{r}\right)\right]\mathbf{E}_0\left(\mathbf{r}\right)\right]\right\}e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} + ik_{\text{vac}}\left[\nabla R\left(\mathbf{r}\right)\right]\cdot\left[1+\chi\left(\mathbf{r}\right)\right]\mathbf{E}_0\left(\mathbf{r}\right)e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} = 0 \\
&\Rightarrow \left[\nabla R\left(\mathbf{r}\right)\right]\cdot\mathbf{E}_0\left(\mathbf{r}\right) = i\lambda_{\text{vac}}\frac{\nabla\cdot\left[\left[1+\chi\left(\mathbf{r}\right)\right]\mathbf{E}_0\left(\mathbf{r}\right)\right]}{2\pi\left[1+\chi\left(\mathbf{r}\right)\right]}
\end{aligned}$$

In the limit of very short wavelength, this becomes

$$\left[\nabla R\left(\mathbf{r}\right)\right]\cdot\mathbf{E}_0\left(\mathbf{r}\right) \to 0$$

Compute the time average of

$$\begin{aligned}
\mathbf{S}_{\text{Poynting}} &= \frac{1}{\mu_0}\text{Re}\left\{\mathbf{E}(\mathbf{r},t)\right\}\times\text{Re}\left\{\mathbf{B}(\mathbf{r},t)\right\} \\
&= \frac{1}{4\mu_0}\left[\mathbf{E}\left(\mathbf{r},t\right)+\mathbf{E}^*\left(\mathbf{r},t\right)\right]\times\left[\mathbf{B}(\mathbf{r},t)+\mathbf{B}^*\left(\mathbf{r},t\right)\right]
\end{aligned}$$

Employ the BAC-CAB rule (see P 0.12).

9.3 Fermat's Principle

P9.3 Use Fermat's Principle to derive the law of reflection (3.6) for a reflective surface.

HINT: Do not consider light that goes directly from A to B; require a single bounce.

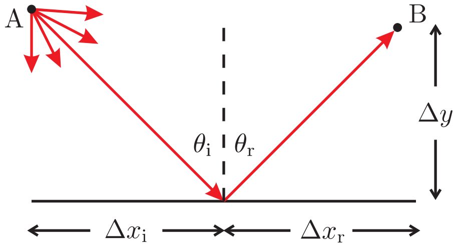


Figure 9.18

P9.4 Show that Fermat's Principle fails to give the correct path for an extraordinary ray entering a uniaxial crystal whose optic axis is perpendicular to the surface. HINT: With the index given by (5.32), show that Fermat's principle leads to an answer that neither agrees with the direction of the **k**-vector (5.35) nor with the direction of the Poynting vector (5.43).

9.5 Reflection and Refraction at Curved Surfaces

P9.5 Derive the ABCD matrix that takes a ray on a *round trip* through a simple laser cavity consisting of a flat mirror and a concave mirror of radius R separated by a distance L. HINT: Start at the flat mirror. Use the matrix in (9.29) to travel a distance L. Use the matrix in (9.39) to represent reflection from the curved mirror. Then use the matrix in (9.29) to return to the flat mirror. The matrix for reflection from the flat mirror is the identity matrix (i.e. $R_\text{flat} \to \infty$).

P9.6 Derive the ABCD matrix for a thick lens made of material n_2 surrounded by a liquid of index n_1. Let the lens have curvatures R_1 and R_2 and thickness d.

Answer:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 + \frac{d}{R_1}\left(\frac{n_1}{n_2} - 1\right) & d\frac{n_1}{n_2} \\ -\left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{d}{R_1 R_2}\left(2 - \frac{n_1}{n_2} - \frac{n_2}{n_1}\right) & 1 - \frac{d}{R_2}\left(\frac{n_1}{n_2} - 1\right) \end{bmatrix}$$

9.6 Image Formation by Mirrors and Lenses

P9.7 (a) Show that the ABCD matrix for a thick lens (see P 9.6) reduces to that of a thin lens (9.55) when the thickness goes to zero. Take the index outside of the lens to be $n_1 = 1$.

(b) Find the ABCD matrix for a thick window (thickness d). Take the index outside of the window to be $n_1 = 1$. HINT: A window is a thick lens with infinite radii of curvature.

P9.8 An object is placed in front of a concave mirror. Find the location of the image d_i and magnification M when $d_\text{o} = R$, $d_\text{o} = R/2$, $d_\text{o} = R/4$, and $d_\text{o} = -R/2$ (virtual object). Make a diagram for each situation, depicting rays traveling from a single off-axis point on the object to a corresponding point on the image. You may want to emphasize especially the ray that initially travels parallel to the axis and the ray that initially travels in a direction intersecting the axis at the focal point $R/2$.

P9.9 An object is placed in front of a concave mirror. Find the location of the image d_i and magnification M when $d_\text{o} = 2f$, $d_\text{o} = f$, $d_\text{o} = f/2$, and $d_\text{o} = -f$ (virtual object). Make a diagram for each situation, depicting rays traveling from a single off-axis point on the object to a corresponding point on the image. You may want to emphasize especially the ray that initially travels parallel to the axis and the ray that initially travels in a direction intersecting the axis at the focal point $R/2$.

9.7 Image Formation by Complex Optical Systems

P9.10 A complicated lens element is represented by an ABCD matrix. An object placed a distance d_1 before the unknown element causes an image to appear a distance d_2 after the unknown element.

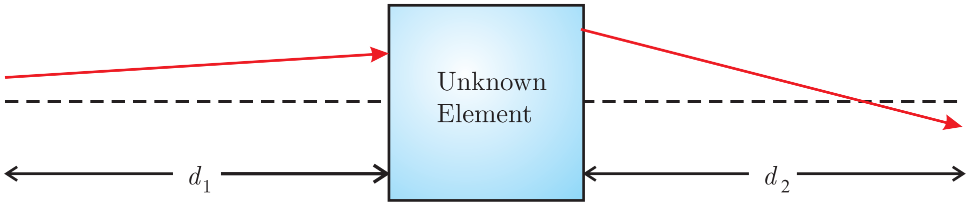


Figure 9.19

Suppose that when $d_1 = \ell$, we find that $d_2 = 2\ell$. Also, suppose that when $d_1 = 2\ell$, we find that $d_2 = 3\ell/2$ with magnification $-1/2$. What is the ABCD matrix for the unknown element?

HINT: Use the conditions for an image (9.59) and (9.60). If the index of refraction is the same before and after, then (9.66) applies. HINT: First find linear expressions for A, B, and C in terms of D. Then put the results into (9.66).

P9.11 (a) Consider a lens with thickness $d = 5$ cm, $R_1 = 5$ cm, $R_2 = -10$ cm, $n = 1.5$. Compute the ABCD matrix of the lens. HINT: See P 9.6.

(b) Where are the principal planes located and what is the effective focal length f_eff for this system?

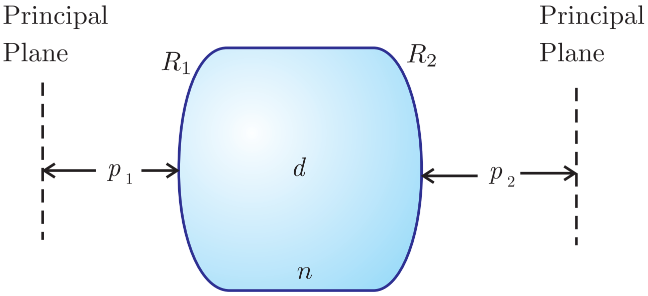


Figure 9.20

L9.12 Deduce the positions of the principal planes and the effective focal length of a compound lens system. Reference the positions of the principal planes to the outside ends of the metal hardware that encloses the lens assembly.

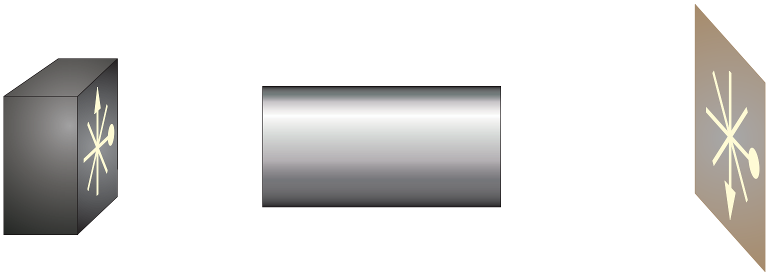


Figure 9.21

HINT: Obtain three sets of distances to the object and image planes and place the data into (9.59) to create three distinct equations for the unknowns A, B, C, and D. Find A, B, and C in terms of D and place the results into (9.66) to obtain the values for A, B, C, and D. The effective focal length and principal planes can then be found through (9.62)–(9.64).

P9.13 Use a computer program to calculate the ABCD matrix for the following compound system known as the "Tessar lens":

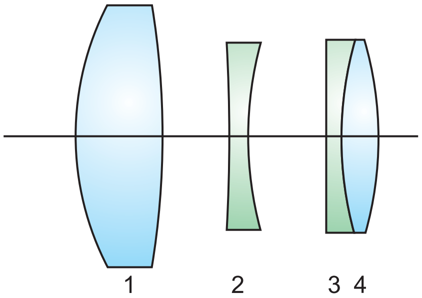


Figure 9.22

The details of this lens are as follows (all distances are in the same units, and *only the magnitude of curvatures are given—you decide the sign*): Convex-convex lens 1 (thickness 0.357, $R_1 = 1.628$, $R_2 = 27.57$, $n = 1.6116$) is separated by 0.189 from concave-concave lens 2 (thickness 0.081, $R_1 = 3.457$, $R_2 = 1.582$, $n = 1.6053$), which is separated by 0.325 from plano-concave lens 3 (thickness 0.217, $R_1 = \infty$, $R2 = 1.920$, $n = 1.5123$), which is directly followed by convex-convex lens 4 (thickness 0.396, $R_1 = 1.920$, $R_2 = 2.400$, $n = 1.6116$).

HINT: You can reduce the number of matrices you need to multiply by using the "thick lens" matrix.

9.8 Stability of Laser Cavities

P9.14 (a) Show that the cavity depicted in Fig. 9.11c is stable if

$$0 < \left(1 - \frac{L}{R_1}\right)\left(1 - \frac{L}{R_2}\right) < 1$$

(b) The two concave mirrors have radii $R_1 = 60$ cm and $R_2 = 100$ cm. Over what range of mirror separation L is it possible to form a stable laser cavity?

HINT: There are two different stable ranges with an unstable range between them.

P9.15 Find the stable ranges for $L_1 = L_2 = L$ for the laser cavity depicted in Fig. 9.11d with focal length $f = 50$ cm.

L9.16 Experimentally determine the stability range of a HeNe laser with adjustable end mirrors. Check that this agrees reasonably well with theory. Can you think of reasons for any discrepancy?

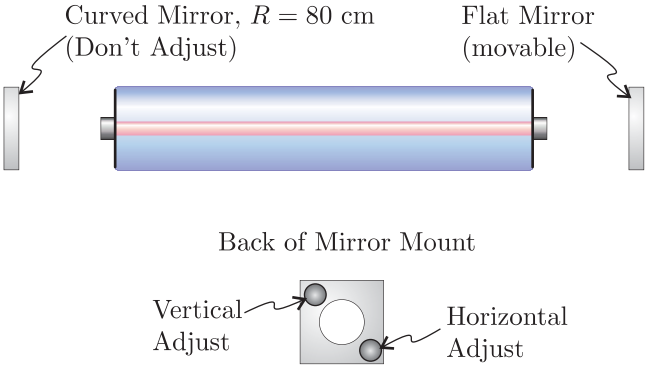


Figure 9.23

Chapter 10

Diffraction

10.1 Huygens' Principle

Christian Huygens developed a wave description for light in the 1600's. However, his ideas were largely overlooked at the time because of Sir Isaac Newton's rejection of the wave description in favor of his corpuscular theory. It was more than a century later that Thomas Young performed his famous two-slit experiment, conclusively demonstrating the wave nature of light. Even then, Young's conclusions were not accepted for many years, a notable exception being a young Frenchman, Augustin Fresnel. The two formed a close friendship through correspondence, and it was Fresnel that followed up on Young's conclusions and dedicated his life to a study of light. Fresnel's skill as a mathematician allowed him to transform physical intuition into powerful and concise ideas. Perhaps Fresnel's greatest accomplishment was the adaptation of *Huygens' principle* into a mathematical formula. Ironically, it was Newton's calculus that made this possible and it settled the debate between the wave and corpuscular theories.

Huygens' principle asserts that a wave front can be thought of as many wavelets, which propagate and interfere to form new wave fronts. Diffraction is then understood as the spilling of wavelets around corners.

Let us examine the calculus that Fresnel applied to the problem of summing up the contributions from the many wavelets originating in an aperture illuminated by a light field. Each point in the aperture is thought of as a source of a *spherical wave*. In our modern notation, such a spherical wave can be written as proportional to e^{ikR}/R, where R is the distance from the source. As a spherical wave propagates, its strength falls off in proportion to the distance traveled and the phase is related to the distance propagated, similar to the phase of a plane wave. Students should be aware that a spherical wave of the form e^{ikR}/R is not a true solution to Maxwell's equations[1] (see P 10.2). Near $R = 0$, this type of wave wave is in fact a very poor solution to Maxwell's equations. However, if R is much larger than a wavelength, this spherical wave satisfies Maxwell's equations to a good

[1] For simplicity, we use the term "spherical wave" in this book to refer to waves of the type imagined by Huygens (i.e. of the form e^{ikR}/R). There is a different family of waves based on spherical harmonics that are also sometimes referred to as spherical waves. These waves have angular as well as radial dependence, and they *are* solutions to Maxwell's equations. For details see pp. 429–432 of Jackson's *Classical Electrodynamics*, 3rd Ed. (Ref. [1]).

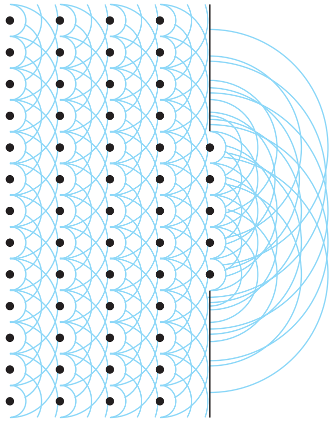


Figure 10.1 Wave fronts depicted as a series of Huygens' wavelets.

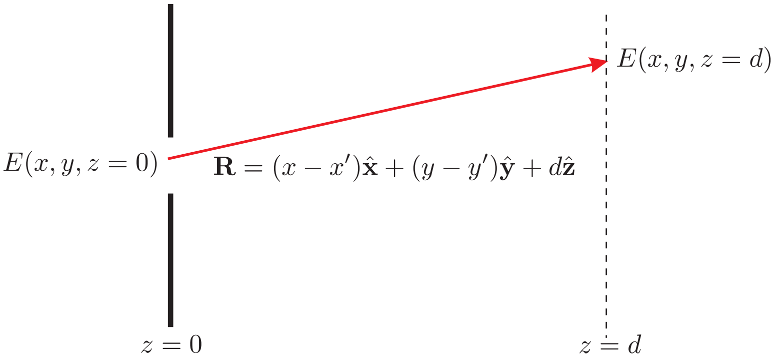

$E(x, y, z = d)$
$E(x, y, z = 0)$
$\mathbf{R} = (x - x')\hat{\mathbf{x}} + (y - y')\hat{\mathbf{y}} + d\hat{\mathbf{z}}$
$z = 0$
$z = d$

Figure 10.2 A wave propagating through an aperture, giving rise to the field at a point downstream.

approximation. In fact, under this approximation a spherical wave can actually be written as a superposition of many plane waves. This is the regime in which Fresnel's diffraction formula (derived in this section and the next) is very successful.

The idea is straightforward. Consider an aperture at $z = 0$ illuminated with a light field distribution $E(x', y', z = 0)$ within the aperture. Then for a point lying somewhere after the aperture, say at $(x, y, z = d)$, the net field is given by adding together spherical waves emitted from each point in the aperture. Each spherical wavelet takes on the strength and phase of the field at the point where it originates. Mathematically, this summation takes the form

$$E(x, y, z = d) = -\frac{i}{\lambda} \iint\limits_{\text{aperture}} E(x', y', z = 0) \frac{e^{ikR}}{R} dx' dy' \tag{10.1}$$

where

$$R = \sqrt{(x - x')^2 + (y - y')^2 + d^2} \tag{10.2}$$

is the radius of each wavelet as it individually intersects the point $(x, y, z = d)$. The constant $-i/\lambda$ in front of the integral in (10.1) ensures the right phase and field strength. We will see how these factors arise in section 10.2. It should be noted that (10.1) considers only a single wavelength of light (i.e. one frequency).

The Fresnel diffraction formula, (10.1), is extremely successful. It was developed a half century before Maxwell assembled his equations. In 1887, Gustav Kirchhoff justified Fresnel's diffraction formula in the context of Maxwell's equations. In doing this he clearly showed the approximations implicit in the theory, and showed that the formula needs to be slightly modified to

$$E\left(x,y,z=d\right)=-\frac{i}{\lambda}\iint\limits_{\text{aperture}}E\left(x',y',z=0\right)\frac{e^{ikR}}{R}\left[\frac{1+\cos\left(\mathbf{r},\hat{\mathbf{z}}\right)}{2}\right]dx'dy' \tag{10.3}$$

The additional factor in square brackets is known as *obliquity factor* ($\cos(\mathbf{r},\hat{\mathbf{z}})$ indicates the cosine of the angle between $\mathbf{r}$ and $\hat{\mathbf{z}}$). Notice that this factor is approximately equal to one when the point $(x, y, z = d)$ is chosen to be in the far-forward direction, and we usually study fields where this approximation holds. The obliquity factor is equal to zero in the case that the field travels in the backwards direction (i.e. in the $-\hat{\mathbf{z}}$ direction). This fixes a problem with Fresnel's earlier version (10.1) based on Huygens' wavelets, which suggests that light can diffract backwards as easily as forwards. In honor of Kirchhoff's work, the formula is now often called the Fresnel-Kirchhoff diffraction formula.

The details of Kirchhoff's derivation are given in Appendix 10.B. Section 10.2 gives a less rigorous derivation, which resorts to the paraxial approximation of the wave equation. In section 10.3, we discuss Babinet's principle, which is a superposition principle for masks and apertures that create diffraction. In section 10.4, we examine Fresnel's approximation made to his own formula (10.1) and find that it is analogous to the paraxial approximation. In section 10.5, we examine the Fraunhofer approximation, a more extreme approximation that only applies to the field at a very large distance after the aperture. We further examine the diffraction integral (in either the Fresnel or the Fraunhofer approximation) in the case of cylindrical symmetry in section 10.6.

10.2 Scalar Diffraction

Consider a light field with a single frequency ω. The light field can be represented by $\mathbf{E}\left(\mathbf{r}\right)e^{-i\omega t}$ which must obey the wave equation

$$\nabla^2\mathbf{E}\left(\mathbf{r}\right)e^{-i\omega t}-\frac{n^2}{c^2}\mathbf{E}\left(\mathbf{r}\right)\frac{\partial^2 e^{-i\omega t}}{\partial t^2}=0 \tag{10.4}$$

Since the temporal part of the field is written explicitly, the time derivative in (10.4) can be performed easily, and the equation reduces to

$$\nabla^2\mathbf{E}\left(\mathbf{r}\right)+k^2\mathbf{E}\left(\mathbf{r}\right)=0 \tag{10.5}$$

where $k \equiv n\omega/c$ is the magnitude of the usual wave vector. Equation (10.5) is called the *Helmholtz equation.* It is the wave equation written for the case of a single frequency, where the trivial time dependence has been removed from the equation. To obtain the full wave solution, the factor $e^{-i\omega t}$ is simply appended to the solution of the Helmholtz equation $\mathbf{E}\left(\mathbf{r}\right)$.

At this point it is convenient to make a significant approximation. We ignore the vectorial nature of (10.5) and consider only the magnitude of $\mathbf{E}(\mathbf{r})$. This is serious! When we use the Fresnel-Kirchhoff diffraction formula we must keep in mind that we have taken this unjustified procedure. The significance of this approximation is discussed in appendix 10.A. Under the scalar approximation, (10.5) becomes the *scalar Helmholtz equation*:

$$\nabla^2 E(\mathbf{r}) + k^2 E(\mathbf{r}) = 0 \tag{10.6}$$

This equation of course is consistent with (10.5) in the case of a plane wave. However, we are interested in so-called spherical waves, which satisfy the vector Helmholtz equation (10.5) only approximately. We can get away with this approximation in the case of a spherical wave only when the radius r is large compared to a wavelength (i.e., $kr \gg 1$) and when the angle is restricted to a narrow angle perpendicular to the polarization.

This highlights an important limitation of the Fresnel-Kirchhoff diffraction formula (10.1), which is a solution to the scalar Helmholtz equation (10.6), but not to the vector Helmholtz equation (10.5). As mentioned in Section 10.6, the Fresnel-Kirchhoff diffraction formula (10.1) can be viewed as a superposition of spherical waves. It turns out that spherical waves of the form $E(r) = E_0 r_0 e^{ikr}/r$ are exact solutions to the scalar Helmholtz equation, (10.6), the proof of which is left as an exercise (see P 10.3). It is therefore not surprising that the Fresnel-Kirchhoff formula satisfies the scalar Helmholtz equation (10.6). The full derivation of the Fresnel-Kirchhoff formula is deferred to Appendix 10.B.

In this section, we will justify the diffraction formula within a simplified context. We will assume that the field that propagates through the aperture is highly directional, such that it propagates mainly in the z-direction. This motivates us to write the field as $E(x,y,z) = \tilde{E}(x,y,z)e^{ikz}$. Upon substitution of this into the scalar Helmholtz equation (10.6), we arrive at

$$\left(\frac{\partial^2 \tilde{E}}{\partial x^2} + \frac{\partial^2 \tilde{E}}{\partial y^2} + 2ik\frac{\partial \tilde{E}}{\partial z} + \frac{\partial^2 \tilde{E}}{\partial z^2}\right) e^{ikz} = 0 \tag{10.7}$$

At this point we make the paraxial wave approximation, which is $|2k\frac{\partial \tilde{E}}{\partial z}| \gg |\frac{\partial^2 \tilde{E}}{\partial z^2}|$. That is, we assume that the amplitude of the field varies slowly in the z-direction such that the wave looks much like a plane wave. We permit the amplitude to change as the wave propagates in the z-direction as long as it does so on a scale much longer than a wavelength. This leads to the paraxial wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + 2ik\frac{\partial}{\partial z}\right) \tilde{E} \cong 0 \tag{10.8}$$

the solution to which is (see P10.5)

$$\tilde{E}(x,y,z) \cong -\frac{i}{\lambda z} \iint\limits_{-\infty}^{\infty} \tilde{E}(x',y',0) e^{i\frac{k}{2z}\left[(x-x')^2+(y-y')^2\right]} dx'dy' \tag{10.9}$$

The field is then given by

$$\begin{aligned} E(x,y,z) &= \tilde{E}(x,y,z)e^{ikz} \\ &\cong -\frac{i}{\lambda z} \iint\limits_{-\infty}^{\infty} \tilde{E}(x',y',0) e^{ik\left[z+\frac{(x-x')^2+(y-y')^2}{2z}\right]} dx'dy' \end{aligned} \tag{10.10}$$

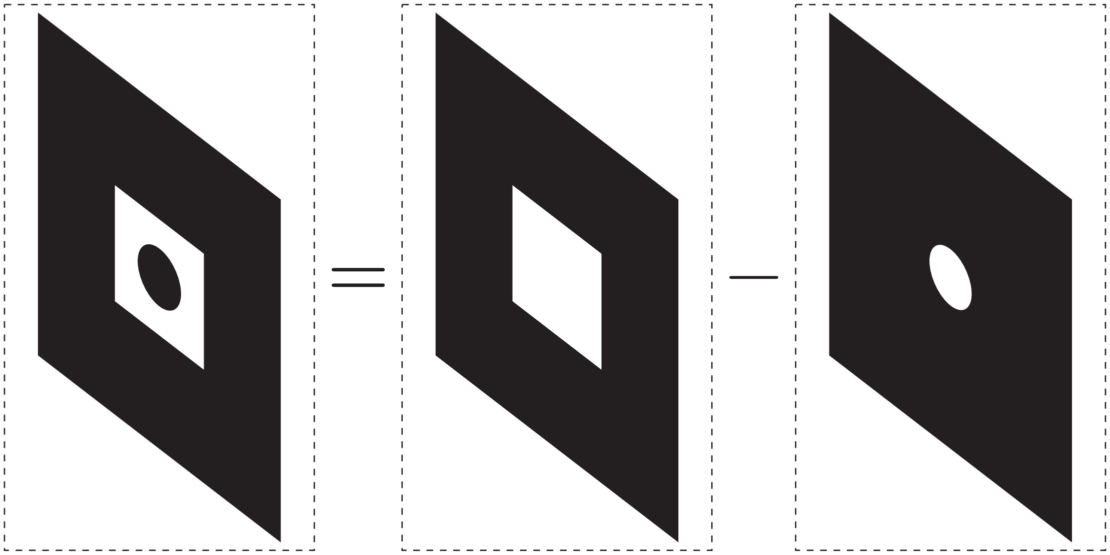


Figure 10.3 Aperture comprised of the region between a circle and a square.

This equation agrees with the Fresnel-Kirchhoff formula (10.1) to the extent that $R \cong z$ in the denominator of the integral and

$$R \cong z + \frac{(x - x')^2 + (y - y')^2}{2z}$$

in the exponent. As we shall see in Section 10.4, this is a good approximation.

10.3 Babinet's Principle

Babinet's principle amounts to a recognition of the linear properties of integration. The principle may be used when a diffraction aperture has a complicated shape so that it is more convenient to break up the diffraction integral (10.3) into several pieces. Students are already used to doing this sort of piecewise approach to integration in other settings. In fact, it is hardly worth giving a name to this approach; perhaps in Babinet's day people were not as comfortable with calculus. As an example of how to use Babinet's principle, suppose that we have an aperture that consists of a circular obstruction within a square opening as depicted in Fig. 10.3. Thus, the light transmits through the region between the circle and the square. One can evaluate the overall diffraction pattern by first evaluating the diffraction integral for the entire square (ignoring the circular block) and then subtracting the diffraction integral for a circular opening having the shape of the block. This removes the unwanted part of the previous integration and yields the overall result. It is important to add and subtract the integrals (i.e. fields), not their squares (i.e. intensity). Remember that it is the electric fields that obey the primary superposition principle.

As trivial as Babinet's principle may seem to the modern student, the principle can also be used to determine diffraction in the shadows behind small obstructions in a wide stream of light. Keep in mind that the diffraction formula (10.3) was derived for finite

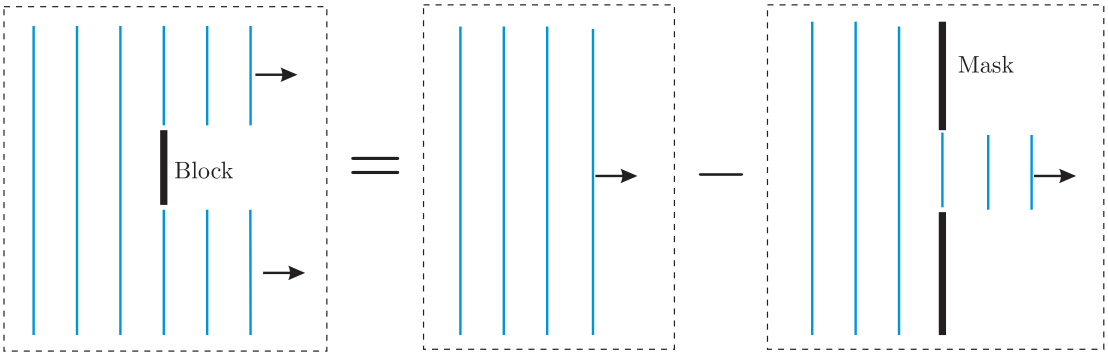


Figure 10.4 A block in a plane wave giving rise to diffraction in the geometric shadow.

apertures or openings in an infinite opaque mask. It therefore may not be obvious that Babinet's principle also applies to an infinitely wide plane wave that is interrupted by finite obstructions. In this case, one simply computes the diffraction of the blocked portions of the field as though these portions were openings in a mask. This result is then subtracted from the uninterrupted field, as depicted in Fig. 10.4.

When Fresnel first presented his diffraction formula to the French Academy of Sciences, a certain judge of scientific papers named Simeon Poisson noticed that the formula predicted that there should be light in the center of the geometric shadow behind a circular obstruction. This seemed so absurd that Fresnel's work was initially disbelieved until the spot was shortly thereafter experimentally confirmed. Needless to say, Fresnel's paper was then awarded first prize, and this spot appearing behind circular blocks has since been known as Poisson's spot.

10.4 Fresnel Approximation

The Fresnel-Kirchhoff diffraction formula (10.3) is valid as long as R and the size of the aperture are both significantly larger than a wavelength. The formula becomes much simpler if we restrict its use to the far-forward direction so that the obliquity factor $[1 + \cos(\mathbf{r}, \hat{\mathbf{z}})]/2$ is approximately equal to one. Even though the Fresnel-Kirchhoff integral looks simple (i.e. a clear implementation of Huygens' superposition of spherical wavelets e^{ikR}/R), it is difficult to evaluate analytically. The integral can be difficult even if the field $E(x', y', z = 0)$ is constant across the aperture.

Fresnel introduced an approximation to his diffraction formula that makes the integration much easier to perform. The approximation is analogous to the paraxial approximation made for rays in chapter 9. Similarly, the Fresnel approximation requires the avoidance of large angles with respect to the z-axis. Besides setting the obliquity factor equal to one, Fresnel made the following simplification to the distance R given in (10.2). In the denominator of (10.3) he approximated R by the distance d. He thereby removed the dependence on x' and y' so that it can be brought out in front of the integral. This is valid to the extent that we restrict ourselves to small angles:

$$R \cong d \qquad \text{(denominator only; paraxial approximation)} \tag{10.11}$$

This approximation is wholly inappropriate in the exponent of (10.3) since small changes in R can result in dramatic variations in e^{ikR}. To approximate R in the exponent, we must proceed with caution. To this end we expand (10.2) under the assumption $d^2 \gg (x-x')^2 + (y-y')^2$. Again, this is consistent with the idea of restricting ourselves to relatively small angles. The expansion of (10.2) is written as

$$\begin{aligned} R &= \sqrt{(x-x')^2 + (y-y')^2 + d^2} \\ &= d\sqrt{1 + \frac{(x-x')^2 + (y-y')^2}{d^2}} \\ &\cong d\left[1 + \frac{(x-x')^2 + (y-y')^2}{2d^2} + \cdots\right] \qquad \text{(paraxial approximation)} \end{aligned} \tag{10.12}$$

Substitution of (10.11) and (10.12) into the Fresnel-Kirchhoff diffraction formula (10.3) and (10.2) yields

$$E(x,y,d) \cong -\frac{ie^{ikd}e^{i\frac{k}{2d}\left(x^2+y^2\right)}}{\lambda d} \iint\limits_{\text{aperture}} E\left(x',y',0\right) e^{i\frac{k}{2d}\left(x'^2+y'^2\right)} e^{-i\frac{k}{d}\left(xx'+yy'\right)} dx'dy' \tag{10.13}$$

This formula is called the Fresnel approximation. It may seem rather complicated, but in terms of being able to perform the integration we are far better off than previously. Notice that the integral can be interpreted as a two-dimensional Fourier transform on $E\left(x',y',0\right) e^{i\frac{k}{2d}\left(x'^2+y'^2\right)}$. The Fresnel approximation to the Fresnel-Kirchhoff formula (10.1) renders an expression which is identical to the exact solution of the paraxial wave equation.

10.5 Fraunhofer Approximation

An additional approximation to the diffraction integral was made famous by Joseph von Fraunhofer. The *Fraunhofer approximation* agrees with the Fresnel approximation in the limiting case when the field is observed at a distance far after the aperture (called the *far field*). The Fraunhofer approximation also requires small angles (i.e. the paraxial approximation). As the diffraction pattern continuously evolves along the z-direction it is described everywhere by the Fresnel approximation. However, it eventually evolves into a final diffraction pattern that maintains itself as it continues to propogate (although it increases its size in proportion to distance). It is this far-away diffraction pattern that is obtained from the Fraunhofer approximation.

In many textbooks, the Fraunhofer approximation is presented first because the formula is easier to use. However, since it is a special case of the Fresnel approximation, it logically should be discussed afterwards as we are doing here. To obtain the diffraction pattern very far after the aperture, we make the following assumption:

$$e^{i\frac{k}{2d}\left(x'^2+y'^2\right)} \cong 1 \qquad \text{(far field)} \tag{10.14}$$

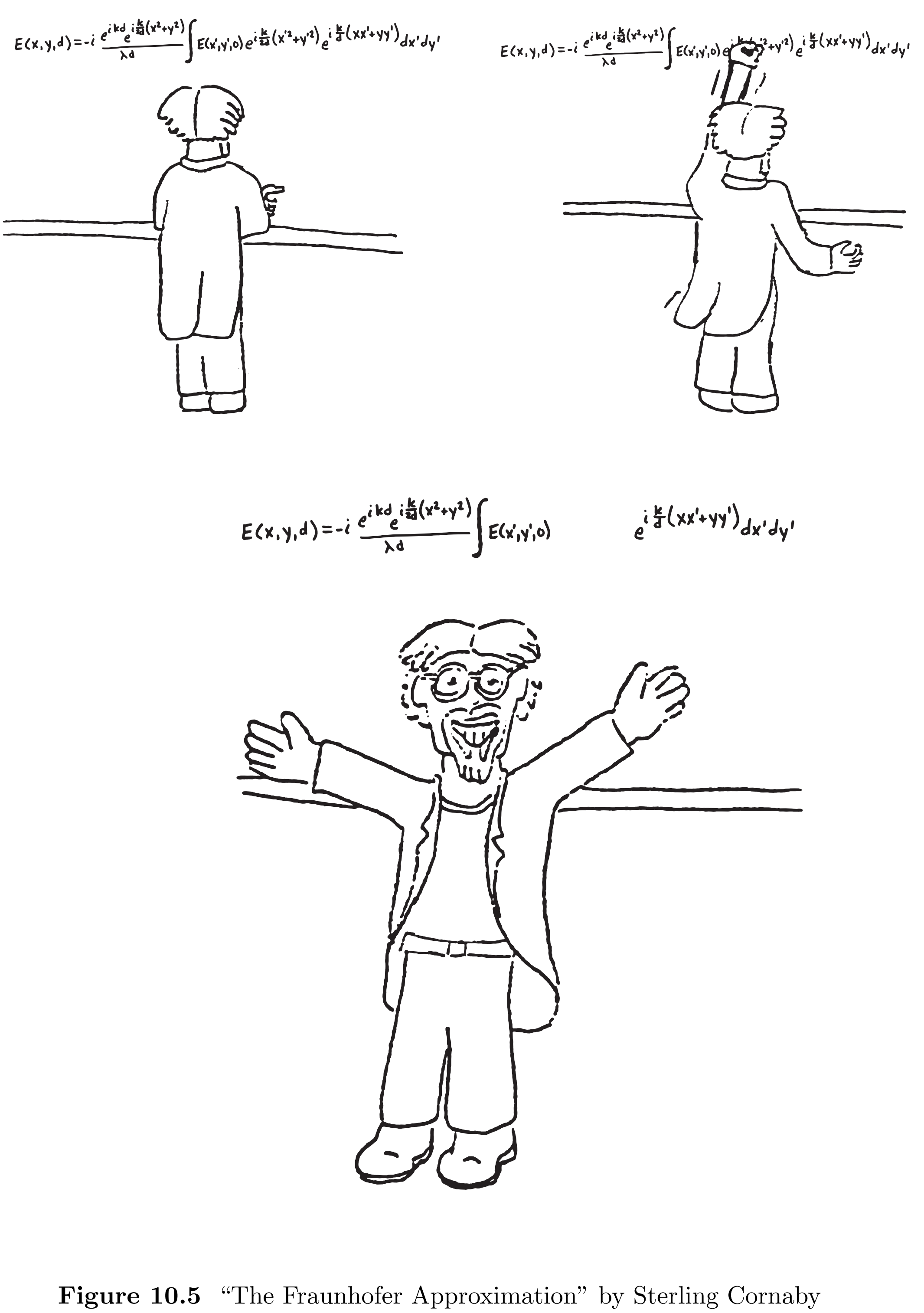

$E(x,y,d) = -i \frac{e^{ikd} e^{i\frac{k}{2d}(x^2+y^2)}}{\lambda d} \int E(x',y',0) e^{i\frac{k}{2d}(x'^2+y'^2)} e^{i\frac{k}{d}(xx'+yy')} dx'dy'$
$E(x,y,d) = -i \frac{e^{ikd} e^{i\frac{k}{2d}(x^2+y^2)}}{\lambda d} \int E(x',y',0) e^{i\frac{k}{2d}(x'^2+y'^2)} e^{i\frac{k}{d}(xx'+yy')} dx'dy'$
$E(x,y,d) = -i \frac{e^{ikd} e^{i\frac{k}{2d}(x^2+y^2)}}{\lambda d} \int E(x',y',0) \quad e^{i\frac{k}{d}(xx'+yy')} dx'dy'$

Figure 10.5 "The Fraunhofer Approximation" by Sterling Cornaby

Joseph von Fraunhofer

(1787–1826, German)

Fraunhofer was orphaned at a young age and was apprenticed to a glass maker. He was treated harshly, but through the help of the Prince of Bavaria he eventually received a good education. He became expert at making optical devices, and invented the diffraction grating. He was the first to observe absorption lines in the sun's spectrum. Fraunhofer passed away at a young age. This was not uncommon for glass makers of his time because of the heavy metal vapors associated with their trade.

This approximation depends on a comparison of the size of the aperture to the distance d where the diffraction pattern is observed. Thus, we need

$$d \gg \frac{k}{2}\,(\text{aperture radius})^2 \qquad \text{(condition for far field)} \tag{10.15}$$

By substituting (10.14) into (10.13), the Fraunhofer approximation yields

$$E\left(x,y,d\right) \cong -\frac{ie^{ikd}e^{i\frac{k}{2d}\left(x^2+y^2\right)}}{\lambda d} \iint\limits_{\text{aperture}} E\left(x',y',0\right)e^{-i\frac{k}{d}\left(xx'+yy'\right)}dx'dy' \tag{10.16}$$

As students will no doubt appreciate, the removal of $e^{i\frac{k}{2d}\left(x'^2+y'^2\right)}$ from the integrand improves our ability to perform the integration. Notice that the integral can now be interpreted as a two-dimensional Fourier transform on the aperture field $E\left(x',y',0\right)$.

Once we are in the Fraunhofer regime, a change in d is not very interesting since it appears in the combination x/d or y/d inside the integral, which in the paraxial approximation indicates a small angle from the axis. At a larger distance d, the same angle is achieved with a proportionately larger value of x or y. The Fraunhofer diffraction pattern thus preserves itself forever as the field propagates. It grows in size as the distance d increases, but the *angular* size defined by x/d or y/d remains the same.

10.6 Diffraction with Cylindrical Symmetry

Often the field transmitted by an aperture is cylindrically symmetric. In this case, the field at the aperture can be written as

$$E(x',y',z=0) = E(\rho',z=0) \tag{10.17}$$

where $\rho \equiv \sqrt{x^2+y^2}$. Under cylindrical symmetry, the two-dimensional integration over x' and y' in (10.13) or (10.16) can be reduced to a single-dimensional integral over a cylindrical coordinate ρ'. The Fresnel diffraction integral (10.13) in this situation is given by

$$E\left(\rho, z=d\right) = -\frac{ie^{ikd}e^{i\frac{k\rho^2}{2z}}}{\lambda d}\int_0^{2\pi} d\theta' \int_{\text{aperture}} \rho' d\rho' E\left(\rho', z=0\right) e^{i\frac{k\rho'^2}{2d}} e^{-i\frac{k}{d}[(\rho\cos\theta)(\rho'\cos\theta')+(\rho\sin\theta)(\rho'\sin\theta')]} \tag{10.18}$$

where

$$\begin{aligned} x &= \rho\cos\theta \\ y &= \rho\sin\theta \\ x' &= \rho'\cos\theta' \\ y' &= \rho'\sin\theta' \end{aligned} \tag{10.19}$$

Notice that in the exponent of (10.18) we have

$$\rho'\rho\left[\cos\theta'\cos\theta + \sin\theta'\sin\theta\right] = \rho'\rho\cos\left(\theta'-\theta\right) \tag{10.20}$$

With this simplification, the diffraction formula (10.18) can be written as

$$E\left(\rho, z=d\right) = -\frac{ie^{ikd}e^{i\frac{k\rho^2}{2d}}}{\lambda d}\int_{\text{aperture}} \rho' d\rho' E\left(\rho', z=0\right) e^{i\frac{k\rho'^2}{2d}} \int_0^{2\pi} d\theta' e^{-i\frac{k\rho\rho'}{d}\cos(\theta-\theta')} \tag{10.21}$$

We are able to perform the integration over θ with the help of the formula (0.54)

$$\int_0^{2\pi} e^{-i\frac{k\rho\rho'}{d}\cos(\theta-\theta')} d\theta' = 2\pi J_0\left(\frac{k\rho\rho'}{d}\right) \tag{10.22}$$

where J_0 is called the zero-order Bessel function. Equation (10.21) then reduces to

$$E\left(\rho, z=d\right) = -\frac{2\pi ie^{ikd}e^{i\frac{k\rho^2}{2d}}}{\lambda d}\int_{\text{aperture}} \rho' d\rho' E\left(\rho', z=0\right) e^{i\frac{k\rho'^2}{2d}} J_0\left(\frac{k\rho\rho'}{d}\right) \tag{10.23}$$

(Fresnel approximation with cylindrical symmetry)

The integral in (10.23) is called a Hankel transform on $E\left(\rho', z=0\right) e^{i\frac{k\rho'^2}{2d}}$.

In the case of the Fraunhofer approximation, the diffraction integral becomes a Hankel transform on just the field $E\left(\rho', z=0\right)$ since $\exp\left(i\frac{k\rho'^2}{2d}\right)$ goes to one. Under cylindrical symmetry, the Fraunhofer approximation is

$$E\left(\rho, z=d\right) = -\frac{2\pi ie^{ikd}e^{i\frac{k\rho^2}{2d}}}{\lambda d}\int_{\text{aperture}} \rho' d\rho' E\left(\rho', z=0\right) J_0\left(\frac{k\rho\rho'}{d}\right) \tag{10.24}$$

(Fraunhofer approximation with cylindrical symmetry)

Just as fast Fourier transform algorithms aid in the numerical evaluation of diffraction integrals in Cartesian coordinates, fast Hankel transforms exist and can be used with cylindrically symmetric diffraction integrals.

Appendix 10.A Significance of the Scalar Wave Approximation

As was mentioned in Sect. 10.2, the arbitrary replacement of the field vector $\mathbf{E}$ with its scalar amplitude in the Helmholtz equation (10.6) is unjustified. Nevertheless, the solution of the scalar Helmholtz equation is not completely unassociated with the solution to the vector Helmholtz equation. In fact, if $E_{\text{scalar}}(\mathbf{r})$ obeys the scalar Helmholtz equation (10.6), then

$$\mathbf{E}(\mathbf{r}) = \mathbf{r} \times \nabla E_{\text{scalar}}(\mathbf{r}) \tag{10.25}$$

obeys the vector Helmholtz equation (10.5).

Consider a spherical wave, which is a solution to the scalar Helmholtz equation:

$$E_{\text{scalar}}(\mathbf{r}) = E_0 r_0 e^{ikr}/r \tag{10.26}$$

Remarkably, when this expression is placed into (10.25) the result is zero. Although zero is in fact a solution to the vector Helmholtz equation, it is not very interesting. A more interesting solution to the scalar Helmholtz equation is

$$E_{\text{scalar}}(\mathbf{r}) = r_0 E_0 \left(1 - \frac{i}{kr}\right) \frac{e^{ikr}}{r} \cos\theta \tag{10.27}$$

which is one of an infinite number of solutions that exist. Notice that in the limit of large r, this expression looks similar to (10.26), aside from the factor $\cos\theta$. The vector form of this field according to (10.25) is

$$\mathbf{E}(\mathbf{r}) = -\hat{\phi} r_0 E_0 \left(1 - \frac{i}{kr}\right) \frac{e^{ikr}}{r} \sin\theta \tag{10.28}$$

This field looks approximately like the scalar spherical wave solution (10.26) in the limit of large r if the angle is chosen to lie near $\theta \cong \pi/2$ (spherical coordinates). Since our use of the scalar Helmholtz equation is in connection with this spherical wave under these conditions, the results are close to those obtained from the vector Helmholtz equation.

Appendix 10.B Fresnel-Kirchhoff Diffraction Formula

To begin our derivation of the Fresnel-Kirchhoff diffraction formula, we employ Green's theorem (proven in appendix 10.C):

$$\oint_S \left[U \frac{\partial V}{\partial n} - V \frac{\partial U}{\partial n}\right] da = \int_V \left[U \nabla^2 V - V \nabla^2 U\right] dv \tag{10.29}$$

The notation $\partial/\partial n$ implies a derivative in the direction normal to the surface. We choose for the functions to be used in this formula

$$\begin{aligned} V &\equiv e^{ikr}/r \\ U &\equiv E(\mathbf{r}) \end{aligned} \tag{10.30}$$

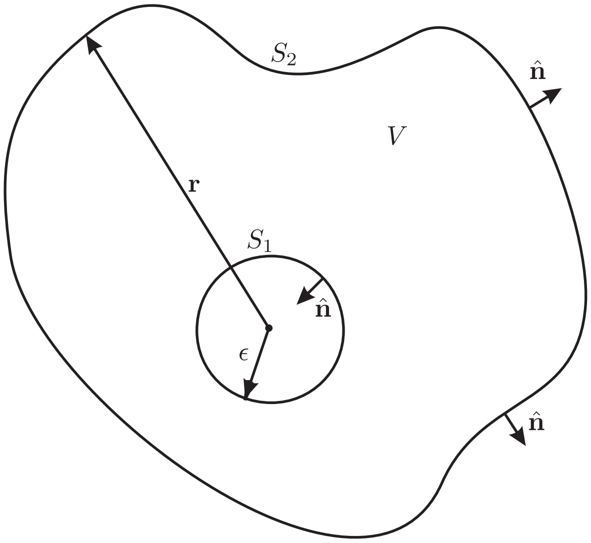


Figure 10.6 A two-part surface enclosing volume V.

where $E(\mathbf{r})$ is assumed to satisfy the scalar Helmholtz equation, (10.6). When these functions are used in Green's theorem (10.29), we obtain

$$\oint_S \left[E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] da = \int_V \left[E \nabla^2 \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \nabla^2 E \right] dv \tag{10.31}$$

The right-hand side of this equation vanishes (as long as we exclude the point $r = 0$; see P 0.13 and P 0.14) since we have

$$E \nabla^2 \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \nabla^2 E = -k^2 E \frac{e^{ikr}}{r} + \frac{e^{ikr}}{r} k^2 E = 0 \tag{10.32}$$

where we have taken advantage of the fact that $E(\mathbf{r})$ and e^{ikr}/r both satisfy (10.6). This is exactly the reason for our judicious choices of the functions V and U since with them we were able to make half of (10.29) disappear. We are left with

$$\oint_S \left[E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] da = 0 \tag{10.33}$$

Now consider a volume between a small sphere of radius ϵ at the origin and an outer surface of whatever shape. The total surface that encloses the volume is comprised of two parts (i.e. $S = S_1 + S_2$ as depicted in Fig. 10.6).

When we apply (10.33) to the surface in Fig. 10.6, we have

$$\oint_{S_2} \left[E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] da = -\oint_{S_1} \left[E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] da \tag{10.34}$$

Our motivation for choosing this geometry with multiple surfaces is that eventually we want to find the field at the origin (inside the little sphere) from knowledge of the field on the outside surface. To this end, we assume that ϵ is small so that $E(\mathbf{r})$ is approximately the same everywhere on the surface S_1. Then the integral over S_1 becomes

$$\oint_{S_1} \left[E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] da = \lim_{r=\epsilon \to 0} \int_0^{2\pi} d\phi \int_0^{\pi} \left[E \left(\frac{\partial}{\partial r} \frac{e^{ikr}}{r} \right) \frac{\partial r}{\partial n} - \frac{e^{ikr}}{r} \left(\frac{\partial E}{\partial r} \right) \frac{\partial r}{\partial n} \right] r^2 \sin\theta d\theta \tag{10.35}$$

where we have used spherical coordinates. Notice that we have employed the chain rule to execute the normal derivative $\partial/\partial n$. Since r always points opposite to the direction of the surface normal $\hat{\mathbf{n}}$, the normal derivative $\partial r/\partial n$ is always equal to -1. (From the definition of the normal derivative we have $\partial r/\partial n \equiv \nabla r \cdot \hat{\mathbf{n}} = -\hat{\mathbf{n}} \cdot \hat{\mathbf{n}} = -1$.) We can now perform the integration in (10.35) as well as take the limit as $\epsilon \to 0$ to obtain

$$\begin{aligned} \lim_{\epsilon \to 0} \oint_{S_1} \left[E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] da &= -4\pi \lim_{\epsilon \to 0} \left[r^2 \left(-\frac{e^{ikr}}{r^2} + ik \frac{e^{ikr}}{r} \right) E - r^2 \frac{e^{ikr}}{r} \left(\frac{\partial E}{\partial r} \right) \right]_{r=\epsilon} \\ &= -4\pi \lim_{\epsilon \to 0} \left[\left(-e^{ik\epsilon} + ik\epsilon e^{ik\epsilon} \right) E - e^{ik\epsilon} \epsilon \left(\frac{\partial E}{\partial r} \right)_{r=\epsilon} \right] \\ &= 4\pi E(0) \end{aligned} \tag{10.36}$$

With the aid of (10.36), Green's theorem applied to our specific geometry (10.34) reduces to

$$E(0) = \frac{1}{4\pi} \oint_{S_2} \left[\frac{e^{ikr}}{r} \frac{\partial E}{\partial n} - E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} \right] da \tag{10.37}$$

The field E on the left is understood to be the value of the field inside the little sphere at the origin. The field E inside the integral is the value of the field on the surface of integration. Hence, if we know the field everywhere on the outer surface S_2, then we can predict the field at the origin. Of course we are free to choose any coordinate system in order to find the field anywhere inside the surface by moving the origin.

Now let us choose a specific surface S_2. We choose an infinite mask with a finite aperture connected to a hemisphere of infinite radius $R \to \infty$. In the end, we will actually be interested in light that enters through the mask and propagates to the origin. In our present coordinate system, the vectors $\mathbf{r}$ and $\hat{\mathbf{n}}$ point opposite to the incoming light. We will transform our coordinate system at a later point.

We must evaluate (10.37) on the surface depicted in the figure. For the portion of S_2 which is on the hemisphere, the integrand tends to zero as R becomes large. To argue this, it is necessary to recognize the fact that at large distances the field must decrease at least as fast as $1/R$. On the mask, we assume, as did Kirchhoff, that both $\partial E/\partial n$ and E are zero. (Later Sommerfeld noticed that these two assumptions actually contradict each other, and he revised Kirchhoff's work to be more accurate. However, the revision in practice makes only a tiny difference as light spills onto the back of the aperture over a distance of only a wavelength. We ignore this and make Kirchhoff's (slightly flawed) assumptions since it

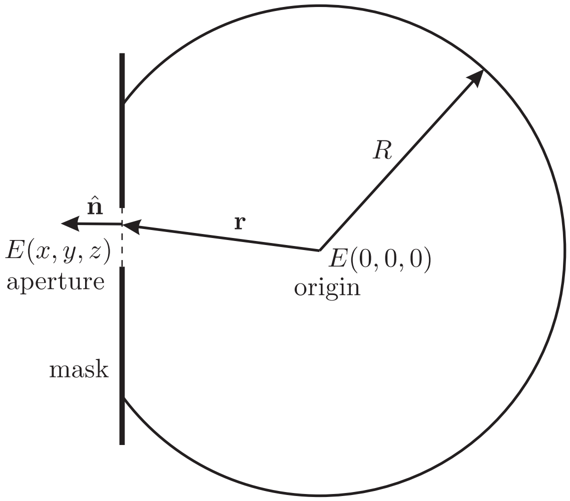


Figure 10.7 Surface S_2 depicted as a mask and a large hemisphere.

saves a lot of work.) Thus, we are left with only the integration over the open aperture:

$$E(0) = \frac{1}{4\pi} \iint\limits_{\text{aperture}} \left[\frac{e^{ikr}}{r} \frac{\partial E}{\partial n} - E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} \right] da \tag{10.38}$$

We have essentially arrived at the result that we are seeking. The field coming through the aperture is integrated to find the field at the origin, which is located beyond the aperture. Let us manipulate the formula a little further. The second term in the integral of (10.38) can be rewritten as follows:

$$\frac{\partial}{\partial n} \frac{e^{ikr}}{r} = \left(\frac{\partial}{\partial r} \frac{e^{ikr}}{r} \right) \frac{\partial r}{\partial n} = \left(\frac{ik}{r} - \frac{1}{r^2} \right) e^{ikr} \cos(\mathbf{r}, \hat{\mathbf{n}}) \underset{r \gg \lambda}{\rightarrow} \frac{ike^{ikr}}{r} \cos(\mathbf{r}, \hat{\mathbf{n}}) \tag{10.39}$$

where $\partial r / \partial n = \cos(\mathbf{r}, \hat{\mathbf{n}})$ indicates the cosine of the angle between $\mathbf{r}$ and $\hat{\mathbf{n}}$. We have also assumed that the distance r is much larger than a wavelength in order to drop a term. Next, we assume that the field in the plane of the aperture can be written as $E \cong \tilde{E}(x, y) e^{ikz}$. This represents a field traveling through the aperture from left to right. Then, we may write the first term in the integral of (10.38) as

$$\frac{\partial E}{\partial n} = \frac{\partial E}{\partial z} \frac{\partial z}{\partial n} = ik\tilde{E}(x, y) e^{ikz} (-1) = -ikE \tag{10.40}$$

Substituting (10.39) and (10.40) into (10.38) yields

$$E(0) = -\frac{i}{\lambda} \iint\limits_{\text{aperture}} E \frac{e^{ikr}}{r} \left[\frac{1 + \cos(\mathbf{r}, \hat{\mathbf{n}})}{2} \right] da \tag{10.41}$$

Finally, we wish to rearrange our coordinate system to that depicted in Fig. 10.2. In our derivation, it was less cumbersome to place the origin at a point after the aperture. Now that we have completed our mathematics, it is convenient to make a change of coordinate system and move the origin to the plane of the aperture as in Fig. 10.2. Then, we can obtain the field at a point lying somewhere after the aperture by computing

$$E\left(x,y,z=d\right)=-\frac{i}{\lambda}\iint\limits_{\text{aperture}}E\left(x',y',z=0\right)\frac{e^{ikR}}{R}\left[\frac{1+\cos\left(\mathbf{r},\hat{\mathbf{z}}\right)}{2}\right]dx'dy' \tag{10.42}$$

where

$$R=\sqrt{\left(x-x'\right)^2+\left(y-y'\right)^2+d^2} \tag{10.43}$$

Equation (10.3) is the same as (10.41) after applying a coordinate transformation. It is called the Fresnel-Kirchhoff diffraction formula and it agrees with (10.1) except for the obliquity factor $[1+\cos\left(\mathbf{r},\hat{\mathbf{z}}\right)]/2$.

Appendix 10.C Green's Theorem

To derive Green's theorem, we begin with the divergence theorem (see (0.26)):

$$\oint\limits_S \mathbf{f}\cdot\hat{\mathbf{n}}\,da=\int\limits_V \nabla\cdot\mathbf{f}\,dv \tag{10.44}$$

The unit vector $\hat{\mathbf{n}}$ always points normal to the surface of volume V over which the integral is taken. Let the vector function $\mathbf{f}$ be $U\nabla V$, where U and V are both analytical functions of the position coordinate $\mathbf{r}$. Then (10.44) becomes

$$\oint\limits_S \left(U\nabla V\right)\cdot\hat{\mathbf{n}}\,da=\int\limits_V \nabla\cdot\left(U\nabla V\right)\,dv \tag{10.45}$$

We recognize $\nabla V\cdot\hat{\mathbf{n}}$ as the directional derivative of V directed along the surface normal $\hat{\mathbf{n}}$. This is often represented in shorthand notation as

$$\nabla V\cdot\hat{\mathbf{n}}=\frac{\partial V}{\partial n} \tag{10.46}$$

The argument of the integral on the right-hand side of (10.45) can be expanded with the chain rule:

$$\nabla\cdot\left(U\nabla V\right)=\nabla U\cdot\nabla V+U\nabla^2 V \tag{10.47}$$

With these substitutions, (10.45) becomes

$$\oint\limits_S U\frac{\partial V}{\partial n}\,da=\int\limits_V \left[\nabla U\cdot\nabla V+U\nabla^2 V\right]dv \tag{10.48}$$

Actually, so far we haven't done much. Equation (10.48) is nothing more than the divergence theorem applied to the vector function $U\nabla V$. Similarly, we can apply the divergence theorem to an alternative vector function given by the combination $V\nabla U$. Thus, we can write an equation similar to (10.48) where U and V are interchanged:

$$\oint_S V\frac{\partial U}{\partial n}\,da = \int_V \left[\nabla V \cdot \nabla U + V\nabla^2 U\right] dv \qquad (10.49)$$

We simply subtract (10.49) from (10.48), and this leads to (10.29) known as Green's theorem.

Exercises

10.1 Huygens' Principle

P10.1 Huygens' principle is often used to describe diffraction through a slits, but it can be also used to describe refraction. Use a drawing program or a ruler and compass to produce a picture similar to Fig. 10.8, which shows that the graphical prediction of refracted angle from the Huygens' principle. Verify that the Huygens picture matches the numerical prediction from Snell's Law for an incident angle of your choice. Use $n_\text{i} = 1$ and $n_\text{t} = 2$.

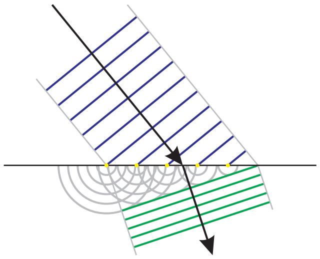


Figure 10.8

HINT: Draw the wavefronts hitting the interface at an angle and treat each point where the wavefronts strike the interface as the source of circular waves propagating into the $n = 2$ material. The wavelength of the circular waves must be exactly half the wavelength of the incident light since $\lambda = \lambda_\text{vac}/n$. Use at least four point sources and connect the matching wavefronts by drawing tangent lines as in the figure.

P10.2 (a) Show that the function

$$f(r) = \frac{A}{r}\cos(kr - \omega t)$$

is a solution to the wave equation in spherical coordinates with only radial dependence,

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}$$

Determine what v is, in terms of k and ω.

(b) If the electric field were a scalar field, we might be done there. However, it's a vector field, and moreover it must satisfy Maxwell's equations. We know from experience that it's generally transverse, and since it's traveling radially let's make a guess that it's oscillating in the $\hat{\phi}$ direction:

$$\mathbf{E}(r) = \frac{A}{r}\cos(kr - \omega t)\,\hat{\phi}$$

Show that this choice for $\mathbf{E}$ unfortunately is not consistent with Maxwell's equations. In particular: (i) show that it does satisfy Gauss's Law (1.1); (ii) compute the curl of **E** use Faraday's Law (1.3) to deduce **B**; (iii) Show that this **B** does satisfy Gauss's Law for magnetism (1.2); (iv) but this **B** it does *not* satisfy Ampere's law (1.4).

(c) A somewhat more complicated "spherical" wave

$$\mathbf{E}(r,\theta) = \frac{A\sin\theta}{r}\left[\cos\left(kr-\omega t\right) - \frac{1}{kr}\sin\left(kr-\omega t\right)\right]\hat{\phi}$$

does satisfy Maxwell's equations. Describe how this wave behaves as a function of r and θ. What conditions need to be satisfied for this equation to reduce to the spherical wave formula used in the diffraction formulas?

10.2 Scalar Diffraction

P10.3 Show that $E\left(r\right) = E_0 r_0 e^{ikr}/r$ is a solution to the scalar Helmholtz equation (10.6).

HINT:

$$\nabla^2\psi = \frac{1}{r}\frac{\partial^2\left(r\psi\right)}{\partial r^2} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}$$

P10.4 Learn by heart the derivation of the Fresnel-Kirchhoff diffraction formula (outlined in Appendix 10.B). Indicate the percentage of how well you understand the derivation. The points for this problem are proportional to your percentage of understanding. If you write 100% percent, it means that you can reproduce the derivation after closing your notes.

P10.5 Check that (10.9) is the solution to the paraxial wave equation (10.8).

P10.6 Apply the Fresnel-Kirchhoff diffraction formula (10.1) to a monochromatic plane wave with intensity I_0, which goes through a circular aperture of *diameter* ℓ. Find the intensity of the light *on axis* (i.e. $x, y = 0$).

HINT: The integral takes on the following form:

$$\begin{aligned} E\left(0,0,d\right) &= -\frac{i}{\lambda}\iint\limits_{\text{aperture}} E\left(x',y',0\right)\frac{e^{ik\sqrt{x'^2+y'^2+d^2}}}{\sqrt{x'^2+y'^2+d^2}}\,dx'dy' \\ &= -\frac{iE_0}{\lambda}\int_0^{2\pi} d\theta' \int_0^{\ell/2} \frac{e^{ik\sqrt{\rho'^2+d^2}}}{\sqrt{\rho'^2+d^2}}\rho'\,d\rho' \end{aligned}$$

Then you will want to make the following change of variables: $\xi \equiv \sqrt{\rho'^2+d^2}$. This will make it easier to accomplish the integration.

Answer: $I\left(0,0,d\right) = 2I_0\left[1-\cos\left[k\sqrt{\left(\ell/2\right)^2+d^2}-kd\right]\right]$.

10.3 Babinet's Principle

P10.7 Subtract the field found in P 10.6 from a plane wave field $E_0 e^{ikd}$ to obtain the on-axis field behind a circular block. Show that the intensity on axis behind the circular block is constant (i.e. independent of d) and is equal to the intensity of the initial plane wave.

L10.8 Why does the on-axis intensity behind a circular opening fluctuate (see P 10.6) whereas the on-axis intensity behind a circular obstruction remains constant (see P 10.7)? Create a collimated laser beam several centimeters wide. Observe the on-axis intensity on a movable screen (e.g. a hand-held card) behind a small circular aperture and behind a small circular obstruction placed in the beam.

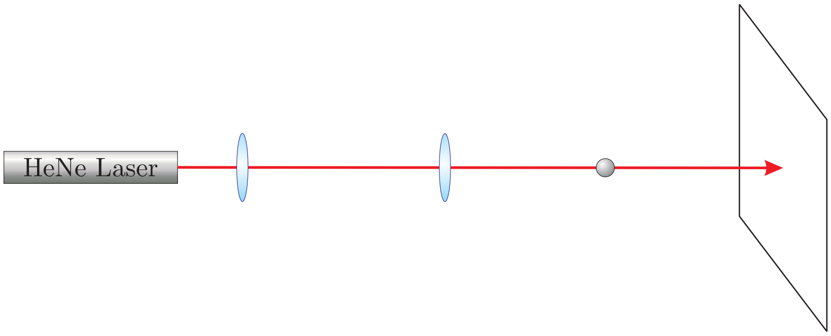


Figure 10.9

10.4 Fresnel Approximation

P10.9 Repeat P 10.6 to find the on-axis intensity after a circular aperture in the Fresnel approximation. HINT: You can make a suitable approximation directly to the answer of P 10.6 to obtain the Fresnel approximation. However, you should also perform the integration under the Fresnel approximation for the sake of gaining experience.

10.5 Fraunhofer Approximation

P10.10 (a) Repeat P 10.6 (or P 10.9) to find the on-axis intensity after a circular aperture in the Fraunhofer approximation.

HINT: You can make a suitable approximation directly to the answer of P 10.9 to obtain the Fraunhofer approximation. However, you should perform the integration under the Fraunhofer approximation for the sake of gaining experience.

(b) Check how well the Fresnel and Fraunhofer approximations work by graphing the three curves (i.e. from P 10.6, P 10.9, and this problem) on a single plot as a function of d. Take $\ell = 10$ μm and $\lambda = 500$ nm. To see the result better, use a log scale on the z-axis.

Answer:

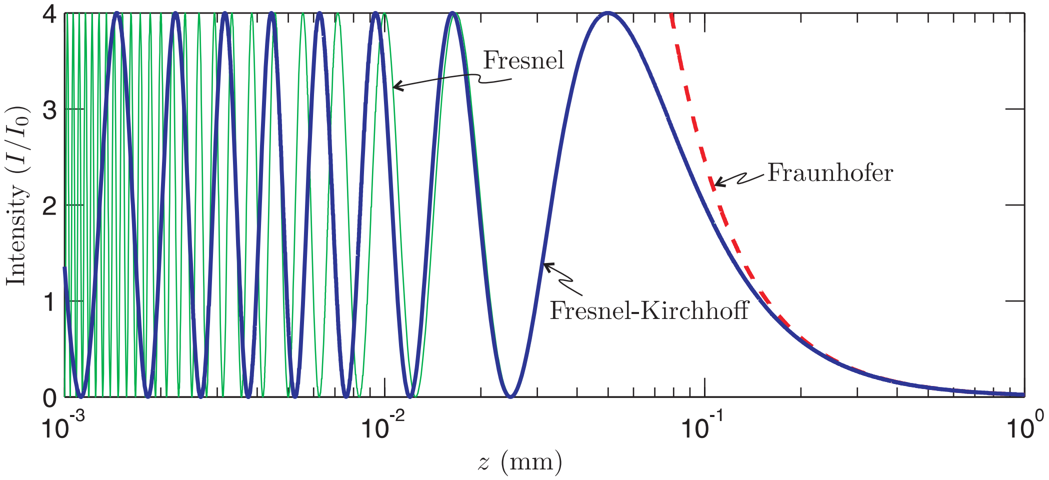


Figure 10.10

P10.11 A single narrow slit has a mask placed over it so the aperture function is not a square pulse but rather a cosine: $E(x', y', 0) = E_0 \cos(x'/L)$ for $-L/2 < x' < L/2$ and $E(x', y', 0) = 0$ otherwise. Calculate the far-field (Fraunhofer) diffraction pattern. Make a plot of intensity as a function of $xkL/2d$; qualitatively compare the pattern to that of a regular single slit.

Chapter 11

Diffraction Applications

11.1 Introduction

In this chapter, we consider a number of practical examples of diffraction. We first examine a Gaussian laser beam. This choice is not arbitrary since most students of optics at some point in their career use laser beams to perform measurements of one kind or another. It is often essential to characterize the laser beam profile and to understand its focusing properties. (Every semester we are contacted by students and faculty from a variety of departments seeking to better understand a laser beam that they are using.) The information presented here will very likely prove valuable to future research activity.

We often think of lasers as collimated beams of light that propagate indefinitely without expanding. However, the laws of diffraction require that every finite beam eventually grow in width. The rate at which a laser beam diffracts depends on its *beam waist* size. Because laser beams usually have narrow divergence angles and therefore obey the paraxial approximation, we can calculate their behavior via the Fresnel approximation discussed in section 10.4. This is done in section 11.2. In section 11.3, we examine the Gaussian field solution as a practical description of simple laser beams. Section 11.A discusses the ABCD law for Gaussian beams, which is a method of computing the effects of optical elements represented by ABCD matrices on Gaussian laser beams.

In section 11.4, we discuss diffraction theory in systems involving lenses. We will find that the Fraunhofer diffraction pattern discussed in section 10.5 for a far-away screen is imaged to the focus of a lens placed in the stream of light. This has important implications for the resolution of instruments such as telescopes or the human eye, as discussed in section 11.5.

The *array theorem* is introduced in section 11.6. This theorem is a powerful mathematical tool that enables one to deal conveniently with diffraction from an array of identical apertures. One of the important uses of the array theorem is in determining diffraction from a grating. As discussed in section 11.7, a diffraction grating can be thought of as an array of narrow slit apertures. In section 11.8, we study the workings of a diffraction spectrometer. To find the resolution limitations, one combines the diffraction properties of gratings with the Fourier properties of lenses discussed in section 11.4.

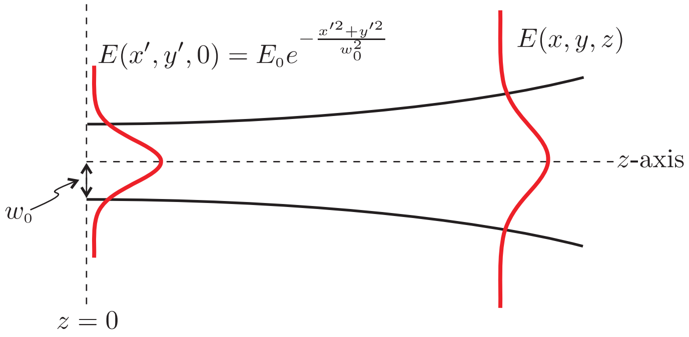


Figure 11.1 Diffraction of a Gaussian field profile.

11.2 Diffraction of a Gaussian Field Profile

Consider the diffraction of a Gaussian field profile. At the plane $z = 0$, we describe the field profile with the functional form

$$E(x', y', 0) = E_0 e^{-\frac{x'^2+y'^2}{w_0^2}} \tag{11.1}$$

where w_0, called the *beam waist*, specifies the radius of beam profile. This beam profile, depicted in Fig. 11.1, is very common for laser beams and is called the zero-order Gaussian mode; more complicated distributions are also possible. To appreciate the meaning of w_0, consider the intensity of the field distribution defined in (11.1):

$$I\left(x', y', 0\right) = I_0 e^{-2\rho'^2/w_0^2} \tag{11.2}$$

where $\rho'^2 \equiv x'^2 + y'^2$. In (11.2) we see that w_0 indicates the radius at which the intensity reduces by the factor $e^{-2} = 0.135$.

We would like to know how this field evolves as it propagates beyond the plane $z = 0$. (We will no longer write $z = d$ as we did in chapter 10; instead we will simply retain the variable z.) We compute the field downstream using the Fresnel approximation (10.13):

$$E\left(x, y, z\right) = -i\frac{e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\int\limits_{-\infty}^{\infty} dx' \int\limits_{-\infty}^{\infty} dy' \left[E_0 e^{-\left(x'^2+y'^2\right)/w_0^2}\right] e^{i\frac{k}{2z}\left(x'^2+y'^2\right)} e^{-i\frac{k}{z}\left(xx'+yy'\right)} \tag{11.3}$$

Notice that we have treated the aperture as being infinitely large. This is not a problem since the Gaussian profile itself limits the dimension of the emission region to a radius on the scale of w_0. Equation (11.3) can be rewritten as

$$E\left(x, y, z\right) = -i\frac{E_0 e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\int\limits_{-\infty}^{\infty} dx' e^{-\left(\frac{1}{w_0^2}-i\frac{k}{2z}\right)x'^2-i\frac{kx}{z}x'} \int\limits_{-\infty}^{\infty} dy' e^{-\left(\frac{1}{w_0^2}+i\frac{k}{2z}\right)y'^2-i\frac{ky}{z}y'} \tag{11.4}$$

The integrals over x' and y' have the identical form and can be done individually with the help of the integral formula (0.52). The algebra is cumbersome, but the integral in the x'

dimension becomes

$$\begin{aligned}
\int_{-\infty}^{\infty} dx' e^{-\left(\frac{1}{w_0^2}-i\frac{k}{2z}\right)x'^2-i\frac{kx}{z}x'} &= \left(\frac{\pi}{\frac{1}{w_0^2}-i\frac{k}{2z}}\right)^{\frac{1}{2}} \exp\left(\frac{\left(-i\frac{kx}{z}\right)^2}{4\left(\frac{1}{w_0^2}-i\frac{k}{2z}\right)}\right) \\
&= \left(\frac{\pi}{-i\frac{k}{2z}\left(1+i\frac{2z}{kw_0^2}\right)}\right)^{\frac{1}{2}} \exp\left(\frac{-kx^2}{2z\left(\frac{2z}{kw_0^2}-i\right)}\right) \\
&= \left(\frac{\lambda z}{\sqrt{1+\left(\frac{2z}{kw_0^2}\right)^2}\, e^{i\tan^{-1}\frac{2z}{kw_0^2}}}\right)^{\frac{1}{2}} \exp\left(\frac{-kx^2\left[\frac{2z}{kw_0^2}+i\right]}{2z\left[1+\left(\frac{2z}{kw_0^2}\right)^2\right]}\right)
\end{aligned} \tag{11.5}$$

A similar expression results from the integration on y'.

When (11.5) and the equivalent expression for the y-dimension are used in (11.4), the result is

$$E\left(x,y,z\right) = E_0 e^{ikz} e^{i\frac{k}{2z}\left(x^2+y^2\right)} \frac{1}{\sqrt{1+\left(\frac{2z}{kw_0^2}\right)^2}} \frac{e^{-\frac{k\left(x^2+y^2\right)}{2z\left[1+\left(\frac{2z}{kw_0^2}\right)^2\right]}\left[\frac{2z}{kw_0^2}+i\right]}}{e^{i\tan^{-1}\frac{2z}{kw_0^2}}} \tag{11.6}$$

This rather complicated expression for the field distribution is in fact very useful and can be directly interpreted, as discussed in the next section.

Before proceeding, we take a moment to mention that this Fresnel integral can also be performed while utilizing the cylindrical symmetry. A Gaussian field profile is one of few diffraction problems that can be handled conveniently in either the Cartesian or the cylindrical coordinate systems. In cylindrical coordinates, the Fresnel diffraction integral (10.23) takes the form

$$E\left(\rho,z\right) = -\frac{2\pi i e^{ikz} e^{i\frac{k\rho^2}{2z}}}{\lambda z} \int_0^{\infty} \rho' d\rho' E_0 e^{-\rho'^2/w_0^2} e^{i\frac{k\rho'^2}{2z}} J_0\left(\frac{k\rho\rho'}{z}\right) \tag{11.7}$$

We can use the integral formula (0.56) to obtain

$$E\left(\rho,z\right) = -i\frac{2\pi E_0 e^{ikz} e^{i\frac{k\rho^2}{2z}}}{\lambda z} \frac{e^{-\frac{\left(\frac{k\rho}{z}\right)^2}{4\left[\frac{1}{w_0^2}-i\frac{k}{2z}\right]}}}{2\left[\frac{1}{w_0^2}-i\frac{k}{2z}\right]} \tag{11.8}$$

which is identical to (11.6).

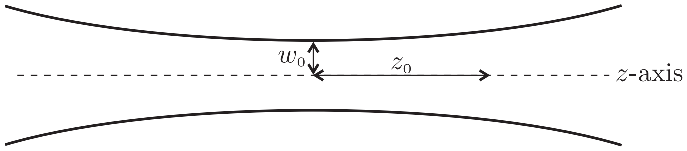


Figure 11.2 A Gaussian laser field profile in the vicinity of its beam waist.

11.3 Gaussian Laser Beams

The rather complicated Gaussian field expression (11.6) can be cleaned up through the judicious introduction of new quantities:

$$E(\rho, z) = E_0 \frac{w_0}{w(z)} e^{-\frac{\rho^2}{w^2(z)}} e^{ikz + i\frac{k\rho^2}{2R(z)} - i\tan^{-1}\frac{z}{z_0}} \tag{11.9}$$

where

$$\rho^2 \equiv x^2 + y^2, \tag{11.10}$$

$$w(z) \equiv w_0\sqrt{1 + z^2/z_0^2}, \tag{11.11}$$

$$R(z) \equiv z + z_0^2/z, \tag{11.12}$$

$$z_0 \equiv \frac{kw_0^2}{2} \tag{11.13}$$

This formula describes the lowest-order Gaussian mode, the most common laser beam profile. (Please be aware that some lasers are *multimode* and exhibit more complicated spatial mode structures—e.g. a high-power YAG laser.) It turns out that (11.9) works equally well for negative values of z. The expression can therefore be used to describe the field of a simple laser beam everywhere (before and after it goes through a focus). In fact, the expression works also near $z = 0$! One might call into question the paraxial approximation for small z since the radius of the beam might be larger than z. Nevertheless, at $z = 0$ the diffracted field (11.9) returns the exact expression for the original field profile (11.1) (see P 11.1). (There is good reason for this since the solution (11.9) obeys the scalar Helmholtz equation (10.6) under the paraxial approximation, where the second derivative with respect to z is neglected.) In short, (11.9) may be used with impunity as long as the divergence angle of the beam is not too wide.

To begin our interpretation of (11.9), consider the intensity profile $I \propto E^*E$ as depicted in Fig. 11.2:

$$I(\rho, z) = I_0 \frac{w_0^2}{w^2(z)} e^{-\frac{2\rho^2}{w^2(z)}} = \frac{I_0}{1 + z^2/z_0^2} e^{-\frac{2\rho^2}{w^2(z)}} \tag{11.14}$$

By inspection we see that $w(z)$ is the radius of the beam as a function of z. At $z = 0$, the *beam waist*, $w(z = 0)$ reduces to w_0, as is seen in (11.11). The parameter z_0, known as the *Rayleigh range*, specifies the distance along the axis from $z = 0$ to the point where the intensity decreases by a factor of 2. Note that w_0 and z_0 are not independent of each other but are connected through the wavelength according to (11.13). There is a tradeoff: a small beam waist means a short *depth of focus*. That is, a small w_0 means a small z_0.

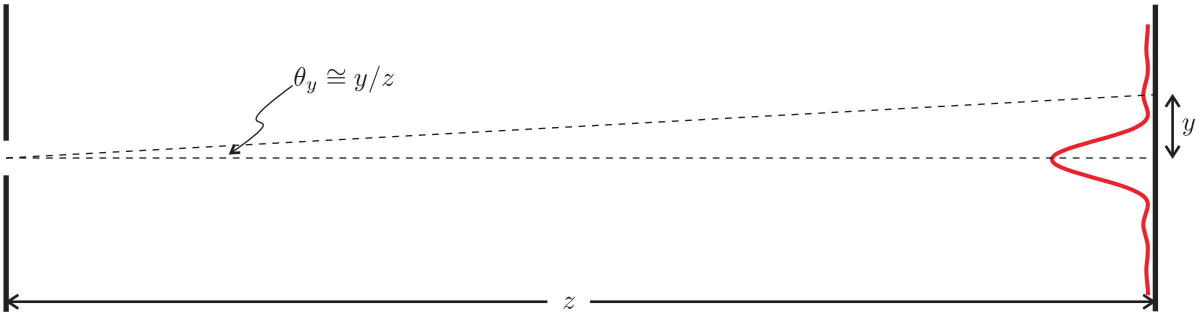


Figure 11.3 Diffraction in the far field.

We now return to an examination of the electric field (11.9). As a reminder, to restore the temporal dependence of the field, we simply append $e^{-i\omega t}$ to the solution, as discussed in connection with (10.5). Let us consider the phase terms that appear in (11.9). The factor $\exp\left\{ikz + ik\rho^2/2R(z)\right\}$ describes the phase of curved wave fronts, where $R(z)$ is the radius of curvature of the wave front at z. At $z = 0$, the radius of curvature is infinite (see (11.12)), meaning that the wave front is flat at the laser beam waist. In contrast, at very large values of z we have $R(z) \cong z$ (see (11.12)). In this case, we may write these phase terms as $kz + \frac{k\rho^2}{2R(z)} \cong k\sqrt{z^2 + \rho^2}$. This describes a spherical wave front emanating from the origin out to point (ρ, z). The Fresnel approximation (same as the paraxial approximation) essentially replaces spherical wave fronts with the former parabolic approximation. Near the origin, the wave fronts are flat. Far from the origin, the wave fronts are spherical.

The phase factor $\exp\left(-i\tan^{-1} z/z_0\right)$ is perhaps the most mysterious. It is called the Gouy shift and is actually present for any light that goes through a focus, not just laser beams. The Gouy shift is not overly dramatic since the expression $\tan^{-1} z/z_0$ ranges from $-\pi/2$ (at $z = -\infty$) to $\pi/2$ (at $z = +\infty$). Nevertheless, when light goes through a focus, it experiences an overall phase shift of π.

11.4 Fraunhofer Diffraction Through a Lens

As has been previously discussed, the Fraunhofer approximation applies to diffraction when the propagation distance from an aperture is sufficiently large (see (10.15) and (10.16)). The intensity of the far-field diffraction pattern is

$$I(x,y,z) = \frac{1}{2}c\epsilon_0 \left| \frac{1}{\lambda z} \iint\limits_{\text{aperture}} E\left(x', y', 0\right) e^{-ik\left(\frac{x}{z}x' + \frac{y}{z}y'\right)} dx'dy' \right|^2 \tag{11.15}$$

Notice that the dependence of the diffraction on x, y, and z comes only through the combinations $\theta_x \cong x/z$ and $\theta_y \cong y/z$. Therefore, the diffraction pattern in the Fraunhofer limit is governed by the two angles θ_x and θ_y, and the pattern preserves itself indefinitely. As the light continues to propagate, the pattern increases in size at a rate proportional to distance traveled so that the angular width is preserved. The situation is depicted in Fig. 11.3.

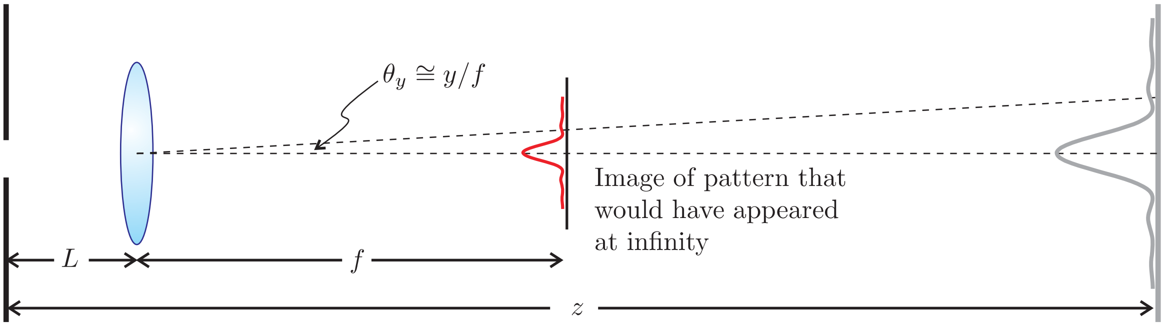


Figure 11.4 Imaging of the Fraunhofer diffraction pattern to the focus of a lens.

The Fraunhofer limit corresponds to the ultimate amount of diffraction that light in an optical system experiences. Mathematically, it is obtained via a two-dimensional Fourier transform as seen in (11.15). The Fraunhofer limit is very important in a variety of optical instruments (e.g. telescopes, spectrometers), discussed later in this chapter.

Recall that in order to use the Fraunhofer diffraction formula we need to satisfy $z \gg \pi\,(\text{aperture radius})^2/\lambda$ (see (10.15)). As an example, if an aperture with a 1 cm radius (not necessarily circular) is used with visible light, the light must travel more than a kilometer in order to reach the Fraunhofer limit. It may therefore seem unlikely to reach the Fraunhofer limit in a typical optical system, especially if the aperture or beam size is relatively large. Nevertheless, spectrometers, which typically utilize diffraction gratings many centimeters wide, depend on achieving the Fraunhofer limit within the confines of a manageable instrument box. This is accomplished using imaging techniques, which is the topic addressed in this section.

Consider a lens with focal length f placed in the path of light following an aperture (see Fig. 11.4). Let the lens be placed an arbitrary distance L after the aperture. The lens produces an image of the Fraunhofer pattern at a new location d_i following the lens according to the imaging formula (see (9.51))

$$\frac{1}{f} = \frac{1}{-(z-L)} + \frac{1}{d_\mathrm{i}}. \tag{11.16}$$

Keep in mind that the lens interrupts the light before the Fraunhofer pattern has a chance to form, at a distance z after the aperture (or a distance $z - L$ after the location of the lens). This means that the Fraunhofer diffraction pattern may be thought of as a *virtual object* for the imaging system. Since the Fraunhofer diffraction pattern occurs at very large distances (i.e. $z \to \infty$) we see that the image of the Fraunhofer pattern must appear at the focus of the lens:

$$d_\mathrm{i} \cong f. \tag{11.17}$$

Thus, a lens makes it very convenient to observe the Fraunhofer diffraction pattern even from relatively large apertures. It is not necessary to let the light propagate for kilometers. We need only observe the pattern at the focus of the lens as shown in Fig. 11.4. Notice that the spacing L between the aperture and the lens is unimportant to this conclusion.

Even though we know that the Fraunhofer diffraction pattern occurs at the focus of a lens, the question remains as to the size of the image. We would like to know how the size of the diffraction pattern compares to what would have occurred on a far-away screen. To find the answer, let us examine the magnification (9.53), which is given by

$$M = -\frac{d_{\mathrm{i}}}{-\left(z - L\right)} \tag{11.18}$$

Taking the limit of very large z and employing (11.17), the magnification becomes

$$M \to \frac{f}{z} \tag{11.19}$$

This is a remarkable result. When the lens is inserted, the size of the diffraction pattern decreases by the ratio of the lens focal length f to the original distance z to a far-away screen. Since in the Fraunhofer regime the diffraction pattern is proportional to distance (i.e. *size* $\propto z$), the image at the focus of the lens scales in proportion to the focal length (i.e. *size* $\propto f$). This means that *the angular width of the pattern is preserved*! With the lens in place, we can rewrite (11.15) straightaway as

$$I\left(x, y, L + f\right) \cong \frac{1}{2} c\epsilon_0 \left| \frac{1}{\lambda f} \iint\limits_{\text{aperture}} E\left(x', y', 0\right) e^{-i\frac{k}{f}\left(xx' + yy'\right)} dx' dy' \right|^2 \tag{11.20}$$

which describes the intensity distribution pattern at the focus of the lens.

Although (11.20) correctly describes the intensity, we cannot easily write the electric field since the imaging techniques that we have used do not render the phase information. To obtain an expression for the field, it is necessary to use repeatedly the Fresnel diffraction formula. First, the Fresnel diffraction formula is used to find the field arriving at the lens. The next task is to determine what the lens does to the field as the light passes through it. Finally, the Fresnel diffraction formula is used again to find the field distribution at the focus of the lens. The result of this lengthy analysis gives an intensity pattern in agreement with (11.20). However, it also gives the full expression for the field, including its phase.

Before proceeding we take a moment to understand how a lens modifies the field of light as it passes through. Consider a monochromatic light field that goes through a *thin* lens with focal length f. In traversing the lens, the field undergoes a phase shift. Let us reference this phase shift to that experienced by the light that goes through the center of the lens. In the Fig. 11.5, R_1 is a positive radius of curvature, and R_2 is a negative radius of curvature, according to our previous convention. We take the distances ℓ_1 and ℓ_2 to be positive.

The light passing through the off-axis portion of the lens experiences less material than the light passing through the center. The difference in optical path length is $(1 - n)\left(\ell_1 + \ell_2\right)$ (see discussion connected with (9.16)). This means that the phase of the field passing through the off-axis portion of the lens relative to the phase of field passing through the center is

$$\Delta\phi = -k\left(n - 1\right)\left(\ell_1 + \ell_2\right). \tag{11.21}$$

The negative sign indicates a phase *advance* (i.e. same sign as $-\omega t$) in contrast to a phase *delay*, which takes a positive sign. Off axis, the phase advances because the light travels

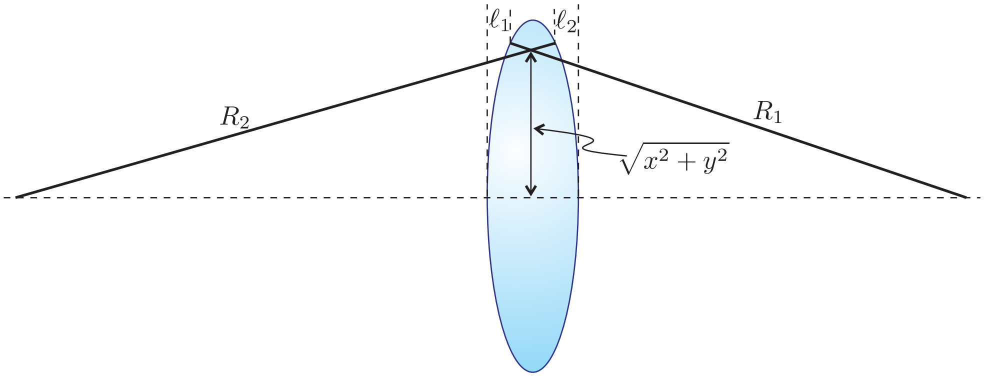


Figure 11.5 A thin lens, which modifies the phase of a field passing through.

through less material and gets ahead of the light traveling through the center of the lens. In (11.21), k represents the wave number in vacuum (i.e. $2\pi/\lambda_{\text{vac}}$).

We can find expressions for ℓ_1 and ℓ_2 from the equations describing the spherical surfaces of the lens:

$$\begin{aligned}(R_1-\ell_1)^2+x^2+y^2&=R_1^2\\(R_2+\ell_2)^2+x^2+y^2&=R_2^2\end{aligned}\tag{11.22}$$

In the Fresnel approximation, the light propagation takes place in the paraxial limit. It is therefore appropriate to neglect the terms ℓ_1^2 and ℓ_2^2 in comparison with the other terms present. Within this approximation, equations (11.22) can be solved, and they render

$$\begin{aligned}\ell_1&\cong\frac{x^2+y^2}{2R_1}\\\ell_2&\cong-\frac{x^2+y^2}{2R_2}\end{aligned}\tag{11.23}$$

We are now able to evaluate the phase advance (11.21) in terms of x and y. Substitution of (11.23) into (11.21) yields

$$\Delta\phi=-k\,(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right)\frac{\left(x^2+y^2\right)}{2}\tag{11.24}$$

As is noticed right away, (11.24) contains the focal length f of a thin lens according to lens-maker's formula (9.56). With this identification, the phase introduced by the lens becomes

$$\Delta\phi=-\frac{k}{2f}\left(x^2+y^2\right)\tag{11.25}$$

In summary, the light traversing a lens experiences a relative phase shift given by

$$E\left(x,y,z_{\text{after lens}}\right)=E\left(x,y,z_{\text{before lens}}\right)e^{-i\frac{k}{2f}\left(x^2+y^2\right)}\tag{11.26}$$

Equation (11.26) introduces a wave-front curvature to the field. For example, if a plane wave (i.e. a uniform field E_0) passes through the lens, the field emerges with a spherical-like wave front converging towards the focus of the lens.

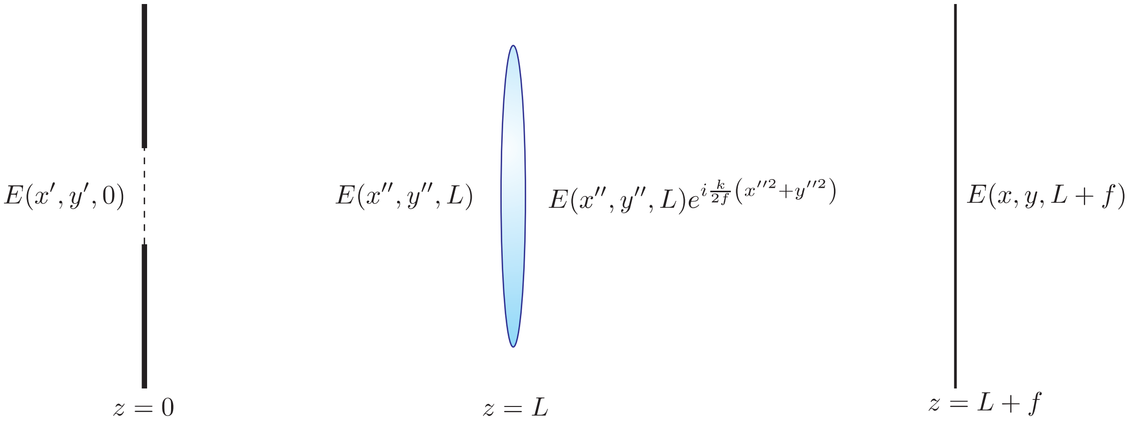


Figure 11.6 Diffraction from an aperture viewed at the focus of a lens.

We now consider the Fresnel diffraction pattern at the focus of a lens inserted a distance L following the aperture (see Fig. 11.6).

Assume that the field $E\left(x', y', 0\right)$ at the aperture is known. We use the Fresnel approximation to compute the field incident on the lens:

$$E(x'', y'', L) = -i\frac{e^{ikL}e^{i\frac{k}{2L}\left(x''^2+y''^2\right)}}{\lambda L}\iint E(x', y', 0)e^{i\frac{k}{2L}\left(x'^2+y'^2\right)}e^{-i\frac{k}{L}\left(x''x'+y''y'\right)}dx'dy' \tag{11.27}$$

(The double primes keep track of distinct variables in the two diffraction problems that are being put together into a system.) Next, the field gains a phase factor according to (11.26) upon transmitting through the lens. Finally, we use the Fresnel diffraction formula to propagate the distance f from the back of the thin lens:

$$E\left(x, y, L+f\right) = -i\frac{e^{ikf}e^{i\frac{k}{2f}\left(x^2+y^2\right)}}{\lambda f}\iint \left[E(x'', y'', L)e^{-i\frac{k}{2f}\left(x''^2+y''^2\right)}\right] \times e^{i\frac{k}{2f}\left(x''^2+y''^2\right)}e^{-i\frac{k}{f}\left(xx''+yy''\right)}dx''dy'' \tag{11.28}$$

As is immediately appreciated by students, the injection of (11.27) into (11.28) makes a rather long formula involving four integrals. Nevertheless, two of the integrals can be performed in advance of choosing the aperture (i.e. those over x'' and y''). This is accomplished with the help of the integral formula (0.52) (even though in this instance the real part of A is zero). After this cumbersome work, (11.28) becomes

$$E\left(x, y, L+f\right) = -i\frac{e^{ik(L+f)}e^{i\frac{k}{2f}\left(x^2+y^2\right)}e^{-i\frac{kL}{2f^2}\left(x^2+y^2\right)}}{\lambda f}\iint E(x', y', 0)e^{-i\frac{k}{f}\left(xx'+yy'\right)}dx'dy' \tag{11.29}$$

Notice that at least the integration portion of this formula looks exactly like the Fraunhofer diffraction formula! This happened even though in the preceding discussion we did not at any time specifically make the Fraunhofer approximation. The result (11.29) implies the intensity distribution (11.20) as anticipated. However, the phase of the field is also revealed in (11.29). In general, the field caries a wave front curvature as it passes through the focal plane of the lens. In the special case $L = f$, the diffraction formula takes a particularly simple form:

$$E(x', y', L + f)\big|_{L=f} = -i\frac{e^{2ikf}}{\lambda f}\iint E(x', y', 0)e^{-i\frac{k}{f}(xx'+yy')}dx'dy' \tag{11.30}$$

When the lens is placed at this special distance following the aperture, the Fraunhofer diffraction pattern viewed at the focus of the lens carries a flat wave front.

11.5 Resolution of a Telescope

In the previous section we learned that the Fraunhofer diffraction pattern appears at the focus of a lens. This has important implications for telescopes and other optical instruments such as the human eye. In essence, any optical instrument involving lenses or mirrors has a built-in aperture, limiting the light that enters. For example, the pupil of the eye acts as an aperture that induces a Fraunhofer diffraction pattern to occur at the retina. Cameras have irises which aperture the light, causing a Fraunhofer diffraction pattern at the film plane. If nothing else, the diameter of the lens itself induces diffraction.

Recall that the Fraunhofer pattern represents the ultimate amount of diffraction caused by an aperture, and this just happens to occur at the focus of any lens. Of course, the focus of the lens is just where one needs to look in order to see images of distant objects. This has the effect of blurring out features in the image, limiting the *resolution.* This illustrates why it is impossible to focus light to a true point.

Suppose you want to image two very distant stars that are close together. An image of each star appears near the focus of the lens. Since the rays traversing the center of the lens from either star are non-deviating (in the thin lens approximation), the angular separation between the two images is the same as the angular separation between the stars. This is seen in Fig. 11.7. A resolution problem occurs when the Fraunhofer diffraction pattern causes each image to blur by more than the angular separation between them. In this case the two images cannot be resolved because they bleed into each other.

The Fraunhofer diffraction pattern from a circular aperture was computed in P 11.6. At the focus of a lens, this pattern is

$$I(\rho, f) = I_0\left(\frac{\pi\ell^2}{4\lambda f}\right)^2\left[2\frac{J_1(k\ell\rho/2f)}{(k\ell\rho/2f)}\right]^2 \tag{11.31}$$

where f is the focal length of the lens and ℓ is its diameter. This pattern contains the first order Bessel function J_1, which behaves somewhat like a sine wave as seen in Fig. 11.8. The main differences are that the zero crossings are not exactly periodic and the function slowly diminishes with larger arguments. The first zero crossing (after $x = 0$) occurs at 1.22π.

The intensity pattern described by (11.31) contains the factor $2J_1(x)/x$, where x represents the combination $k\ell\rho/2f$. As noticed in Fig. 11.8, $J_1(x)$ goes to zero at $x = 0$. Thus, we have a zero-divided-by-zero situation when evaluating $2J_1(x)/x$ at the origin. This is similar to the sinc function (i.e. $\sin(x)/x$), which approaches one at the origin. In fact, $2J_1(x)/x$ is sometimes called the jinc function because it also approaches one at the origin.

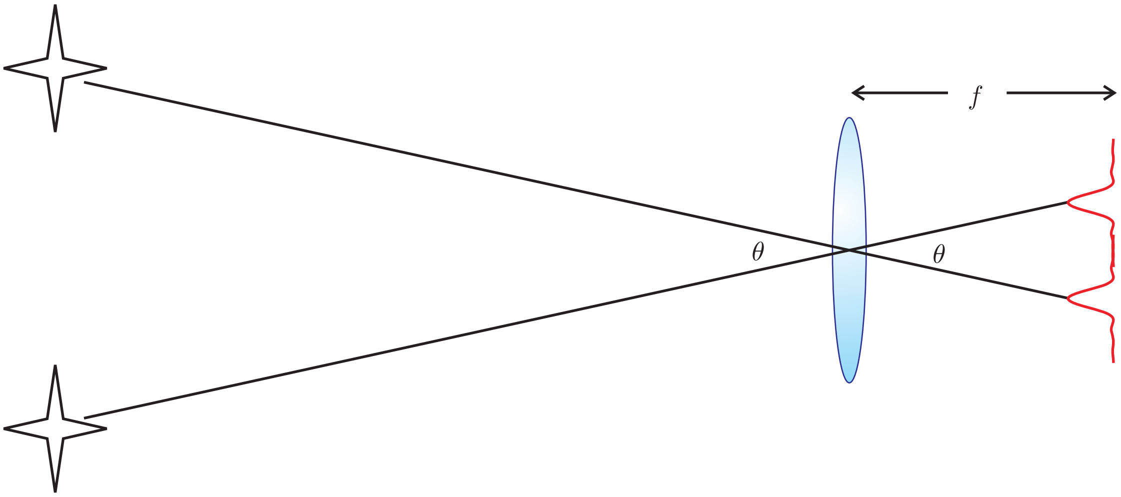


Figure 11.7 To resolve distinct images at the focus of a lens, the angular separation must exceed the width of the Fraunhofer diffraction patterns.

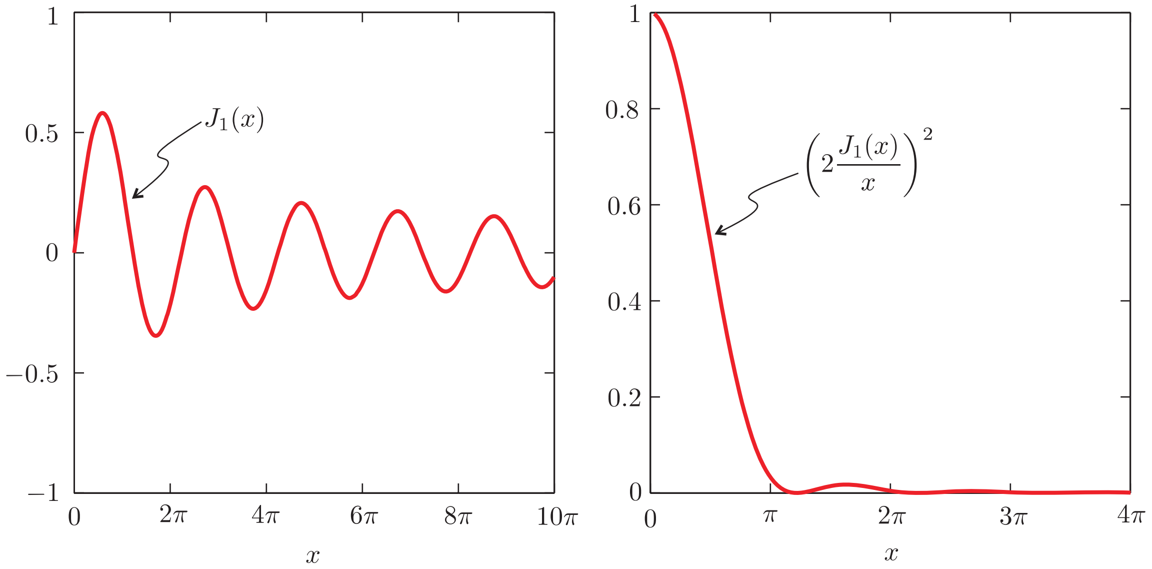


Figure 11.8 (a) First-order Bessel function. (b) Square of the Jinc function.

The square of the jinc is shown in Fig. 11.8b. This curve is proportional to the intensity described in (11.31). This pattern is sometimes called an Airy pattern after Sir George Biddell Airy (English, 18011892) who first described the pattern. As can be seen in Fig. 11.8b, the intensity quickly drops at larger radii.

We now return to the question of whether the images of two nearby stars as depicted in Fig. 11.7 can be distinguished. Since the peak in Fig. 11.8b is the dominant feature in the diffraction pattern, we will say that the two stars are resolved if the angle between them is enough to keep their respective diffraction peaks from seriously overlapping. We will use the criterion suggested by Lord Rayleigh for this purpose. His criterion for well-separated diffraction patterns requires the peak of one pattern to be no closer than the first zero of the other. This situation is shown in Fig. 11.9.

It is straightforward to find the angle that corresponds to this separation of diffraction patterns. Since the width of the diffraction patterns depends on the diameter ℓ of the lens as well as on the wavelength of the light, we expect the minimum angle between resolvable objects to depend on these parameters. To find this angle we set the argument of (11.31) equal to 1.22π, the location of the first zero:

$$\frac{k\ell\rho}{2f} = 1.22\pi \tag{11.32}$$

With a little rearranging we have

$$\theta_{\text{min}} \cong \frac{\rho}{f} = \frac{1.22\lambda}{\ell} \tag{11.33}$$

Here we have associated the ratio ρ/f (i.e. the radius of the diffraction pattern compared to the distance from the lens) with an angle. This angle is a measure of the angular extent of the diffraction pattern. The Rayleigh criterion requires that the diffraction patterns associated with two images be separated by at least this amount before we say that they are resolved. We therefore label the angle as θ_{min}.

11.6 The Array Theorem

In this section we develop the array theorem, which is used for calculating the Fraunhofer diffraction from an array of N identical apertures. The array theorem is remarkable in itself, but our purpose in studying it is for its application to diffraction gratings, discussed in the next section. Conceptually, a grating may be thought of as a mask with an array of identical slits. This is similar to a Young's double-slit setup, only an arbitrarily large number of slits may be used. As far as the array theorem is concerned, however, the apertures can have any shape, as suggested by Fig. 11.10.

Consider N apertures in a mask, each with the identical field distribution described by

$$E_{\text{aperture}}(x', y', 0) \tag{11.34}$$

Each identical aperture has a unique location on the mask. Let the location of the n^{th} aperture be designated by the coordinates (x'_n, y'_n). Since each aperture is identical, we can

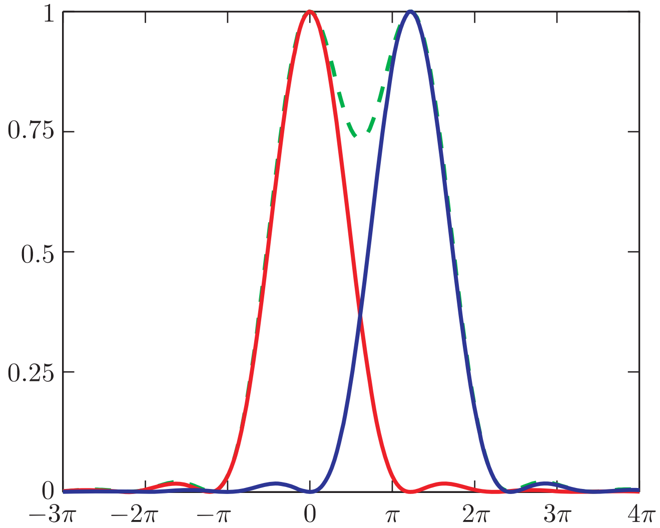


Figure 11.9 The Rayleigh criterion for a circular aperture.

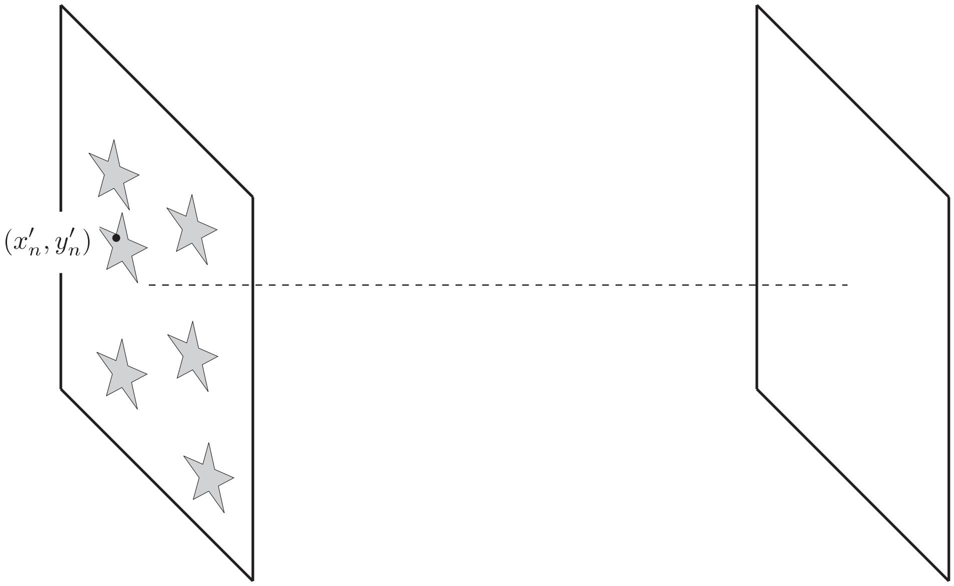


Figure 11.10 Array of identical apertures.

conveniently write the total field by summing over the same individual pattern displaced repeatedly according to the locations of the individual apertures:

$$E\left(x',y',0\right)=\sum_{n=1}^{N}E_{\text{aperture}}(x'-x'_n,y'-y'_n,0) \tag{11.35}$$

Let us compute the Fraunhofer diffraction pattern produced by the field described by (11.35). However, we want to do this in a general case so that we can delay picking the specific aperture shape until a later time (i.e. (11.34)). Upon inserting (11.35) into the Fraunhofer diffraction formula (10.16) we obtain

$$E\left(x,y,z\right)=-i\frac{e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\sum_{n=1}^{N}\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'E_{\text{aperture}}\left(x'-x'_n,y'-y'_n,0\right)e^{-i\frac{k}{z}(xx'+yy')} \tag{11.36}$$

where we have taken the summation out in front of the integral.

To proceed further, let us make the following change of variables:

$$\begin{aligned}x''&\equiv x'-x'_n\\ y''&\equiv y'-y'_n\end{aligned} \tag{11.37}$$

With the use of these new variables, (11.36) becomes

$$E\left(x,y,z\right)=-i\frac{e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\sum_{n=1}^{N}\int_{-\infty}^{\infty}dx''\int_{-\infty}^{\infty}dy''E_{\text{aperture}}\left(x'',y'',0\right)e^{-i\frac{k}{z}[x(x''+x'_n)+y(y''+y'_n)]} \tag{11.38}$$

We next pull a constant factor out of the integrals and we arrive at our final result. With a slight re-arrangement (and with a trivial exchange of x' for x'' and y' for y''), (11.38) can be rewritten as

$$\begin{aligned}E\left(x,y,z\right)=&\left[\sum_{n=1}^{N}e^{-i\frac{k}{z}(xx'_n+yy'_n)}\right]\\ &\times\left[-i\frac{e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'E_{\text{aperture}}\left(x',y',0\right)e^{-i\frac{k}{z}(xx'+yy')}\right]\end{aligned} \tag{11.39}$$

Equation (11.39) is known as the array theorem. Note that the second factor in brackets is exactly the Fraunhofer diffraction pattern from just one of the identical apertures. When more than one identical aperture is present, we need only evaluate the Fraunhofer diffraction formula for an individual aperture. Then, the single-aperture result is multiplied by the summation in front, which contains the information about the number of apertures and their respective positions.

11.7 Diffraction Grating

In this section we will use the array theorem to calculate the diffraction from a grating comprised of an array of equally spaced identical slits. An array of uniformly spaced slits

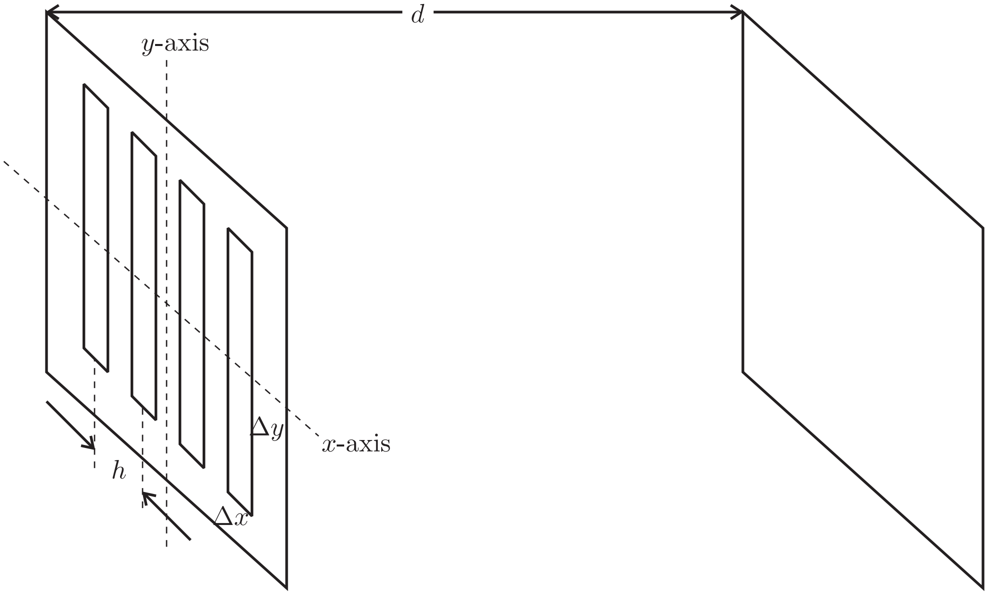


Figure 11.11 Transmission grating.

is called a transmission grating (see Fig. 11.11). Reflection gratings are similar, being composed of an array of narrow rectangular mirrors that behave similarly to the slits.

As was calculated in P 11.5, the Fraunhofer diffraction pattern from a single aperture is given by

$$E_{\text{aperture}}\left(x,y,z\right) = -iE_0 \frac{\Delta x \Delta y e^{ikz}}{\lambda z} e^{i\frac{k}{2z}\left(x^2+y^2\right)} \text{sinc}\left(\frac{\pi \Delta x}{\lambda z} x\right) \text{sinc}\left(\frac{\pi \Delta y}{\lambda z} y\right) \tag{11.40}$$

The only part of (11.39) that remains to be evaluated is the summation out in front. Let the apertures be positioned at

$$x'_n = \left(n - \frac{N+1}{2}\right) h, \qquad y'_n = 0 \tag{11.41}$$

where N is the total number of slits. Then the summation in the array theorem, (11.39), becomes

$$\sum_{n=1}^{N} e^{-i\frac{k}{z}\left(xx'_n + yy'_n\right)} = e^{i\frac{khx}{z}\left(\frac{N+1}{2}\right)} \sum_{n=1}^{N} e^{-i\frac{khx}{z}n} \tag{11.42}$$

This summation is recognized as a geometric sum, which can be performed using formula (0.59).

Equation (11.42) then simplifies to

$$\begin{aligned}\sum_{n=1}^{N} e^{-i\frac{k}{z}(xx'_n+yy'_n)} &= e^{i\frac{k}{z}\left(\frac{N+1}{2}\right)xh}e^{-i\frac{khx}{z}}\frac{e^{-i\frac{khx}{z}N}-1}{e^{-i\frac{khx}{z}}-1}\\ &= \frac{e^{-i\frac{khx}{2z}N}-e^{i\frac{khx}{2z}N}}{e^{-i\frac{khx}{2z}}-e^{i\frac{khx}{2z}}}\\ &= \frac{\sin\left(N\frac{khx}{2z}\right)}{\sin\left(\frac{khx}{2z}\right)}\end{aligned} \tag{11.43}$$

By combining (11.40) and (11.43) we obtain the full Fraunhofer diffraction pattern for a diffraction grating. The expression for the field is

$$E\left(x,y,z\right)=\frac{\sin\left(N\frac{khx}{2z}\right)}{\sin\left(\frac{khx}{2z}\right)}\left[-iE_0\frac{\Delta x\Delta y e^{ikz}}{\lambda z}e^{i\frac{k}{2z}\left(x^2+y^2\right)}\text{sinc}\left(\frac{\pi\Delta x}{\lambda z}x\right)\text{sinc}\left(\frac{\pi\Delta y}{\lambda z}y\right)\right] \tag{11.44}$$

Let's consider a grating with the slits oriented in the y-direction, and $\Delta y \gg \lambda$ so that the last sinc function in Eq. (11.44) goes to one.[1] The intensity pattern in the horizontal direction can then be written in terms of the peak intensity of the diffraction pattern on the screen:

$$I\left(x\right)=I_{\text{peak}}\text{sinc}^2\left(\frac{\pi\Delta x}{\lambda z}x\right)\frac{\sin^2\left(N\frac{\pi hx}{\lambda z}\right)}{N^2\sin^2\left(\frac{\pi hx}{\lambda z}\right)} \tag{11.45}$$

Note that $\lim\limits_{\alpha\to 0}\frac{\sin N\alpha}{\sin\alpha}=N$ so we have placed N^2 in the denominator when introducing our definition of I_{peak}, which represents the intensity on the screen at $x=0$. In principle, the intensity I_{peak} is a function of y and depends on the exact details of how the slits are illuminated as a function of y, but this is usually not of interest as long as we stay with a given value of y as we scan along x.

It is left as an exercise to study the functional form of (11.45), especially how the number of slits N influences the behavior. The case of $N=2$ describes the diffraction pattern for a Young's double slit experiment. We now have a description of the Young's two-slit pattern in the case that the slits have finite openings of width Δx rather than infinitely narrow ones.

A final note: You may wonder why we are interested in Fraunhofer diffraction from a grating. The reason is that we are actually interested in separating different wavelengths by observing their distinct diffraction patterns separated in space. In order to achieve good spatial separation between light of different wavelengths, it is necessary to allow the light to propagate a far distance. Optimal separation (the maximum possible) occurs therefore in the Fraunhofer regime.

[1] This is mostly the right idea, but is still a bit of a fake. In fact, the field often does not have a uniform phase along the entire slit in the y-dimension, so our use of the function sinc $[(\pi\Delta y/\lambda z)\,y]$ was inappropriate to begin with. The energy in a real spectrometer is usually spread out in a diffuse pattern in the y-dimension. However, its form in y is of little relevance; the spectral information is carried in the x-dimension only.

11.8 Spectrometers

The formula (11.45) can be exploited to make wavelength measurements. This forms the basis of a diffraction grating spectrometer. A spectrometer has relatively poor resolving power compared to a Fabry-Perot interferometer. Nevertheless, a spectrometer is not hampered by the serious limitation imposed by free spectral range. Therefore, it is able to measure a wide range of wavelengths simultaneously. The Fabry-Perot interferometer and the grating spectrometer in this sense are complementary, the one being able to make very precise measurements within a narrow wavelength range and the other being able to characterize wide ranges of wavelengths simultaneously.

To appreciate how a spectrometer works, consider the Fraunhofer diffraction from a grating, described by (11.45). The structure of the diffraction pattern gives rise to peaks. For example, Fig. 11.12a shows the diffraction peaks from a Young's double slit (i.e. $N = 2$). The diffraction pattern is comprised of the typical Young's double-slit pattern multiplied by the diffraction pattern of a single slit, according to the array theorem. (Note that $\sin^2\left(2\frac{\pi hx}{\lambda z}\right)/\sin^2\left(\frac{\pi hx}{\lambda z}\right) = 4\cos^2\left(\frac{\pi hx}{\lambda z}\right)$.)

As the number of slits N is increased, the peaks seen in the Young's double-slit pattern tend to sharpen with additional smaller peaks appearing in between. Figure 11.12b shows the case for $N = 5$. The more significant peaks occur when $\sin(\pi hx/\lambda z)$ in the denominator of (11.45) goes to zero. Keep in mind that the numerator goes to zero at the same places, creating a zero-over-zero situation, so the peaks are not infinitely tall.

With larger values of N, the peaks can become extremely sharp, and the small secondary peaks in between are smaller in comparison. Fig. 11.12c shows the case of $N = 10$ and Fig. 11.12d, shows the case of $N = 100$.

When very many slits are used, the diffraction pattern becomes very useful for measuring spectra of light. Keep in mind that the position of the diffraction peaks depends on wavelength (except for the center peak at $x = 0$). If light of different wavelengths is simultaneously present, then the diffraction peaks associated with different wavelengths appear in different locations. It helps to have very many slits involved (i.e. large N) so that the diffraction peaks are sharply defined. Then closely spaced wavelengths can be more easily distinguished.

Consider the inset in Fig. 11.12d, which gives a close-up view of the first-order diffraction peak for $N = 100$. The location of this peak on a distant screen varies with the wavelength of the light. How much must the wavelength change to cause the peak to move by half of its "width" as marked in the inset of Fig. 11.12d? We will say that this is the minimum separation of wavelengths that still allows the two peaks to be distinguished. Let us solve for this minimum distinguishable wavelength difference.

As mentioned, the main diffraction peaks occur when the denominator of (11.45) [i.e. $\sin^2(\pi hx/\lambda z)$] goes to zero. The location of the m^{th} peak is therefore located at

$$\frac{\pi hx}{\lambda_0 z} = m\pi \Rightarrow \lambda_0 = \frac{hx}{mz} \tag{11.46}$$

The numerator of (11.45) $\sin^2(N\pi hx/\lambda z)$ also goes to zero at this same location, so the expression avoids going to infinity. The first zero to the sides of the main peaks (see

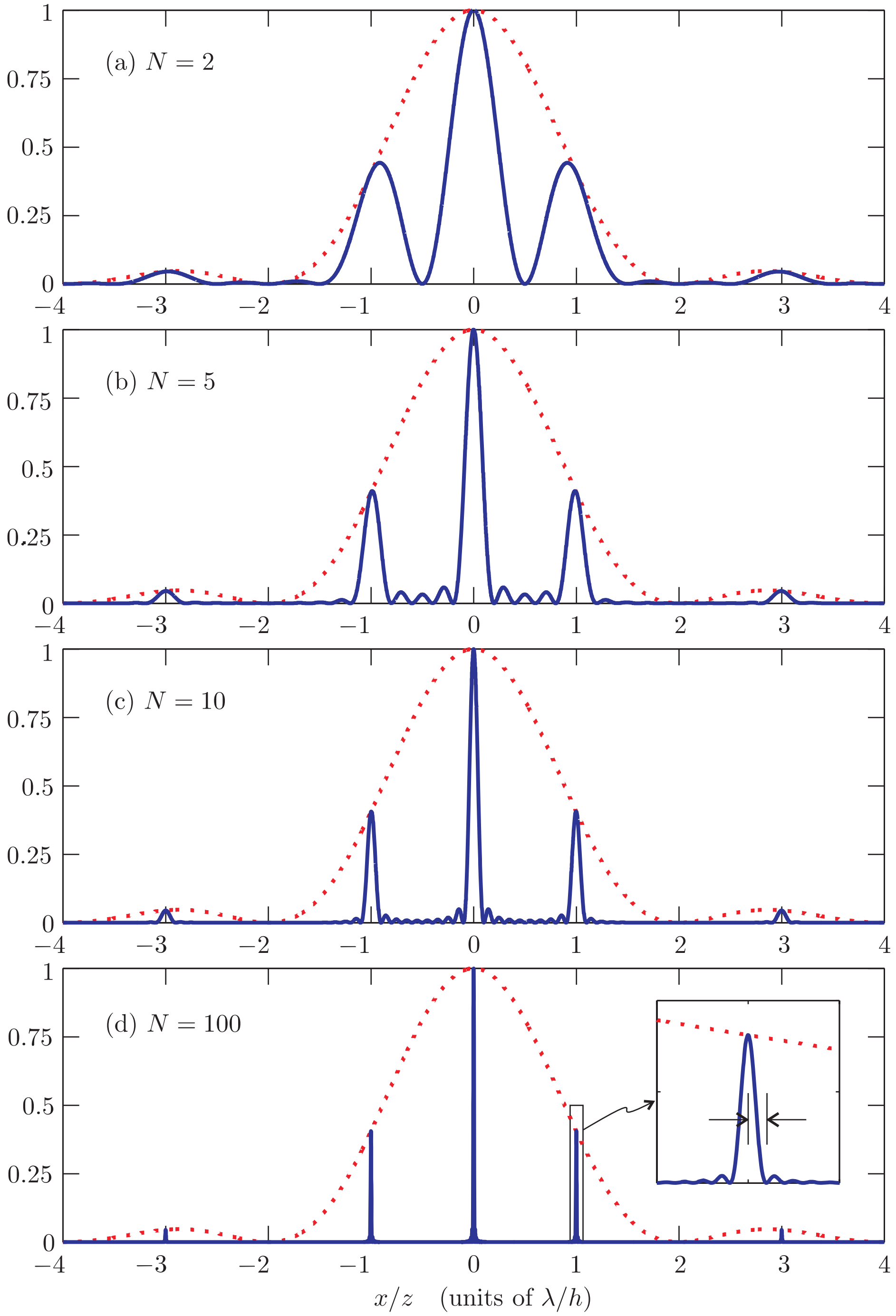


Figure 11.12 Diffraction through various numbers of slits, each with $\Delta x = h/2$ (slit widths half the separation). The dotted line shows the single slit diffraction pattern. (a) Diffraction from a double slit. (b) Diffraction from 5 slits. (c) Diffraction from 10 slits. (d) Diffraction from 100 slits.

Fig. 11.12d) occurs when

$$N\frac{k'hx}{2z} = Nm\pi + \pi \Rightarrow \lambda_0 - \Delta\lambda = N\frac{hx}{(Nm+1)\,z} \tag{11.47}$$

The wavelength difference that shifts the peak by this amount (from peak center to the adjacent zero) is then

$$\Delta\lambda = \frac{hx}{mz} - \frac{Nhx}{(Nm+1)\,z} = \frac{(Nm+1) - Nm}{(Nm+1)}\frac{hx}{mz} \cong \frac{hx}{Nm^2 z} = \frac{\lambda_0}{Nm} \tag{11.48}$$

This is the minimum difference in wavelength that we can hope to distinguish if two peaks of the different wavelengths are together side by side.

As we did for the Fabry-Perot interferometer, we can define the resolving power of the diffraction grating as

$$RP \equiv \frac{\lambda}{\Delta\lambda} = mN \tag{11.49}$$

The resolving power is proportional to the number of slits illuminated on the diffraction grating. The resolving power also improves by using larger diffraction orders m.

Appendix 11.A ABCD Law for Gaussian Beams

In this section we discuss and justify the ABCD law for Gaussian beams. The law enables one to predict the parameters of a Gaussian beam that exits from an optical system, given the parameters of an input Gaussian beam. To make the prediction, one needs only the ABCD matrix for the optical system, taken as a whole. The system may be arbitrarily complex with many optical components.

At first, it may seem unlikely that such a prediction should be possible since ABCD matrices were introduced to describe the propagation of rays. On the other hand, Gaussian beams are governed by the laws of diffraction. As an example of this dichotomy, consider a collimated Gaussian beam that traverses a converging lens. By ray theory, one expects the Gaussian beam to focus near the focal point of the lens. However, a collimated beam by definition is already in the act of going through focus. In the absence of the lens, there is a tendency for the beam to grow via diffraction, especially if the beam waist is small. This tendency competes with the focusing effect of the lens, and a new beam waist can occur at a wide range of locations, depending on the exact outcome of this competition.

A Gaussian beam is characterized by its Rayleigh range z_0. From this, the beam waist radius w_0 may be extracted via (11.13), assuming the wavelength is known. Suppose that a Gaussian beam encounters an optical system at position z, referenced to the position of the beam's waist as shown in Fig. 11.13. The beam exiting from the system, in general, has a new Rayleigh range z_0'. The waist of the new beam also occurs at a different location. Let z' denote the location of the exit of the optical system, referenced to the location of the waist of the new beam. If the exiting beam diverges as in Fig. 11.13, then it emerges from a *virtual* beam waist located before the exit point of the system. In this case, z' is taken to be positive. On the other hand, if the emerging beam converges to an actual waist, then z' is taken to be negative since the exit point of the system occurs before the focus.

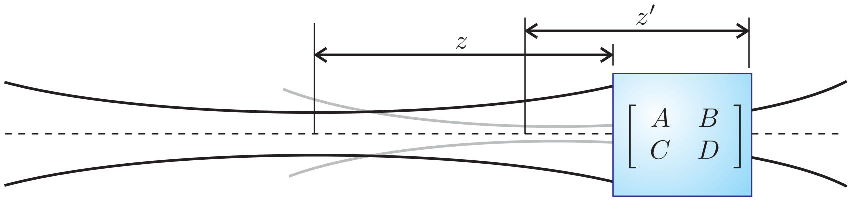


Figure 11.13 Gaussian laser beam traversing an optical system described by an ABCD matrix. The dark lines represent the incoming and exiting beams. The gray line represents where the exiting beam *appears* to have been.

The ABCD law is embodied in the following relationship:

$$z' - iz'_0 = \frac{A\,(z - iz_0) + B}{C\,(z - iz_0) + D} \tag{11.50}$$

where A, B, C, and D are the matrix elements of the optical system. The imaginary number $i \equiv \sqrt{-1}$ imbues the law with complex arithmetic. It makes two equations from one, since the real and imaginary parts of (11.50) must separately be equal.

We now prove the ABCD law. We begin by showing that the law holds for two specific ABCD matrixes. First, consider the matrix for propagation through a distance d:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \tag{11.51}$$

We know that simple propagation has minimal effect on a beam. The Rayleigh range is unchanged, so we expect that the ABCD law should give $z'_0 = z_0$. The propagation through a distance d modifies the beam position by $z' = z + d$. We now check that the ABCD law agrees with these results by inserting (11.51) into (11.50):

$$z' - iz'_0 = \frac{1\,(z - iz_0) + d}{0\,(z - iz_0) + 1} = z + d - iz_0 \qquad \text{(propagation through distance d)} \tag{11.52}$$

Thus, the law holds in this case.

Next we consider the ABCD matrix of a thin lens (or a curved mirror):

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \tag{11.53}$$

A beam that traverses a thin lens undergoes the phase shift $-k\rho^2/2f$, according to (11.26). This modifies the original phase of the wave front $k\rho^2/2R\,(z)$, seen in (11.9). The phase of the exiting beam is therefore

$$\frac{k\rho^2}{2R\,(z')} = \frac{k\rho^2}{2R\,(z)} - \frac{k\rho^2}{2f} \tag{11.54}$$

where we do not keep track of unimportant overall phases such as kz or kz'. With (11.12) this relationship reduces to

$$\frac{1}{R\,(z')} = \frac{1}{R\,(z)} - \frac{1}{f} \Rightarrow \frac{1}{z' + z'^2_0/z'} = \frac{1}{z + z_0^2/z} - \frac{1}{f} \tag{11.55}$$

In addition to this relationship, the local radius of the beam given by (11.11) cannot change while traversing the "thin" lens. Therefore,

$$w\left(z'\right) = w\left(z\right) \Rightarrow z'_0\left(1+\frac{z'^2}{z'^2_0}\right) = z_0\left(1+\frac{z^2}{z_0^2}\right) \tag{11.56}$$

On the other hand, the ABCD law for the thin lens gives

$$z' - iz'_0 = \frac{1\left(z-iz_0\right)+0}{-\left(1/f\right)\left(z-iz_0\right)+1} \qquad \text{(traversing a thin lens with focal length f)} \tag{11.57}$$

It is left as an exercise (see P 11.18) to show that (11.57) is consistent with (11.55) and (11.56).

So far we have shown that the ABCD law works for two specific examples, namely propagation through a distance d and transmission through a thin lens with focal length f. From these elements we can derive more complicated systems. However, the ABCD matrix for a thick lens cannot be constructed from just these two elements. However, we can construct the matrix for a thick lens if we sandwich a thick *window* (as opposed to empty space) between two thin lenses. The proof that the matrix for a thick window obeys the ABCD law is left as an exercise (see P 11.21). With these relatively few elements, essentially any optical system can be constructed, provided that the beam propagation begins and ends up in the same index of refraction.

To complete our proof of the general ABCD law, we need only show that when it is applied to the compound element

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} = \begin{bmatrix} A_2A_1+B_2C_1 & A_2B_1+B_2D_1 \\ C_2A_1+D_2C_1 & C_2B_1+D_2D_1 \end{bmatrix} \tag{11.58}$$

it gives the same answer as when the law is applied sequentially, first on

$$\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix}$$

and then on

$$\begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}$$

Explicitly, we have

$$\begin{aligned} z'' - iz''_0 &= \frac{A_2\left(z'-iz'_0\right)+B_2}{C_2\left(z'-iz'_0\right)+D_2} \\ &= \frac{A_2\left[\frac{A_1(z-iz_0)+B_1}{C_1(z-iz_0)+D_1}\right]+B_2}{C_2\left[\frac{A_1(z-iz_0)+B_1}{C_1(z-iz_0)+D_1}\right]+D_2} \\ &= \frac{A_2\left[A_1\left(z-iz_0\right)+B_1\right]+B_2\left[C_1\left(z-iz_0\right)+D_1\right]}{C_2\left[A_1\left(z-iz_0\right)+B_1\right]+D_2\left[C_1\left(z-iz_0\right)+D_1\right]} \\ &= \frac{\left(A_2A_1+B_2C_1\right)\left(z-iz_0\right)+\left(A_2B_1+B_2D_1\right)}{\left(C_2A_1+D_2C_1\right)\left(z-iz_0\right)+\left(C_2B_1+D_2D_1\right)} \\ &= \frac{A\left(z-iz_0\right)+B}{C\left(z-iz_0\right)+D} \end{aligned} \tag{11.59}$$

Thus, we can construct any ABCD matrix that we wish from matrices that are known to obey the ABCD law. The resulting matrix also obeys the ABCD law.

Exercises

11.3 Gaussian Laser Beams

P11.1 (a) Confirm that (11.9) reduces to (11.1) when $z = 0$.

(b) Take the limit $z \gg z_0$ to find the field far from the laser focus.

(c) Define the ratio of z to the (far-away) beam diameter as the f-number:

$$f^{\#} \equiv \lim_{z \to \pm\infty} \frac{z}{2w(z)})$$

Write the beam waist w_0 in terms of the f-number and the wavelength.

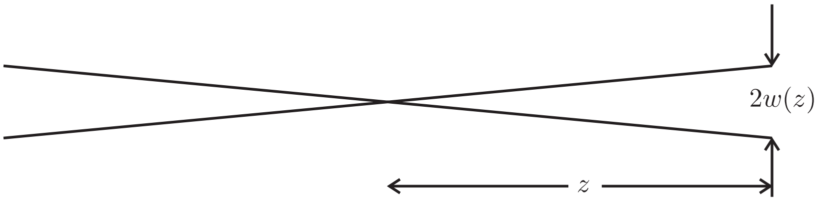


Figure 11.14

NOTE: You now have a convenient way to predict the size of a laser focus by measuring the cone angle of the beam. However, in an experimental setting you may be very surprised at how badly a beam focuses compared to the theoretical prediction (due to aberrations, etc.). It is always good practice to actually measure your focus if its size is important to the experiment.

P11.2 Use the Fraunhofer integral formula (either (10.16) or (10.24)) to determine the far-field pattern of a Gaussian laser focus (11.1).

HINT: The answer should agree with P 11.1 part (b).

L11.3 Consider the following setup where a diverging laser beam is collimated using an uncoated lens. A double reflection from both surfaces of the lens (known as a ghost) comes out in the forward direction, focusing after a short distance. Use a CCD camera to study this focused beam. The collimated beam serves as a reference to reveal the phase of the focused beam through interference. Because the weak ghost beam concentrates near its focus, the two beams can have similar intensities for optimal interference effects.

The ghost beam $E_1(\rho, z)$ is described by (11.9), where the origin is at the focus. Let the collimated beam be approximated as a plane wave $E_2 e^{ikz+i\phi}$, where ϕ is the relative phase between the two beams. The net intensity is then $I_t(\rho, z) \propto \left|E_1(\rho, z) + E_2 e^{ikz+i\phi}\right|^2$ or

$$I_t(\rho, z) = \left[I_2 + I_1(\rho, z) + 2\sqrt{I_2 I_1(\rho, z)} \cos\left(\frac{k\rho^2}{2R(z)} - \tan^{-1}\frac{z}{z_0} - \phi\right)\right]$$

where $I_1(\rho, z)$ is given by (11.14). We now have a formula that retains both $R(z)$ and the Gouy shift $\tan^{-1} z/z_0$, which are not present in the intensity distribution of a single beam (see (11.14)).

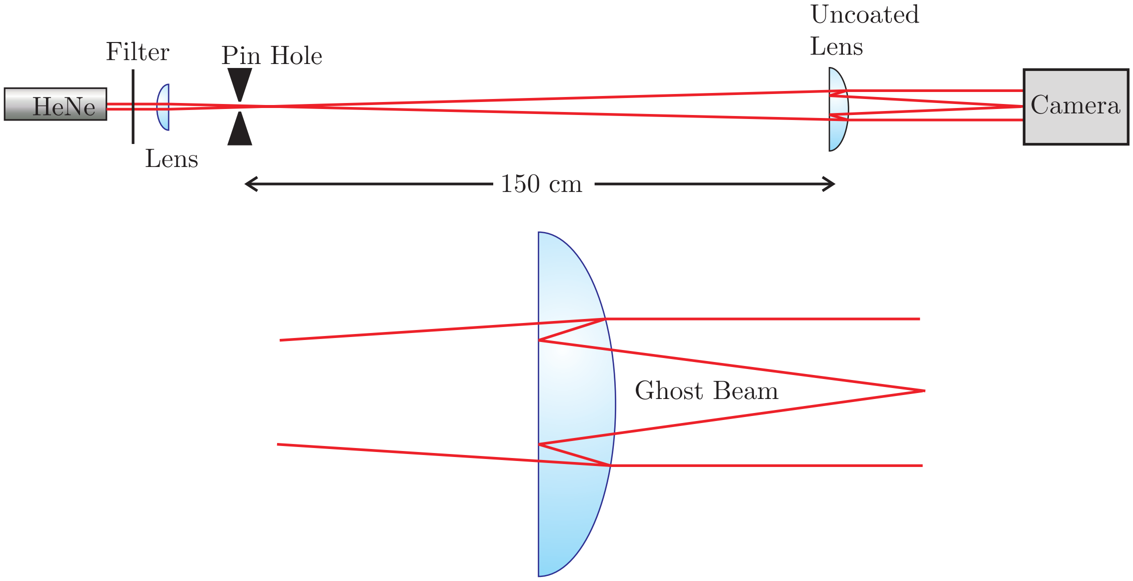


Figure 11.15

(a) Determine the f-number for the ghost beam (see P 11.1 part (c)). Use this measurement to predict a value for w_0. HINT: You know that at the lens, the focusing beam is the same size as the collimated beam.

(b) Measure the actual spot size w_0 at the focus. How does it compare to the prediction?

HINT: Before measuring the spot size, make a minor adjustment to the tilt of the lens. This controls the relative phase between the two beams, which you will set to $\phi = \pm\pi/2$ so that *at the focus* the cosine term vanishes and the two beams don't interfere. This is accomplished if the center of the interference pattern is as dark as possible either far before or far after the focus. In this case, the intensity of the individual beams at the focus simply added together (only at $z = 0$), the small profile "on top" of the wave profile.

(c) Observe the effect of the Gouy shift. Since $\tan^{-1} z/z_0$ varies over a range of π, you should see that the ring pattern inverts before and after the focus. The bright rings exchange with the dark ones.

(d) Predict the Rayleigh range z_0 and check that the radius of curvature $R(z) \equiv z + z_0^2/z$ agrees with measurement.

HINT: As you look at different radii ρ, the only interference term that varies is $k\rho^2/2R(z)$. If you count N fringes out to a radius ρ, then $k\rho^2/2R(z)$ has varied by $2\pi N$. You can then compute $R(z)$ and compare it to the prediction. You should see pictures like the following:

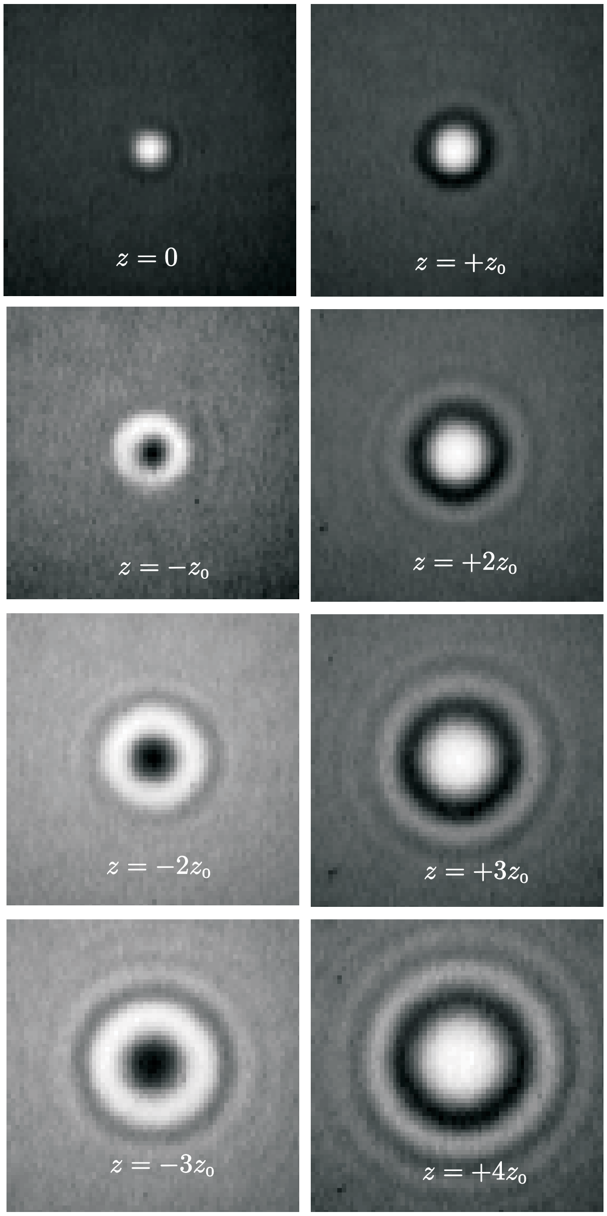


Figure 11.16

11.4 Fraunhofer Diffraction Through a Lens

P11.4 Fill in the steps leading to (11.29) from (11.28). Show that the intensity distribution (11.20) is consistent (11.29).

P11.5 Calculate the Fraunhofer diffraction *field* and *intensity* patterns for a rectangular aperture (dimensions Δx by Δy) illuminated by a plane wave E_0.

HINT: Use (10.16)

$$E(x,y,z) = -iE_0 \frac{e^{ikz}}{\lambda z} e^{i\frac{k}{2z}(x^2+y^2)} \int_{-\Delta x/2}^{\Delta x/2} dx' e^{-i\frac{kx}{z}x'} \int_{-\Delta y/2}^{\Delta y/2} dy' e^{-i\frac{ky}{z}y'}$$

Answer: $I(x,y,z) = I_0 \frac{\Delta x^2 \Delta y^2}{\lambda^2 z^2} \operatorname{sinc}^2\left(\frac{\pi \Delta x}{\lambda z} x\right) \operatorname{sinc}^2\left(\frac{\pi \Delta y}{\lambda z} y\right)$

P11.6 Calculate the Fraunhofer diffraction *intensity* pattern for a circular aperture (diameter ℓ) illuminated by a plane wave E_0.

HINT: Use (10.24) and (0.55).

Answer: $I(\rho,z) = I_0 \left(\frac{\pi \ell^2}{4\lambda z}\right)^2 \left[2\frac{J_1(k\ell\rho/2z)}{(k\ell\rho/2z)}\right]^2$. The function $\frac{2J_1(x)}{x}$ (sometimes called the jinc function) looks similar to the sinc function except that its first zero is at $x = 1.22\pi$ rather than at π. Note that $\lim_{x\to 0} \frac{2J_1(x)}{x} = 1$.

L11.7 Set up a collimated "plane wave" in the laboratory using a HeNe laser ($\lambda =$ 633 nm) and appropriate lenses.

(a) Choose a rectangular aperture (Δx by Δy) and place it in the plane wave. Observe the Fraunhofer diffraction on a very far away screen (i.e., where $z \gg \frac{k}{2}\,(\text{aperture radius})^2$ is satisfied). Check that the location of the "zeros" agrees with the result from P 11.5.

(b) Place a lens in the beam after the aperture. Use a CCD camera to observe the Fraunhofer diffraction profile at the focus of the lens. Check that the location of the "zeros" agrees with the result from P 11.5, replacing z with f.

(c) Repeat parts (a) and (b) using a circular aperture with diameter ℓ. Check the position of the first "zero."

11.5 Resolution of a Telescope

P11.8 (a) What minimum telescope diameter would be required to distinguish a Jupiter-like planet (orbital radius 8×10^8 km) from its star if they are 10 light-years away? Take the wavelength to be $\lambda = 500$ nm. NOTE: The unequal brightness is the biggest technical challenge.

(b) On the night of April 18, 1775, a signal was sent from the Old North Church steeple to Paul Revere, who was 1.8 miles away: "One if by land, two if by sea." If in the dark, Paul's pupils had 4 mm diameters, what is the minimum possible separation between the two lanterns that would allow him to correctly interpret the signal? Assume that the predominant wavelength of the lanterns was 580 nm.

HINT: In the eye, the index of refraction is about 1.33 so the wavelength is shorter. This leads to a smaller diffraction pattern on the retina. However, in accordance with Snell's law, two rays separated by an angle 580 nm outside of the eye are separated by an angle $\theta/1.33$ inside the eye. The two rays then hit on the retina closer together. As far as resolution is concerned, the two effects exactly compensate.

L11.9 Simulate two stars with laser beams ($\lambda = 633$ nm). Align them nearly parallel with a small lateral displacement. Send the beams down a long corridor until diffraction causes both beams to grow into one another so that it is no longer apparent that they are from two distinct sources. Use a lens to image the two sources onto a CCD camera. The camera should be placed close to the focal plane of the lens. Use a variable iris near the lens to create different pupil openings.

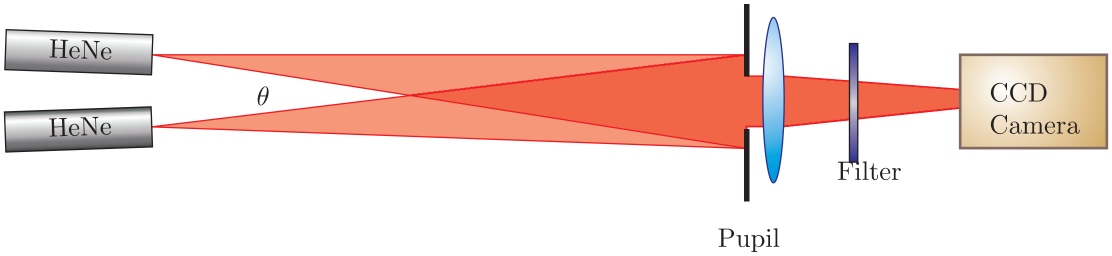


Figure 11.17

Experimentally determine the pupil diameter that just allows you to resolve the two sources according to the Rayleigh criterion. Check your measurement against theoretical prediction.

HINT: The angular separation between the two sources is obtained by dividing propagation distance into the lateral separation of the beams.

P11.10 (a) A monochromatic plane wave with intensity I_0 and wavelength λ is incident on a circular aperture of diameter ℓ followed by a lens of focal length f. Write the intensity distribution at a distance f behind the lens.

(b) You wish to *spatially filter* the beam such that, when it emerges from the focus, it varies smoothly without diffraction rings or hard edges. A pinhole is placed at the focus, which transmits only the central portion of the Airy pattern (inside of the first zero). Calculate the intensity pattern at a distance f after the pinhole using the approximation given in the hint below.

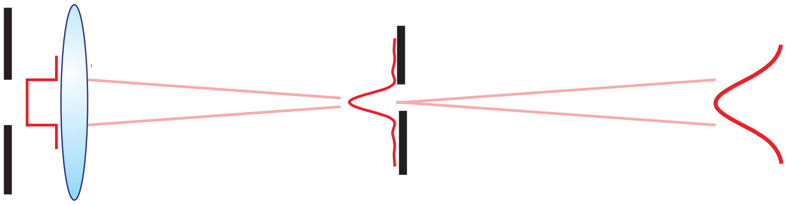


Figure 11.18

HINT: A reasonably good approximation of the transmitted field is that of a Gaussian $E(\rho, 0) = E_f e^{-\rho^2/w_0^2}$, where E_f is the magnitude of the field at the center of the focus found in part (a), and the width is $w_0 = 2\lambda f^{\#}/\pi$ and $f^{\#} \equiv f/\ell$. The figure below shows how well the Gaussian approximation fits the actual curve. We have assumed that the first aperture is a distance f before the lens so that at the focus after the lens the wave front is flat at the pinhole. To avoid integration, you may want to use the result of P 11.2 or P 11.1(b) to get the Fraunhofer limit of the Gaussian profile. (See figure below.)

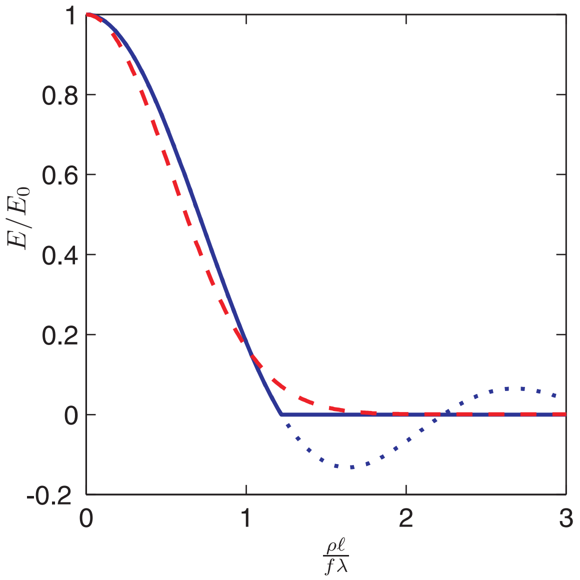


Figure 11.19

11.6 The Array Theorem

P11.11 Find the diffraction pattern created by an array of nine circles, each with radius a, which are centered at the following (x', y') coordinates: $(-b, b)$, $(0, b)$, (b, b), $(-b, 0)$, $(0, 0)$, $(b, 0)$, $(-b, -b)$, $(0, -b)$, $(b, -b)$ (a is less than b). Make a plot of the result for the situation where (in some choice of units) $a = 1$, $b = 5a$, and $k/d = 1$. View the plot at different "zoom levels" to see the finer detail.

P11.12 A diffraction screen with apertures as arranged in Fig. 11.20 is illuminated with a plane wave. All the circles are of radius b, and the square which is centered in the upper circle, is of side a. Light comes through all the circles and the square, but in the shaded regions labeled with "π" the light coming through is shifted to be 180° out of phase with light coming through the other regions. (Thus the square is 180° out of phase with the rest of the upper circle, and the left circle is 180° out of phase with the right circle.) The distances c and L as indicated below are also given.

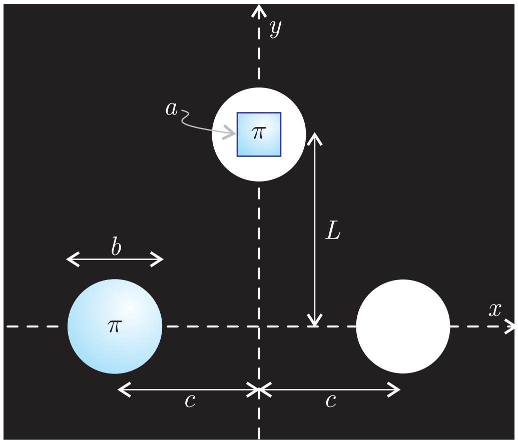


Figure 11.20

(a) Find the fair-field (Fraunhofer) diffraction pattern for the upper aperture alone-that is, the circle-square combination.

(b) Find the fair-field (Fraunhofer) diffraction pattern for the lower two apertures alone (omitting the upper square-circle combination)

(c) Find the diffraction pattern for all the apertures together.

11.7 Diffraction Grating

P11.13 Consider Fraunhofer diffraction from a grating of N slits having widths Δx and equal separations h. Make plots (label relevant points and scaling) of the intensity pattern for $N = 1$, $N = 2$, $N = 5$, and $N = 1000$ in the case where $h = 2\Delta x$, $\Delta x = 5\ \mu\text{m}$, and $\lambda = 500$ nm. Let the Fraunhofer diffraction be observed at the focus of a lens with focal length $f = 100$ cm. Do you expect I_{peak} to be the same value for all of these cases?

P11.14 For the case of $N = 1000$ in P 11.13, you wish to position a narrow slit at the focus of the lens so that it transmits only the first-order diffraction peak (i.e. at $khx/(2f) = \pm\pi$). (a) How wide should the slit be if it is to be *half* the separation between the first intensity zeros to either side of the peak?

(b) What small change in wavelength (away from $\lambda = 500$ nm) will cause the intensity peak to shift by the width of the slit found in part (a)?

P11.15 (a) A plane wave is incident on a screen of N^2 uniformly spaced identical rectangular apertures of dimension Δx by Δy (see figure below). Their positions are described by $x_n = h\left(n - \frac{N+1}{2}\right)$ and $y_m = s\left(m - \frac{N+1}{2}\right)$. Find the far-field (Fraunhofer) pattern of the light transmitted by the grid.

(b) You are looking at a distant sodium street lamp (somewhat monochromatic) through a curtain made from a fine mesh fabric with crossed threads. Make a sketch of what you expect to see (how the lamp will look to you).

HINT: Remember that the lens of your eye causes the Fraunhofer diffraction of the mesh to appear at the retina.

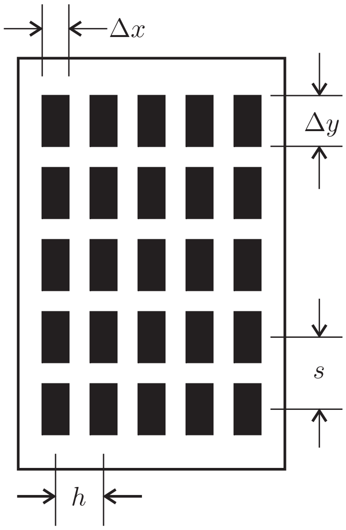


Figure 11.21

11.8 Spectrometers

L11.16 (a) Use a HeNe laser to determine the period h of a reflective grating.

(b) Give an estimate of the blaze angle ϕ on the grating. HINT: Assume that the blaze angle is optimized for first-order diffraction of the HeNe laser (on one side). The blaze angle enables a mirror-like reflection of the diffracted light on each groove.

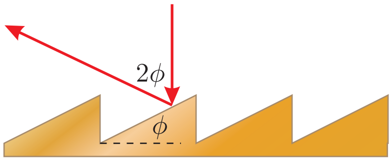


Figure 11.22

(c) You have two mirrors of focal length 75 cm and the reflective grating in the lab. You also have two very narrow adjustable slits and the ability to "tune" the angle of the grating. Sketch how to use these items to make a monochromator (scans through one wavelength at a time). If the beam that hits the grating is 5 cm wide, what do you expect the ultimate resolving power of the monochromator to be in the wavelength range of 500 nm? Do not worry about aberration such as astigmatism from using the mirrors off axis.

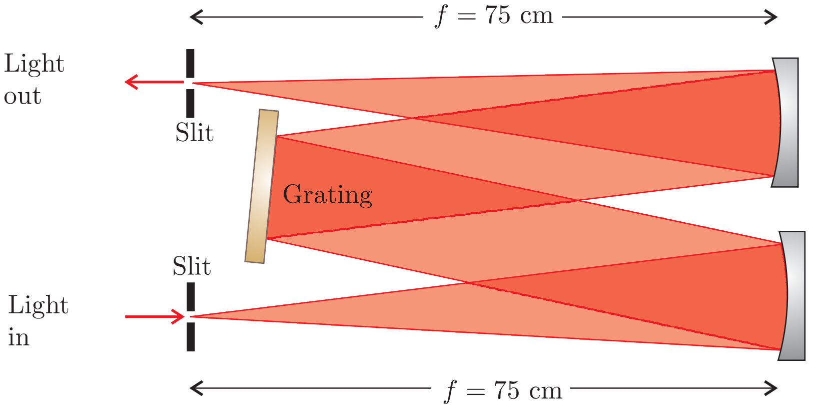


Figure 11.23

L11.17 Study the Jarrell Ash monochromator. Use a tungsten lamp as a source and observe how the instrument works by taking the entire top off. Do not breathe or touch when you do this. In the dark, trace the light inside of the instrument with a white plastic card and observe what happens when you change the wavelength setting. Place the top back on when you are done.

(a) Predict the best theoretical resolving power that this instrument can do assuming 1200 lines per millimeter.

(b) What should the width Δx of the entrance and exit slits be to obtain this resolving power? Assume $\lambda = 500$ nm.

HINT: Set Δx to be the distance between the peak and the first zero of the diffraction pattern at the exit slit for monochromatic light.

11.A ABCD Law for Gaussian Beams

P11.18 Find the solutions to (11.57) (i.e. find z' and z'_0 in terms of z and z_0). Show that the results are in agreement with (11.55) and (11.56).

P11.19 Assuming a collimated beam (i.e. $z = 0$ and beam waist w_0), find the location $L = -z'$ and size w'_0 of the resulting focus when the beam goes through a thin lens with focal length f.

L11.20 Place a lens in a HeNe laser beam soon after the exit mirror of the cavity. Characterize the focus of the resulting laser beam, and compare the results with the expressions derived in P 11.19.

P11.21 Prove the ABCD law for a beam propagating through a thick window of material with matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix}$$

Review, Chapters 9–11

True and False Questions

R48 T or F: The eikonal equation and Fermat's principle depend on the assumption that the wavelength is relatively small compared to features of interest.

R49 T or F: The eikonal equation and Fermat's principle depend on the assumption that the index of refraction varies only gradually.

R50 T or F: The eikonal equation and Fermat's principle depend on the assumption that the angles involved must not be too big.

R51 T or F: The eikonal equation and Fermat's principle depend on the assumption that the polarization is important to the problem.

R52 T or F: Spherical aberration can be important even when the paraxial approximation works well.

R53 T or F: Chromatic aberration (the fact that refractive index depends on frequency) is an example of the violation of the paraxial approximation.

R54 T or F: The Fresnel approximation falls within the paraxial approximation.

R55 T or F: The imaging relation $1/f = 1/d_\text{o} + 1/d_\text{i}$ relies on the paraxial ray approximation.

R56 T or F: Spherical waves of the form e^{ikR}/R are exact solutions to Maxwell's equations.

R57 T or F: Spherical waves can be used to understand diffraction from apertures that are relatively large compared to λ.

R58 T or F: Fresnel was the first to conceive of spherical waves.

R59 T or F: Spherical waves were accepted by Poisson immediately without experimental proof.

R60 T or F: The array theorem is useful for deriving the *Fresnel* diffraction from a grating.

R61 T or F: A diffraction grating with a period h smaller than a wavelength is ideal for making a spectrometer.

R62 T or F: The blaze on a reflection grating can improve the amount of energy in a desired order of diffraction.

R63 T or F: The resolving power of a spectrometer used in a particular diffraction order depends *only* on the number of lines illuminated (not wavelength or grating period).

R64 T or F: The central peak of the Fraunhofer diffraction from two narrow slits separated by spacing h has the same width as the central diffraction peak from a single slit with width $\Delta x = h$.

R65 T or F: The central peak of the Fraunhofer diffraction from a circular aperture of diameter ℓ has the same width as the central diffraction peak from a single slit with width $\Delta x = \ell$.

R66 T or F: The Fraunhofer diffraction pattern appearing at the focus of a lens varies in *angular* width, depending on the focal length of the lens used.

R67 T or F: Fraunhofer diffraction can be viewed as a spatial Fourier transform (or inverse transform if you prefer) on the field at the aperture.

Problems

R68 (a) Derive Snell's law using Fermat's principle.

(b) Derive the law of reflection using Fermat's principle.

R69 (a) Consider a ray of light emitted from an object, which travels a distance d_o before traversing a lens of focal length f and then traveling a distance d_i.

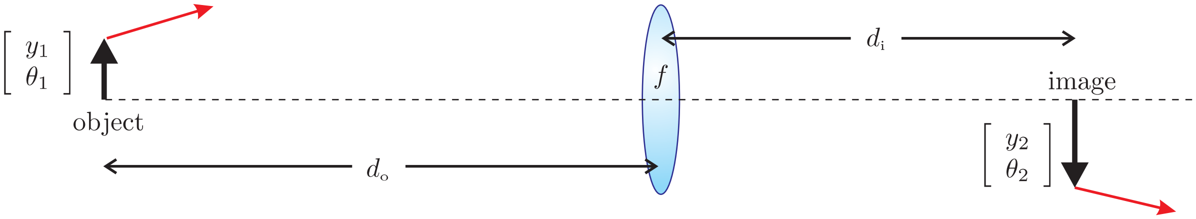

$\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$
object
d_o
f
d_i
image
$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$

Figure 11.24

Write a vector equation relating $\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$ to $\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$. Be sure to simplify the equation so that only one ABCD matrix is involved.

HINT: $\begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}, \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$

(b) Explain the requirement on the ABCD matrix in part (a) that ensures that an image appears for the distances chosen. From this requirement, extract a familiar constraint on d_o and d_i. Also, make a reasonable definition for magnification M in terms of y_1 and y_2, then substitute to find M in terms of d_o and d_i.

(c) A telescope is formed with two thin lenses separated by the sum of their focal lengths f_1 and f_2. Rays from a given far-away point all strike the first lens with essentially the same angle θ_1. Angular magnification M_θ quantifies the telescope's purpose of enlarging the apparent angle between points in the field of view.

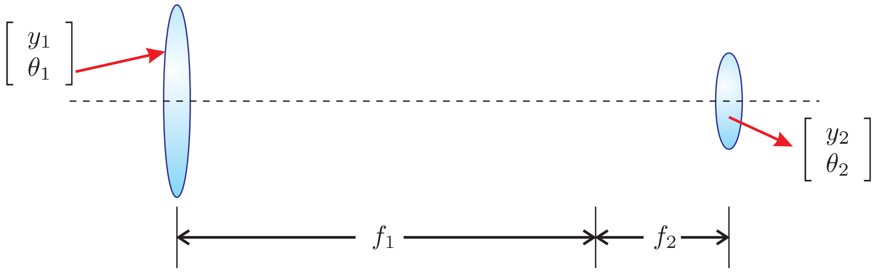

$\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$
$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$
f_1
f_2

Figure 11.25

Give a sensible definition for angular magnification in terms of θ_1 and θ_2. *Use ABCD-matrix formulation* to derive the angular magnification of the telescope in terms of f_1 and f_2.

R70 (a) Show that a system represented by a matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ (beginning and ending in the same index of refraction) can be made to look like the matrix for a thin lens if the beginning and ending positions along the z-axis are referenced from two principal planes, located distances p_1 and p_2 before and after the system.

HINT: $\begin{vmatrix} A & B \\ C & D \end{vmatrix} = 1$.

(b) Where are the principal planes located and what is the effective focal length for two identical thin lenses with focal lengths f that are separated by a distance $d = f$?

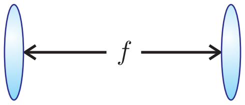


Figure 11.26

R71 Derive the on-axis intensity (i.e. $x, y = 0$) of a Gaussian laser beam if you know that at $z = 0$ the electric field of the beam is $E(\rho', z = 0) = E_0 e^{-\frac{\rho'^2}{w_0^2}}$.

Fresnel:

$$E(x, y, d) \cong -\frac{ie^{ikd}e^{i\frac{k}{2d}(x^2+y^2)}}{\lambda d}\iint E(x', y', 0)\, e^{i\frac{k}{2d}(x'^2+y'^2)}e^{-i\frac{k}{d}(xx'+yy')}dx'dy'$$

$$\int_{-\infty}^{\infty} e^{-Ax^2+Bx+C}dx = \sqrt{\frac{\pi}{A}}e^{\frac{B^2}{4A}+C}.$$

R72 (a) You decide to construct a simple laser cavity with a flat mirror and another mirror with concave curvature of $R = 100$ cm. What is the longest possible stable cavity that you can make?

HINT: Sylvester's theorem is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^N = \frac{1}{\sin\theta}\begin{bmatrix} A\sin N\theta - \sin\left(N-1\right)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin\left(N-1\right)\theta \end{bmatrix}$$

where $\cos\theta = \frac{1}{2}\left(A+D\right)$.

(b) The amplifier is YLF crystal, which lases at $\lambda = 1054$ nm. You decide to make the cavity 10 cm shorter than the longest possible (i.e. found in part (a)). What is the value of w_0, and where is the beam waist located inside the cavity (the place we assign to $z = 0$)?

HINT: One can interpret the parameter $R\left(z\right)$ as the radius of curvature of the wave front. For a mode to exist in a laser cavity, the radius of curvature of *each* of the end mirrors must match the radius of curvature of the beam at that location.

$$\begin{aligned}
&E\left(\rho,z\right) = E_0\frac{w_0}{w\left(z\right)}e^{-\frac{\rho^2}{w^2(z)}}e^{ikz+i\frac{k\rho^2}{2R(z)}}e^{-i\tan^{-1}\frac{z}{z_0}}\\
&\rho^2 \equiv x^2+y^2\\
&w\left(z\right) \equiv w_0\sqrt{1+z^2/z_0^2}\\
&R\left(z\right) \equiv z+z_0^2/z\\
&z_0 \equiv \frac{kw_0^2}{2}
\end{aligned}$$

R73 (a) Compute the Fraunhofer diffraction intensity pattern for a uniformly illuminated circular aperture with diameter ℓ.

HINT:

$$E\left(x,y,d\right) \cong -\frac{ie^{ikd}e^{i\frac{k}{2d}\left(x^2+y^2\right)}}{\lambda d}\iint E\left(x',y',0\right)e^{-i\frac{k}{d}\left(xx'+yy'\right)}dx'dy'$$

$$J_0\left(\alpha\right) = \frac{1}{2\pi}\int_0^{2\pi} e^{\pm i\alpha\cos\left(\theta-\theta'\right)}d\theta'$$

$$\int_0^a J_0\left(bx\right)xdx = \frac{a}{b}J_1\left(ab\right)$$

$$J_1\left(1.22\pi\right) = 0$$

$$\lim_{x\to 0}\frac{2J_1\left(x\right)}{x} = 1$$

(b) The first lens of a telescope has a diameter of 30 cm, which is the only place where light is clipped. You wish to use the telescope to examine two stars in a binary system. The stars are approximately 25 light-years away. How far apart need the stars be (in the perpendicular sense) for you to distinguish them in the visible range of $\lambda = 500$ nm? Compare with the radius of Earth's orbit, 1.5×10^8 km.

R74 (a) Derive the Fraunhofer diffraction pattern for the *field* from a uniformly illuminated single slit of width Δx. (Don't worry about the y-dimension.)

(b) Find the Fraunhofer *intensity* pattern for a grating of N slits of width Δx positioned on the mask at $x'_n = h\left(n - \frac{N+1}{2}\right)$ so that the spacing between all slits is h.

HINT: The array theorem says that the diffraction pattern is $\sum\limits_{n=1}^{N} e^{-i\frac{k}{d}xx'_n}$ times the diffraction pattern of a single slit. You will need

$$\sum_{n=1}^{N} r^n = r\frac{r^N - 1}{r - 1}$$

(c) Consider Fraunhofer diffraction from the grating in part (b). The grating is 5.0 cm wide and is uniformly illuminated. For best resolution in a monochromator with a 50 cm focal length, what should the width of the exit slit be? Assume a wavelength of $\lambda = 500$ nm.

Selected Answers

R72: (a) 100 cm (b) 0.32 mm.

R73: (b) 4.8×10^8 km.

R74: (c) 5 μm.

Chapter 12

Interferograms and Holography

12.1 Introduction

In chapter 7, we studied a Michelson interferometer in an idealized sense: 1) The light entering the instrument was considered to be a planewave. 2) The retro-reflecting mirrors were considered to be aligned perpendicular to the beams impinging on them. 3) All reflective surfaces were taken to be perfectly flat. If any of these conditions are relaxed, the result is an interference or *fringe pattern* in the beam emerging from the interferometer. A recorded fringe pattern (on a CCD or photographic film) is called an *interferogram*. In section 12.2, we shall examine typical fringe patterns that can be produced in an interferometer. Such patterns are very useful for testing the prescription and quality of optical components. Some examples of how to do this are addressed in section 12.3.

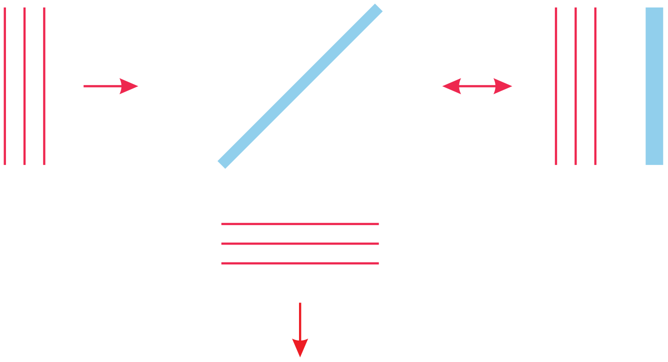


Figure 12.1 Michelson interferometer.

The technique of holography was conceived of by Dennis Gabor in the late 1940's. In optical holography, light interference patterns (or fringe patterns) are recorded and then later used to diffract light, much like gratings diffract light.[1] The recorded fringe pattern, when used for the purpose of diffracting light, is called a hologram. When the light diffracts from the hologram, it can mimic the light field originally used to generate the previously recorded fringe pattern. This is true even for very complex fields generated when light is scattered from arbitrary three-dimensional objects. When the light field is re-created through diffraction by the fringe pattern, an observer perceives the presence of the original object. The image looks three-dimensional since the holographic fringes re-construct the original light pattern simultaneously for a wide range of viewing angles. Holograms are studied in sections 12.4 and 12.5.

12.2 Interferograms

Consider the Michelson interferometer seen in Fig. 12.1. Suppose that the beamspliter divides the fields evenly, so that the overall output intensity is given by (8.1):

$$I_{\text{det}} = 2I_0\left[1 + \cos\left(\omega\tau\right)\right] \tag{12.1}$$

where τ is the roundtrip delay time of one path relative to the other. This equation is based on the idealized case, where the amplitude and phase of the two beams are uniform and perfectly aligned to each other following the beamsplitter. The entire beam "blinks" on and off as the delay path τ is varied.

What happens if one of the retro-reflecting mirrors is misaligned by a small angle θ? The fringe patterns seen in Fig. 12.2 (b)-(d) are the result. By the law of reflection, the beam returning from the misaligned mirror deviates from the "ideal" path by an angle 2θ. This puts a relative phase term of

$$\phi = kx\sin\left(2\theta_x\right) + ky\sin\left(2\theta_y\right) \tag{12.2}$$

on the misaligned beam (in addition to $\omega\tau$). Here θ_x represents the tilt of the mirror in the x-dimension and θ_y represents the amount of tilt in the y-dimension.

When the two plane waves join, the resulting intensity pattern is

$$I_{\text{det}} = 2I_0\left[1 + \cos\left(\phi + \omega\tau\right)\right] \tag{12.3}$$

Of course, the phase term ϕ depends on the local position within the beam through x and y. Regions of uniform phase, called fringes (in this case individual stripes), "blink" on and off together as the delay τ is varied. As the delay is varied, the fringes seem to "move" across the detector, owing to the fact that the phase of the "blinking" varies smoothly across the beam. The fringes emerge from one edge of the beam and disappear at the other.

Another interesting situation arises when the beams in a Michelson interferometer are diverging. A fringe pattern of concentric circles will be seen at the detector when the two beam paths are unequal (see Fig. 12.2 (e)). The radius of curvature for the beam traveling

[1] In fact, a grating can be considered to be a hologram and holographic techniques are often employed to produce gratings.

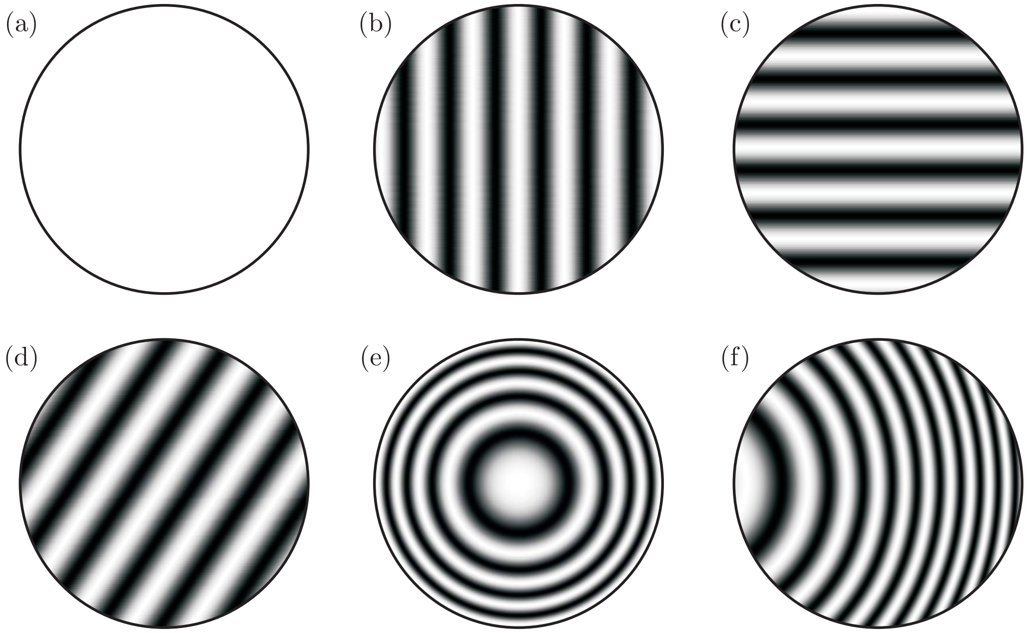


Figure 12.2 Fringe patterns for a Michelson interferometer: (a) perfectly aligned beams. (b) Horizontally misaligned beams. (c) Vertically misaligned beams. (d) Both vertically and horizontally misaligned beams. (e) Diverging beam with unequal paths. (f) Diverging beam with unequal paths and horizontal misalignment.

the longer path is increased by the added amount of delay $d = \tau/c$. Thus, if beam 1 has radius of curvature R_1 when returning to the beam splitter, then beam 2 will have radius $R_2 = R_1 + d$ upon return (assuming flat mirrors). The relative phase between the two beams is

$$\phi = k\rho^2/2R_1 - k\rho^2/2R_2 \tag{12.4}$$

and the intensity pattern at the detector is given as before by (12.3).

12.3 Testing Optical Components

A Michelson interferometer is ideal for testing the quality of optical surfaces. If any of the flat surfaces (including the beam splitter) in the interferometer are distorted, the fringe pattern readily reveals it. Fig. 12.3 shows an example of a fringe pattern when one of the mirrors in the interferometer has an arbitrary deformity in the surface figure. A new fringe stripe occurs for every half wavelength that the surface varies. (The round trip turns a half wavelength into a whole wavelength.) This makes it possible to determine the flatness of a surface with very high precision. Of course, in order to test a given surface in an interferometer, the quality of the other surfaces must first be ensured.

A typical industry standard for research-grade optics is to specify the surface flatness to within one tenth of an optical wavelength (633 nm HeNe laser). This means that the interferometer should reveal no more than one fifth of a fringe variation across the substrate. The fringe pattern tells the technician how the surface should continue to be polished in

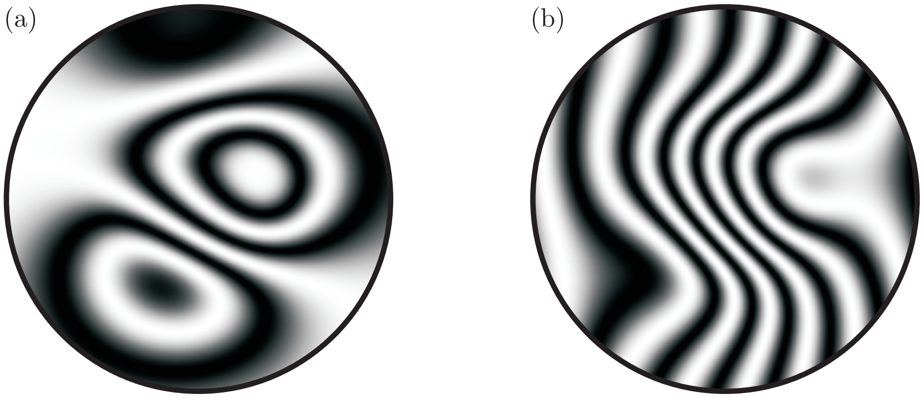


Figure 12.3 (a) Fringe pattern arising from an arbitrarily distorted mirror in a perfectly aligned interferometer with plane wave beams. (b) Fringe pattern from the same mirror as (a) when the mirror is tilted (still plane wave beams). The distortion due to surface variation is still easily seen.

order to achieve the desired surface flatness. When testing a surface, it is not necessary to remove all tilt from the alignment in order to see fringe effects due to surface variations. In fact, it is sometimes helpful to observe the effects of a distorted surface figure as deviations in a regular striped fringe pattern.

Other types of optical surfaces and optical component besides flat mirrors can also be tested with an interferometer. Fig. 12.4 shows how a lens can be tested using a convex mirror to compensate for the focusing action of the lens. With appropriate spacing, the lens-mirror combination can act like a flat surface. Distortions in the lens figure are revealed in the fringe pattern. In this case, the surfaces of the lens are tested together, and variations in optical path length are observed. In order to record fringes, say with a CCD camera, it is often convenient to image a larger beam onto a relatively small active area of the detector. The imaging objective should be adjusted to produce an image of the test optic on the detector screen. The diameter of the objective lens needs to accommodate the whole beam.

12.4 Generating Holograms

Consider a coherent monochromatic beam of light that is split in half by a beamsplitter, similar to that in a Michelson interferometer. Let one beam, called the reference beam, proceed directly to a recording film, and let the other beam scatter from an arbitrary object back towards the same film. The two beams interfere at the recording film. It may be advantageous to split the beam initially into unequal intensities such that the light scattered from the object has an intensity similar to the reference beam at the film.

The purpose of the film is to record the interference pattern. It is important that the coherence length of the light be much longer than the difference in path length starting from the beam splitter and ending at the film. In addition, during exposure to the film, it is important that the whole setup be stable against vibrations on the scale of a wavelength

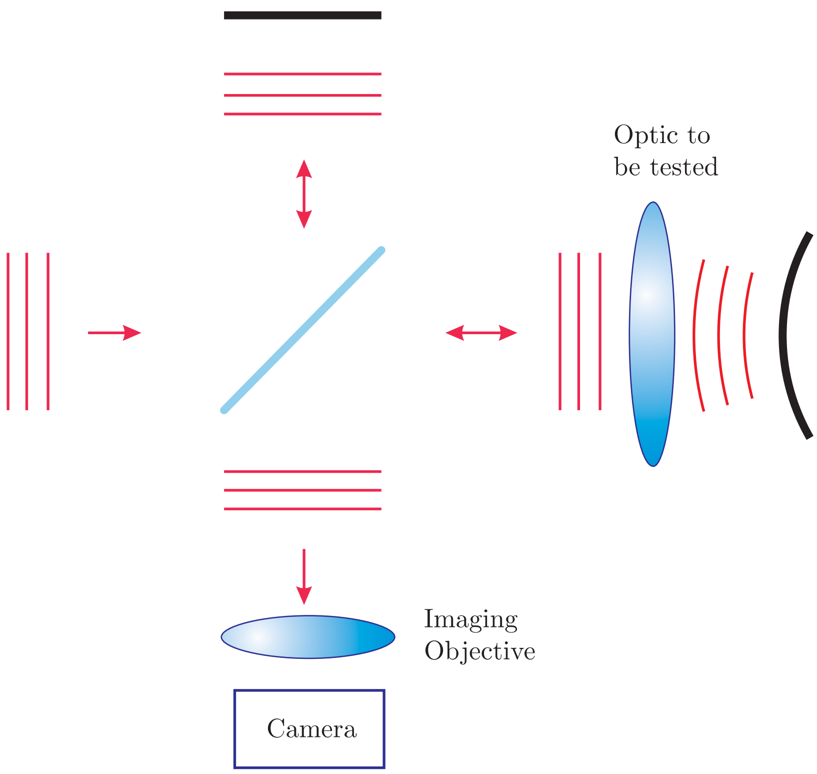


Figure 12.4 Twyman-Green setup for testing lenses.

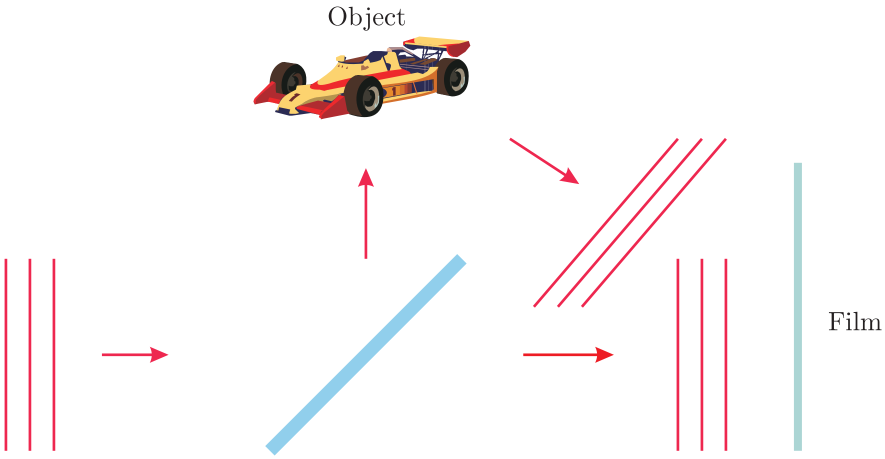
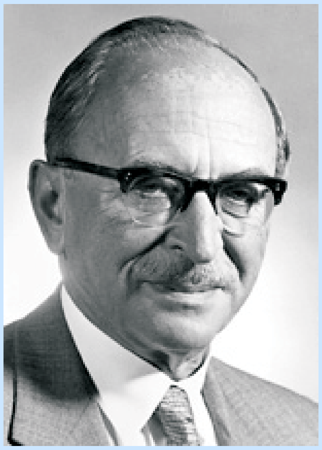


Figure 12.5 Exposure of holographic film.



Dennis Gabor

(1900–1979, Hungarian)

Gabor was educated and worked in Germany. However, when Hitler came to power, he left and eventually went to England. While there Gabor invented holography in the early 1950s, but it would not become practical until the invention of the laser.

since this will cause the fringes to washout. For simplicity, we neglect the vector nature of the electric field, assuming that the scattering from the object for the most part preserves polarization and that the angle between the two beams incident on the film is modest (so that the electric fields of the two beams are close to parallel). To the extent that the light scattered from the object contains the polarization component orthogonal to that of the reference beam, it provides a uniform (unwanted) background exposure to the film on top of which the fringe pattern is recorded.

In general terms, we may write the electric field arriving at the film as

$$E_{\text{film}}(\mathbf{r})\, e^{-i\omega t} = E_{\text{object}}(\mathbf{r})\, e^{-i\omega t} + E_{\text{ref}}(\mathbf{r})\, e^{-i\omega t} \tag{12.5}$$

Here, the coordinate $\mathbf{r}$ indicates locations on the film surface, which may have arbitrary shape. The field $E_{\text{object}}(\mathbf{r})$, which is scattered from the object, is in general very complicated. The field $E_{\text{ref}}(\mathbf{r})$ may be equally complicated, but typically it is convenient if it has a simple form such as a plane wave, since this beam must be re-created later in order to view the hologram.

The intensity of the field (12.5) is given by

$$\begin{aligned} I_{\text{film}}(\mathbf{r}) &= \frac{1}{2}c\epsilon_0 \left|E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}(\mathbf{r})\right|^2 \\ &= \frac{1}{2}c\epsilon_0 \left[\left|E_{\text{object}}(\mathbf{r})\right|^2 + \left|E_{\text{ref}}(\mathbf{r})\right|^2 + E^*_{\text{ref}}(\mathbf{r})\, E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}(\mathbf{r})\, E^*_{\text{object}}(\mathbf{r})\right] \end{aligned} \tag{12.6}$$

For typical photographic film, the exposure of the film is proportional to the intensity of the light hitting it. This is known as the linear response regime. That is, after the film is developed, the transmittance T of the light through the film is proportional to the intensity of the light that exposed it I_{film}. However, for low exposure levels, or for film specifically designed for holography, the transmission of the light through the film can be proportional to the square of the intensity of the light that exposes the film. Thus, after the film is exposed to the fringe pattern and developed, the film acquires a spatially varying transmission function according to

$$T(\mathbf{r}) \propto I_{\text{film}}^2(\mathbf{r}) \tag{12.7}$$

This means that a field that is later incident on the film has its amplitude modified by

$$E_{\text{transmitted}}(\mathbf{r}) = t(\mathbf{r})\, E_{\text{incident}}(\mathbf{r}) \propto I_{\text{film}}(\mathbf{r})\, E_{\text{incident}}(\mathbf{r}) \tag{12.8}$$

as it emerges from the other side of the film.

12.5 Holographic Wavefront Reconstruction

To see a holographic image, we re-illuminate film (previously exposed and developed) with the original reference beam. That is, we send in

$$E_{\text{incident}}(\mathbf{r}) = E_{\text{ref}}(\mathbf{r}) \tag{12.9}$$

and view the light that is transmitted. According to (12.6) and (12.8), the transmitted field is proportional to

$$\begin{aligned} E_{\text{transmitted}}(\mathbf{r}) &\propto I_{\text{film}}(\mathbf{r})\, E_{\text{ref}}(\mathbf{r}) \\ &= \left[|E_{\text{object}}(\mathbf{r})|^2 + |E_{\text{ref}}(\mathbf{r})|^2\right] E_{\text{ref}}(\mathbf{r}) + |E_{\text{ref}}(\mathbf{r})|^2\, E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}^2(\mathbf{r})\, E_{\text{object}}^*(\mathbf{r}) \end{aligned} \tag{12.10}$$

Although this expression looks fairly complicated, each of the three terms has a direct interpretation. The first term is just the reference beam $E_{\text{ref}}(\mathbf{r})$ with an amplitude modified by the transmission through the film. It is the residual undeflected beam, similar to the zero-order diffraction peak for a transmission grating. The second term is interpreted as a reconstruction of the light field originally scattered from the object $E_{\text{object}}(\mathbf{r})$. Its amplitude is modified by the intensity of the reference beam, but if the reference beam is uniform across the film, this hardly matters. An observer looking into the film sees a wavefront identical to the one produced by the original object. Thus, the observer sees a virtual image at the location of the original object. Since the wavefront of the original object has genuinely been recreated, the image looks “three-dimensional,” because the observer is free to view from different perspectives.

The final term in (12.10) is proportional to the complex conjugate of the original field from the object. It also contains twice the phase of the reference beam, which we can overlook if the reference beam is uniform on the film. In this case, the complex conjugate of the object field actually converges to a real image of the original object. This image is located on the observer’s side of the film, but it is often of less interest since the image is inside out. An ideal screen for viewing the real image would be an item shaped identical to the original object, which of course defeats the purpose of the hologram! To the extent that the film is not flat or to the extent that the reference beam is not a plane wave, the phase of $E_{\text{ref}}^2(\mathbf{r})$ severely distorts the image. The virtual image never suffers from this problem.

As an example, consider a hologram made from a point object, as depicted in Fig. 12.7. Presumably, the point object is illuminated sufficiently brightly so as to make the scattered light have an intensity similar to the reference beam at the film.

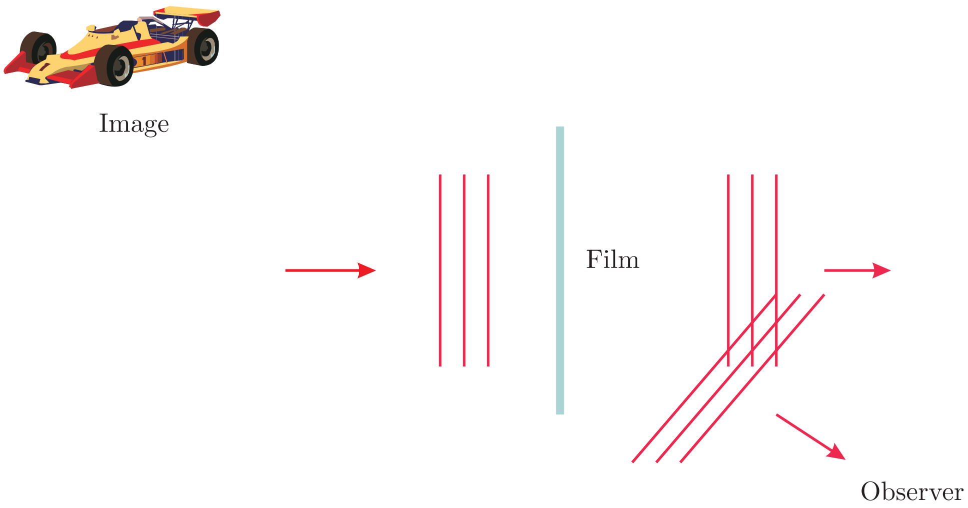


Figure 12.6 Holographic reconstruction of wavefront through diffraction from fringes on film. Compare with Fig. 12.2.

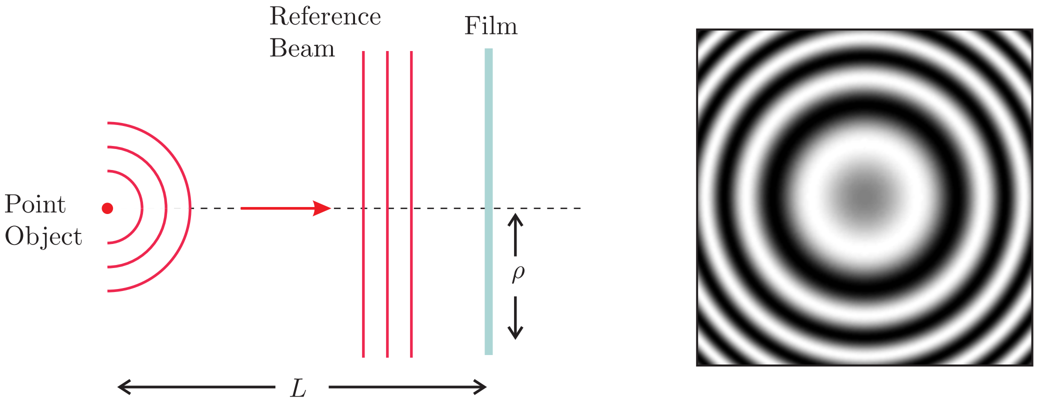


Figure 12.7 Exposure to holographic film by a point source and a reference plane wave. The holographic fringe pattern for a point object and a plane wave reference beam exposing a flat film is shown on the right.

Let the reference plane wave strike the film at normal incidence. Then the reference field will have constant amplitude and phase across it; call it E_{ref}. The field from the point object can be treated as a spherical wave:

$$E_{\text{object}}(\rho) = \frac{E_{\text{ref}} L}{\sqrt{L^2 + \rho^2}} e^{ik\sqrt{L^2+\rho^2}} \qquad \text{(point source example)} \tag{12.11}$$

Here ρ represents the radial distance from the center of the film to some other point on the film. We have taken the amplitude of the object field to match E_{ref} in the center of the film.

After the film is exposed, developed, and re-illuminated by the reference beam, the field emerging from the right-hand-side of the film, according to (12.10), becomes

$$\begin{aligned} E_{\text{transmitted}}(\rho) \propto \left[\frac{E_{\text{ref}}^2 L^2}{L^2+\rho^2} + E_{\text{ref}}^2\right] E_{\text{ref}} + E_{\text{ref}}^2 \frac{E_{\text{ref}} L}{\sqrt{L^2+\rho^2}} e^{ik\sqrt{L^2+\rho^2}} \\ + E_{\text{ref}}^2 \frac{E_{\text{ref}} L}{\sqrt{L^2+\rho^2}} e^{-ik\sqrt{L^2+\rho^2}} \qquad \text{(point source example)} \end{aligned} \tag{12.12}$$

We see the three distinct waves that emerge from the holographic film. The first term in (12.12) is merely the plane wave reference beam passing straight through the film (with some variation in amplitude), which is depicted in Fig. 12.8 (a). The second term in (12.12) has the identical form as the field from the original object (aside from an overall amplitude factor). It describes an outward-expanding spherical wave, which gives rise to a virtual image at the location of the original point object, as depicted in Fig. 12.8 (b). The final term in (12.12) corresponds to a converging spherical wave, which focuses to a point at a distance L from the observer's side of the screen (depicted in Fig. 12.8 (c)).

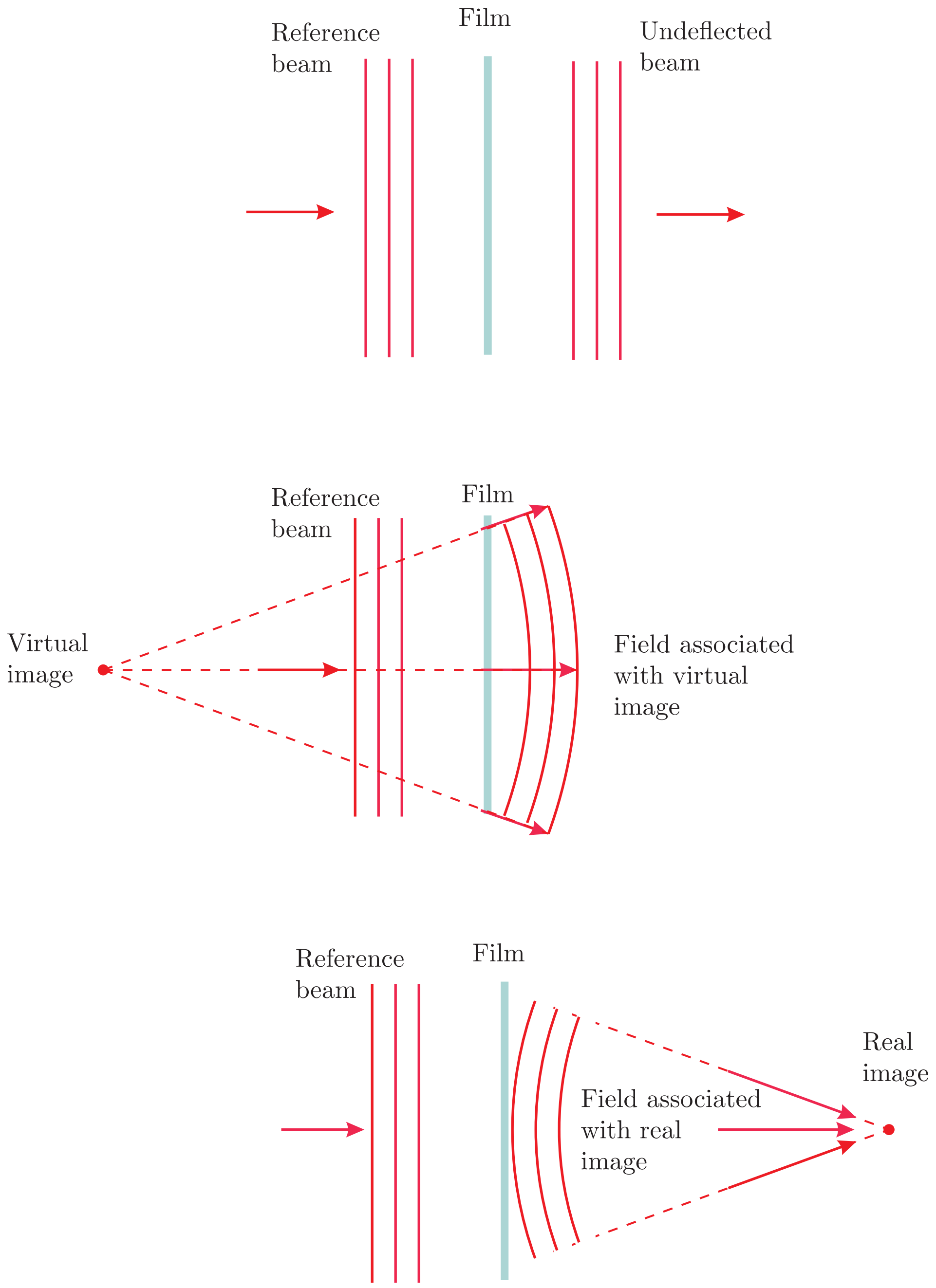


Figure 12.8 Reference beam incident on previously exposed holographic film. (a) Part of the beam goes through. (b) Part of the beam takes on the field profile of the original object. undeflected. (c) Part of the beam converges to a real image of the original object.

Exercises

12.4 Generating Holograms

P12.1 An ideal Michelson interferometer that uses flat mirrors is perfectly aligned to a wide collimated laser beam. Suppose that one of the mirrors is then misaligned by 0.1°. What is the spacing between adjacent fringes on the screen if the wavelength is $\lambda = 633$ nm? What would happen if the the angle of the input beam (before the beamsplitter) was tilted by 0.1°?

P12.2 An ideal Michelson interferometer uses flat mirrors perfectly aligned to an expanding beam that diverges from a point 50 cm before the beamsplitter. Suppose that one mirror is 10 cm away from the beam splitter, and the other is 11 cm. Suppose also that the center of the resulting bull's-eye fringe pattern is dark. If a screen is positioned 10 cm after the beam splitter, what is the radial distance to the next dark fringe on the screen if the wavelength is $\lambda = 633$ nm?

L12.3 Set up an interferometer and observe distortions to a mirror substrate when the setscrew is over tightened.

P12.4 Consider a diffraction grating as a simple hologram. Let the light from the "object" be a plane wave (object placed at infinity) directed onto a flat film at angle θ. Let the reference beam strike the film at normal incidence, and take the wavelength to be λ.

(a) What is the period of the fringes?

(b) Show that when re-illuminated by the reference beam, the three terms in (12.10) give rise zero-order and 1st-order diffraction to either side of center.

(c) Check that it matches predictions in the previous section.

P12.5 Consider the holographic pattern produced by the point object described in section 12.5.

(a) Show that the phase of the real image in (12.12) may be approximated as $\Delta\phi = -k\rho^2/2L$, aside from a spatially independent overall phase. Compare with (11.25) and comment.

(b) This hologram is similar to a Fresnel zone plate, used to focus extreme ultraviolet light or x-rays, for which it is difficult to make a lens. Graph the field transmission for the hologram as a function of ρ and superimpose a similar graph for a "best-fit" mask that has regions of either 100% or 0% transmission. Use $\lambda = 633$ nm and $L = (5 \times 10^5 - \frac{1}{4})\lambda$ (this places the point source about a 32 cm before the screen). See Fig. 12.9.

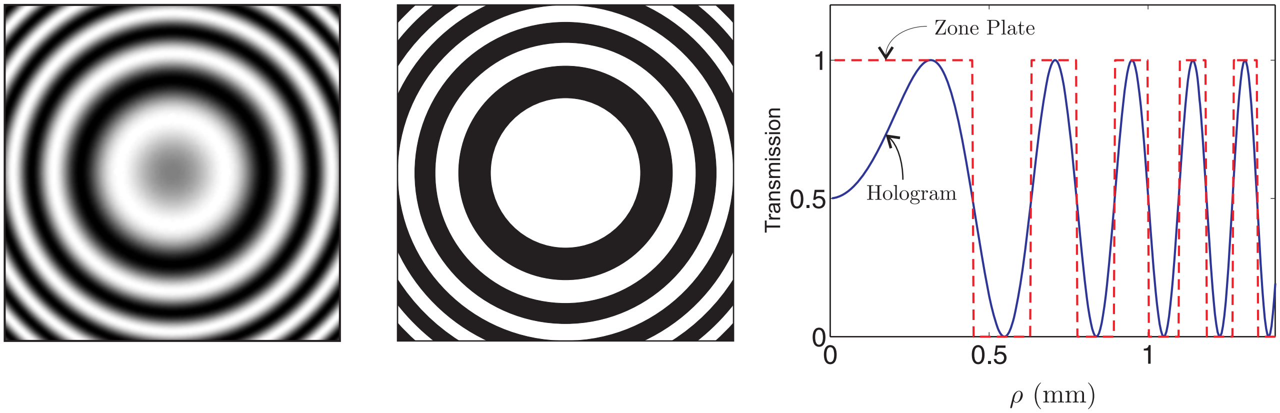


Figure 12.9 Field transmission for a point-source hologram (left) and a Fresnel zone plate (middle), and a plot of both as a function of radius (right).

L12.6 Make a hologram.

Chapter 13

Blackbody Radiation

13.1 Introduction

Hot objects glow. In 1860, Kirchhoff proposed that the radiation emitted by hot objects as a function of frequency is approximately the same for all materials. (An important exception is atomic vapors, which have relatively few discrete spectral lines. However, Kirchhoff's assumption holds quite well for most solids, which are sufficiently complex.) The notion that all materials behave similarly led to the concept of an ideal "blackbody" radiator. Most materials have a certain shininess that causes light to reflect or scatter in addition to being absorbed and reemitted. However, light that falls upon an *ideal blackbody* is absorbed perfectly before the possibility of reemission, hence the name blackbody. The distribution of frequencies emitted by a blackbody radiator is related to its temperature. The key concept of a blackbody radiator is that the light surrounding it is in thermal equilibrium with the radiation. If some of the light escapes to the environment, the object inevitably must cool as it continually moves towards a new thermal equilibrium.

The Sun is a good example of a blackbody radiator. The light emitted from the Sun is associated with its surface temperature. Any light that arrives to the Sun from outer space is virtually 100% absorbed, however little light that might be. Mostly, light escapes to the much colder surrounding space, and the temperature of the Sun's surface is maintained by the fusion process within.

Experimentally, a near perfect blackbody radiator can be constructed from a hollow object. As the object is heated, the light present inside the internal cavity can only come from the walls. Also, any radiation in the interior cavity is eventually absorbed (before being potentially reemitted), if not on the first bounce then on subsequent bounces. In this case, the walls of the cavity and light field are in thermal equilibrium. A small hole can be drilled through the wall into the interior to observe the radiation there without significantly disturbing the system. A glowing tungsten filament also makes a reasonably good example of a blackbody radiator. However, if not formed into a cavity, one must take surface reflections into account because the *emissivity* is less than unity.

In this chapter, we develop a theoretical understanding of blackbody radiation and provide some historical perspective. One of the earliest properties deduced about blackbody radiation is known as the Stefan-Boltzmann law, derived from thermodynamic ideas in 1879, long before blackbody radiation was fully understood. This law says that the total intensity

Gustav Kirchhoff

(1824–1887, German)

Kirchhoff studied the spectra emitted by various objects. He coined the term "blackbody" radiation. He understood that an excited gas gives off a discrete spectrum, and that an unexcited gas surrounding a blackbody emitter produces dark lines in the blackbody spectrum.

I of radiation (including all frequencies) that flows outward from a blackbody radiator is given by

$$I = e\sigma T^4, \tag{13.1}$$

where σ is called the Stefan-Boltzmann constant and T is the absolute temperature (in Kelvin) of the blackbody. The value of the Stefan-Boltzmann constant is $\sigma = 5.6696 \times 10^{-8}\ \mathrm{W/m^2 \cdot K^4}$. The dimensionless parameter e called the emissivity is equal to one for an ideal blackbody surface. However, it is less than one for actual materials because of surface reflections. For example, the emissivity of tungsten is approximately $e = 0.4$.

It is sometimes useful to express intensity in terms of the energy density of the light field u_{field} (given by (2.52) in units of energy per volume). This connection between *outward-going* intensity and energy density of the field is given by

$$I = \frac{cu_{\text{field}}}{4} \Rightarrow u_{\text{field}} = e\frac{4\sigma T^4}{c} \tag{13.2}$$

since the energy travels at speed c equally in all directions (for example, inside a cavity within a solid object). A factor of 1/2 occurs because only half of the energy travels away from rather than towards any given surface (e.g. the wall of the cavity). The remaining factor of 1/2 occurs because the energy that flows outward through a given surface is directionally distributed over a hemisphere as opposed to flowing only in the direction of the surface normal $\hat{\mathbf{n}}$. The average over the hemisphere is carried out as follows:

$$\frac{\int_0^{2\pi} d\phi \int_0^{\pi/2} \mathbf{r}\cdot\hat{\mathbf{n}}\sin\theta d\theta}{\int_0^{2\pi} d\phi \int_0^{\pi/2} r\sin\theta d\theta} = \frac{\int_0^{2\pi} d\phi \int_0^{\pi/2} r\cos\theta\sin\theta d\theta}{\int_0^{2\pi} d\phi \int_0^{\pi/2} r\sin\theta d\theta} = \frac{1}{2} \tag{13.3}$$

The thermodynamic derivation of the Stefan-Boltzmann law is given in appendix 13.A.

Although (13.1) describes the total intensity of the light that leaves a blackbody surface, it does not describe what frequencies make up the radiation field. This frequency distribution was not fully described for another two decades when Max Planck developed his famous formula. Planck first arrived at the blackbody radiation formula empirically in an effort to match experimental data. He then attempted to explain it, which marks the birth of quantum mechanics. Even Planck was uncomfortable with and perhaps disbelieved the assumptions that his formula implied, but he deserves credit for recognizing and articulating those assumptions. In section 13.3, we study how Planck's blackbody radiation formula implies the existence of electromagnetic quanta, which we now call photons. In section 13.2 we first examine the failure of classical ideas to explain blackbody radiation (even though this failure was only appreciated years after Planck developed his formula). Section 13.4 gives an analysis of blackbody radiation developed by Einstein where he introduced the concept of stimulated and spontaneous emission. In this sense, Einstein can be thought of as the father of light amplification by stimulated emission of radiation (LASER).

13.2 Failure of the Equipartition Principle

In the latter part of the 1800's as spectrographic technology improved, experimenters acquired considerable data on the spectra of blackbody radiation. Experimentalists were able to make detailed maps of the intensity per frequency associated with blackbody radiation over a fairly wide wavelength range. The results appeared to be independent of the material as long as the object was black and rough, and this suggested general underlying physical reasons for the behavior. The intensity per frequency depended only on temperature and when integrated over all frequencies agreed with the Stefan-Boltzmann law (13.1).

In 1900, Rayleigh (and later Jeans in 1905) attempted to explain the blackbody spectral distribution (intensity per frequency) as a function of temperature by applying the equipartition theorem to the problem. Recall, the equipartition theorem states the energy in a system on the average is distributed equally among all degrees of freedom in the system. For example, a system composed of oscillators (say, electrons attached to "springs" representing the response of the material on the walls of a blackbody radiator) has an energy of $k_{\mathrm{B}}T/2$ for each degree of freedom, where $k_{\mathrm{B}} = 1.38 \times 10^{-23}\mathrm{J/K}$ is Boltzmann's constant. Rayleigh and Jeans supposed that each unique mode of the electromagnetic field should carry energy $k_{\mathrm{B}}T$ just as each mechanical spring in thermal equilibrium carries energy $k_{\mathrm{B}}T$ ($k_{\mathrm{B}}T/2$ as kinetic and $k_{\mathrm{B}}T/2$ as potential energy). The problem then reduces to that of finding the number of unique modes for the radiation at each frequency. They anticipated that requiring each mode of electromagnetic energy to hold energy $k_{\mathrm{B}}T$ should reveal the spectral shape of blackbody radiation.

A given frequency is associated with a specific wave number $k = \sqrt{k_x^2 + k_y^2 + k_z^2}$. Notice that there are many ways (i.e. combinations of k_x, k_y, and k_z) to come up with the same wave number $k = 2\pi\nu/c$ (corresponding to a single frequency ν). To count these ways properly, we can let our experience with Fourier series guide us. Consider a box with each side of length L. The Fourier theorem (0.31) states that the total field inside the box (no matter how complicated the distribution) can always be represented as a superposition of sine (and cosine) waves. The total field in the box can therefore be written as

$$\mathrm{Re}\left\{ \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} E_{n,m,\ell} e^{i(nk_0x+mk_0y+\ell k_0z)} \right\} \tag{13.4}$$

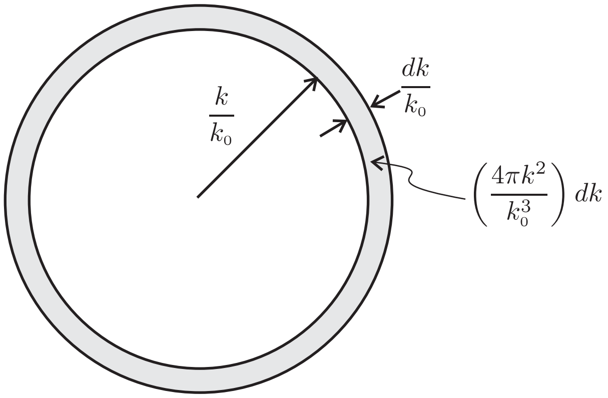


Figure 13.1 The volume of a thin spherical shell in n, m, ℓ space.

where each component of the wave number in any of the three dimensions is always an integer times

$$k_0 = 2\pi/L \tag{13.5}$$

We must keep in mind that (13.4) does not account for the two distinct polarizations for each wave. To find the total number of modes associated with a given frequency, we should double the number of *terms* in (13.4) that have that frequency. It is important to note that we have not artificially made any restrictions by considering the box of size L since we may later take the limit $L \to \infty$ so that our box represents the entire universe. In fact, L naturally disappears from our calculation as we consider the *density of modes.*

We can think of a given wave number k as specifying the equation of a sphere in a coordinate system with axes labeled n, m, and ℓ:

$$n^2 + m^2 + \ell^2 = \left(\frac{k}{k_0}\right)^2 \tag{13.6}$$

We need to know how many more ways there are to choose n, m, and ℓ when the wave number k/k_0 is replaced by $(k + dk)/k_0$. The answer is the difference in the volume of the two spheres as shown in Fig. 13.1:

$$\#\text{ modes in } (k,k+dk) = \left(4\pi\frac{k^2}{k_0^2}\right)\frac{dk}{k_0} \tag{13.7}$$

This represents the number of ways to come up with a wave number between k and $k + dk$. Again, this is the number of terms in (13.4) with a wave number between k and $k + dk$. Recall that n, m, and ℓ are integers. Notice that we have included the possibility of negative integers. This automatically takes into account the fact that for each mode (defined by a set n, m, and ℓ) the field may travel in the forwards or the backwards direction.

Since according to the Rayleigh-Jeans assumption each mode carries energy of $k_B T$, the energy *density* (energy per volume) associated with a specified range of wave numbers dk is $k_B T/L^3$ times (13.7), the number of modes within that range. Thus, the total energy density in the field for all wave numbers is

$$u_{\text{field}} = \int_0^\infty 2 \times \frac{k_{\text{B}}T}{L^3} \times \frac{4\pi k^2}{k_0^3} dk = k_{\text{B}}T \int_0^\infty \frac{k^2}{\pi^2} dk \tag{13.8}$$

where the extra factor of 2 accounts for two independent polarizations for each mode. The dependence on L has disappeared from (13.8).

We can see that (13.8) disagrees drastically with the Stefan-Boltzmann law (13.2), since (13.8) is proportional to temperature rather than to its fourth power. In addition, the integral in (13.8) is seen to diverge, meaning that regardless of the temperature, the light carries infinite energy density! This has since been named the ultraviolet catastrophe since the divergence occurs on the short wavelength end of the spectrum. This is a clear failure of classical physics to explain blackbody radiation. Nevertheless, Rayleigh emphasized the fact that his formula worked well for the longer wavelengths and he did not necessarily want to abandon classical physics. Such dramatic changes take time.

It is instructive to make the change of variables $k = 2\pi\nu/c$ in the integral to write

$$u_{\text{field}} = k_{\text{B}}T \int_0^\infty \frac{8\pi\nu^2}{c^3} d\nu \tag{13.9}$$

The important factor $8\pi\nu^2/c^3$ can now be understood to be the number of modes per frequency. Then (13.9) is rewritten as

$$u_{\text{field}} = \int_0^\infty \rho\left(\nu\right) d\nu \tag{13.10}$$

where

$$\rho_{\text{Rayleigh-Jeans}}\left(\nu\right) = k_{\text{B}}T \frac{8\pi\nu^2}{c^3} \tag{13.11}$$

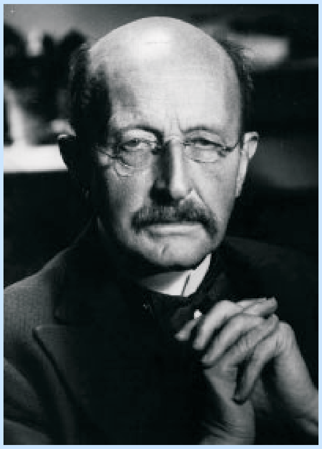
describes (incorrectly) the *spectral energy density* of the radiation field associated with blackbody radiation.

13.3 Planck's Formula

In the late 1800's Wien considered various physical and mathematical constraints on the spectrum of blackbody radiation and tried to find a function to fit the experimental data. The form for the energy distribution of blackbody radiation that Wien proposed was

$$\rho_{\text{Wien}}\left(\nu\right) = \frac{8\pi h\nu^3 e^{-h\nu/k_{\text{B}}T}}{c^3} \tag{13.12}$$

It is important to note that the constant h had not yet been introduced by Planck. The actual way that Wien wrote his distribution was $\rho_{\text{Wien}}\left(\nu\right) = a\nu^3 e^{-b\nu/T}$, where a and b were parameters used to fit the data.



Max Planck

(1858–1947, German)

Planck's work on thermodynamics led him to study the equilibrium between hot objects and electromagnetic radiation, which led to his introduction of the energy quantum in 1900. While he won the Nobel prize in 1918 for this contribution, he had serious reservations about the course that quantum mechanics theory took. He rejected the Copenhagen interpretation of quantum mechanics.

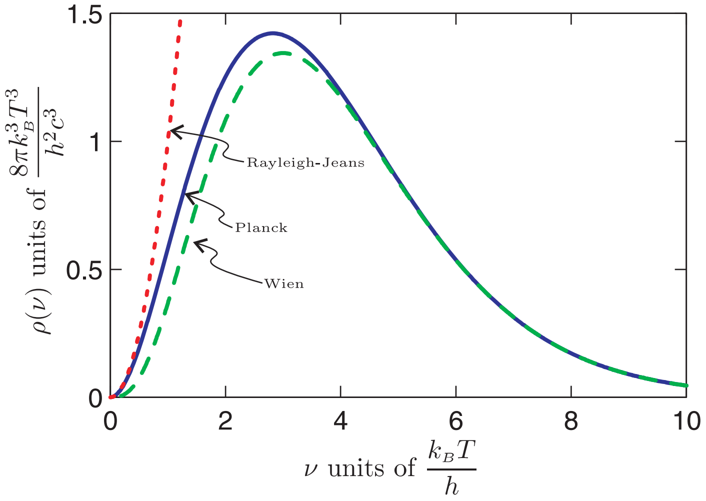


Figure 13.2 Energy density per frequency according to Planck, Wien, and Rayleigh-Jeans.

Wien's formula did a good job of fitting experimental data. However, in 1900 Lummer and Pringshein reported experimental data that deviated from the Wien distribution at long wavelengths (infrared). Max Planck was privy to this information and later that year came up with a revised version of Wien's formula that fit the data beautifully everywhere:

$$\rho_{\text{Planck}}(\nu) = \frac{8\pi h\nu^3}{c^3\left[e^{h\nu/k_\text{B}T} - 1\right]} \tag{13.13}$$

where $h = 6.626 \times 10^{-34}\text{J} \cdot \text{s}$ is an experimentally determined constant.

As seen in Fig. 13.2, the Rayleigh-Jeans curve, (13.11), and the Wien curve, (13.12), both fit the Planck's distribution function asymptotically on opposite ends. The Wien distribution does a good job nearly everywhere. However, at long wavelengths it was off by just enough for the experimentalists to notice that something was wrong.

At this point, it may seem fair to ask, what did Planck do that was so great? After all, he simply guessed a function that was only a slight modification of Wien's distribution. And he knew the "answer from the back of the book," namely Lummer's and Pringshein's well done experimental results. (At the time, Planck was unaware of the work by Rayleigh.) What Planck did that was so great was to interpret the meaning of his new formula. His interpretation was what he called an "act of desperation." While Planck was able to explain the implications of his formula, he did not assert that the implications were necessarily right; in fact, he presented them somewhat apologetically. It was several years later that the young Einstein published his paper explaining the photoelectric effect in terms of the implications of Planck's formula. Planck's insight was an enormous step towards understanding the quantum nature of light. The full theory of quantum electrodynamics would not be developed until nearly three decades later. Students should appreciate that the very people who developed quantum mechanics were also bothered by its confrontation with deep-seated intuition. If quantum mechanics bothers you, you should feel yourself in good company!

Planck found that he could derive his formula only if he made the following strange assumption: A given mode of the electromagnetic field is not able to carry an arbitrary amount of energy (for example, $k_{\mathrm{B}}T$ which varies continuously as the temperature varies). Rather, the field can only carry discrete amounts of energy separated by spacing $h\nu$. Under this assumption, the probability P_n that a mode of the field is excited to the n^{th} level is proportional to the Boltzmann statistical weighting factor $e^{-nh\nu/k_{\mathrm{B}}T}$. We can normalize this factor by dividing by the sum of all such factors to obtain the probability of having energy $nh\nu$ in a particular mode:

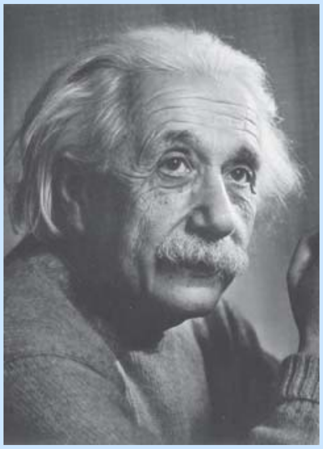
$$P_n = \frac{e^{-nh\nu/k_{\mathrm{B}}T}}{\sum\limits_{m=0}^{\infty} e^{-mh\nu/k_{\mathrm{B}}T}} = e^{-nh\nu/k_{\mathrm{B}}T}\left[1 - e^{-h\nu/k_{\mathrm{B}}T}\right] \tag{13.14}$$

Then, the energy in each mode of the field is expected to be

$$\begin{aligned}\sum_{n=0}^{\infty} h\nu n P_n &= h\nu\left[1 - e^{-h\nu/k_{\mathrm{B}}T}\right]\sum_{n=0}^{\infty} n e^{-nh\nu/k_{\mathrm{B}}T}\\ &= h\nu\left[e^{-h\nu/k_{\mathrm{B}}T} - 1\right]\frac{\partial}{\partial\left(h\nu/k_{\mathrm{B}}T\right)}\sum_{n=0}^{\infty} e^{-nh\nu/k_{\mathrm{B}}T}\\ &= \frac{h\nu}{e^{h\nu/k_{\mathrm{B}}T} - 1}\end{aligned} \tag{13.15}$$

Equation (13.15) is interpreted as the expectation of the energy (associated with an individual frequency) based on probabilities consistent with thermal equilibrium. Finally, we multiply this expected energy by the mode density $8\pi\nu^2/c^3$, obtained in the derivation of the Rayleigh-Jeans formula. In other words, we substitute (13.15) for $k_{\mathrm{B}}T$ in (13.10) to obtain the Planck distribution (13.13).

It is interesting that we are now able to *derive* the constant in the Stefan-Boltzmann law (13.2) in terms of Planck's constant h (see P 13.3). The Stefan-Boltzmann law is obtained by integrating the spectral density function (13.13) over all frequencies to obtain the total



Albert Einstein
(1879–1955, German)

Einstein is without a doubt the most famous scientist in history, and he made significant contributions to the field of optics. Einstein took Planck's notion of energy quanta and used them to explain the photoelectric effect. In addition, he developed a description that predicted the possibility of lasers years before quantum theory was fully developed.

field energy density, which is in thermal equilibrium with the blackbody radiator:

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Plank}}(\nu) d\nu = \frac{4}{c} \frac{2\pi^5 k_B^4}{15c^2h^3} T^4 = \frac{4}{c}\sigma T^4 \tag{13.16}$$

The Stefan-Boltzmann constant is thus calculated in terms of Planck's constant. However, Planck's constant was not introduced for several decades after the Stefan-Boltzmann law was developed. Thus, one may say that the Stefan-Boltzmann constant pins down Planck's constant.

13.4 Einstein's A and B Coefficients

More than a decade after Planck introduced his formula, and after Bohr had proposed that electrons occupy discrete energy states in atoms, Einstein reexamined blackbody radiation in terms of Bohr's new idea. If the material of a blackbody radiator interacts with a mode of the field with frequency ν, then electrons in the material must make transitions between two energy levels with energy separation $h\nu$. Since the radiation of a blackbody is in thermal equilibrium with the material, Einstein postulated that the field *stimulates* electron transitions between the states. In addition, he postulated that some transitions must occur spontaneously. (If the possibility of spontaneous transitions is not included, then there can be no way for a field mode to receive energy if none is present to begin with.)

Einstein wrote down rate equations for populations of the two levels N_1 and N_2 associated with the transition $h\nu$:

$$\begin{aligned} \dot{N}_1 &= A_{21}N_2 - B_{12}\rho(\nu) N_1 + B_{21}\rho(\nu) N_2, \\ \dot{N}_2 &= -A_{21}N_2 + B_{12}\rho(\nu) N_1 - B_{21}\rho(\nu) N_2 \end{aligned} \tag{13.17}$$

The coefficient A_{21} is the rate of spontaneous emission from state 2 to state 1, $B_{12}\rho(\nu)$

is the rate of stimulated absorption from state 1 to state 2, and $B_{21}\rho(\nu)$ is the rate of stimulated emission from state 2 to state 1.

In thermal equilibrium, the rate equations (13.17) are both equal to zero (i.e., $\dot{N}_1 = \dot{N}_2 = 0$) since the relative populations of each level must remain constant. We can then solve for the spectral density $\rho(\nu)$ at the given frequency. Either expression in (13.17) yields

$$\rho(\nu) = \frac{A_{21}}{\frac{N_1}{N_2}B_{12} - B_{21}} \tag{13.18}$$

In thermal equilibrium, the spectral density must match the Planck spectral density formula (13.13). In making the comparison, we should first rewrite the ratio N_1/N_2 of the populations in the two levels using the Boltzmann probability factor:

$$\frac{N_1}{N_2} = \frac{e^{-E_1/k_\mathrm{B}T}}{e^{-E_2/k_\mathrm{B}T}} = e^{(E_2-E_1)/k_\mathrm{B}T} = e^{h\nu/k_\mathrm{B}T} \tag{13.19}$$

Then when equating (13.18) to the Planck blackbody spectral density (13.13) we get

$$\frac{A_{21}}{e^{h\nu/k_\mathrm{B}T}B_{12} - B_{21}} = \frac{8\pi h\nu^3}{c^3\left[e^{h\nu/k_\mathrm{B}T} - 1\right]} \tag{13.20}$$

From this expression we deduce that

$$B_{12} = B_{21} \tag{13.21}$$

and

$$A_{21} = \frac{8\pi h\nu^3}{c^3}B_{21} \tag{13.22}$$

We see from (13.21) that the rate of stimulated absorption is the same as the rate of stimulated emission. In addition, if one knows the rate of stimulated emission between a pair of states, it follows from (13.22) that one also knows the rate of spontaneous emission. This is remarkable because to derive A_{21} directly, one needs to use the full theory of quantum electrodynamics (the complete photon description). However, to obtain B_{21}, it is actually only necessary to use the semiclassical theory, where the light is treated classically and the energy levels in the material are treated quantum-mechanically using the Schrödinger equation. The usual semiclassical theory cannot explain spontaneous emission, but it can explain stimulated emission and the rate of sponaneous emission can then be obtained indirectly through (13.22). It should be mentioned that (13.21) and (13.22) assume that the energy levels 1 and 2 are non-degenerate. Some modifications must be made in the case of degenerate levels, but the procedure is similar.

In writing the rate equations, (13.17), Einstein predicted the possibility of creating lasers fifty years in advance of their development. These rate equations are still valid even if the light is not in thermal equilibrium with the material. The equations suggest that if the population in the upper state 2 can be made artificially large, then amplification will result via the stimulated transition. The rate equations also show that a population inversion (more population in the upper state than in the lower one) cannot be achieved by "pumping" the material with the same frequency of light that one hopes to amplify. This is because the stimulated absorption rate is balanced by the stimulated emission rate. The material-dependent parameters A_{21} and $B_{12} = B_{21}$ are called the Einstein A and B coefficients.

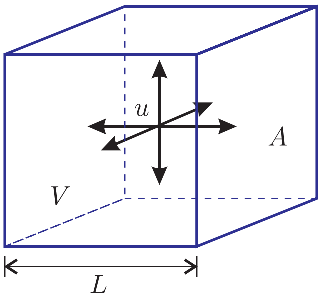


Figure 13.3 Field inside a blackbody radiator.

Appendix 13.A Thermodynamic Derivation of the Stefan-Boltzmann Law

In this appendix, we derive the Stefan-Boltzmann law. This derivation is included for historical interest and may be a little difficult to follow. The derivation relies on the 1st and 2nd laws of thermodynamics. Consider a container whose walls are all at the same temperature and in thermal equilibrium with the radiation field inside, according to the properties of an ideal blackbody radiator.

Notice that the units of energy density u_{field} (energy per volume) are equivalent to force per area, or in other words pressure. The radiation exerts a pressure of

$$P = u_{\text{field}}/3 \tag{13.23}$$

on each wall of the box. This can be derived from the fact that radiation of energy ΔE imparts a momentum

$$\Delta p = \frac{2\Delta E}{c}\cos\theta \tag{13.24}$$

when it is absorbed and reemitted from a wall at an angle θ. The fact that light carries momentum was understood well before the development of the theory of relativity and the photon description of light. The total pressure (force per area averaged over all angles) on a wall averages to be

$$P = \frac{\int\limits_0^{\pi/2} \frac{\Delta p}{\Delta t}\frac{1}{A}\sin\theta\; d\theta}{\int\limits_0^{\pi/2} \sin\theta\; d\theta} \tag{13.25}$$

where A is the area of the wall and $\Delta E = u_{\text{field}}AL$ is the total energy in the box, which makes a round trip during the interval $\Delta t = 2L/(c\cos\theta)$. L is the length of the box in the direction perpendicular to the surface. Upon performing the integration in (13.25), the simple result (13.23) is obtained.

To derive the Stefan-Boltzmann law, consider entropy which is defined in differential form by the quantity

$$dS = \frac{d\text{Q}}{T} \tag{13.26}$$

where $d\text{Q}$ is the injection of heat (or energy) into the radiation field in the box and T is the temperature at which that injection takes place. We would like to write $d\text{Q}$ in terms of u_{field}, V, and T. Then we may invoke the fact that S is a state variable, which implies

$$\frac{\partial^2 S}{\partial T \partial V} = \frac{\partial^2 S}{\partial V \partial T} \tag{13.27}$$

This is a mathematical statement of the fact that S is fully defined if the internal energy, temperature, and volume of system are specified. In other words, S does not depend on past temperature and volume history of a system, but is completely parameterized by the present state of the system.

To obtain $d\text{Q}$ in the form that we need, we can use the 1st law of thermodynamics, which is a statement of energy conservation:

$$\begin{aligned} d\text{Q} &= dU + PdV = d\left(u_{\text{field}} V\right) + PdV \\ &= V du_{\text{field}} + u_{\text{field}} dV + \frac{1}{3} u_{\text{field}} dV \\ &= V \frac{du_{\text{field}}}{dT} dT + \frac{4}{3} u_{\text{field}} dV \end{aligned} \tag{13.28}$$

Notice that we have used energy density times volume to obtain the total energy U in the radiation field in the box. We have also used (13.23) to obtain the work accomplished by pressure as the volume changes. A change in internal energy $dU = d\left(u_{\text{field}} V\right)$ can take place by the injection of heat $d\text{Q}$ or by doing work $dW = PdV$ as the volume increases. We can use (13.28) to rewrite (13.26):

$$d\text{S} = \frac{V}{T} \frac{du_{\text{field}}}{dT} dT + \frac{4u_{\text{field}}}{3T} dV \tag{13.29}$$

When we differentiate (13.29) with respect to temperature or volume we get

$$\begin{aligned} \frac{\partial S}{\partial T} &= \frac{V}{T} \frac{du_{\text{field}}}{dT} \\ \frac{\partial S}{\partial V} &= \frac{4u_{\text{field}}}{3T} \end{aligned} \tag{13.30}$$

We are now able to evaluate the partial derivatives in (13.27), which give

$$\begin{aligned} \frac{\partial^2 S}{\partial T \partial V} &= \frac{4}{3} \frac{\partial}{\partial T} \frac{u_{\text{field}}}{T} = \frac{4}{3} \frac{1}{T} \frac{\partial u_{\text{field}}}{\partial T} - \frac{4}{3} \frac{u_{\text{field}}}{T^2} \\ \frac{\partial^2 S}{\partial V \partial T} &= \frac{1}{T} \frac{du_{\text{field}}}{dT} \end{aligned} \tag{13.31}$$

Finally, (13.27) becomes a differential equation relating the internal energy of the system to the temperature:

$$\frac{4}{3} \frac{1}{T} \frac{\partial u_{\text{field}}}{\partial T} - \frac{4}{3} \frac{u_{\text{field}}}{T^2} = \frac{1}{T} \frac{du_{\text{field}}}{dT} \Rightarrow \frac{\partial u_{\text{field}}}{\partial T} = \frac{4u_{\text{field}}}{T} \tag{13.32}$$

The solution to this differential equation is (13.2), where $4\sigma/c$ is a constant to be determined experimentally (or derived from the Planck blackbody formula as was done in (13.16)).

Exercises

13.1 Introduction

P13.1 The Sun has a radius of $R_{\text{S}} = 6.96 \times 10^8$ m. What is the total power that it radiates, given a surface temperature of 5750 K?

P13.2 A 1 cm-radius spherical ball of polished gold hangs suspended inside an evacuated chamber that is at room temperature (20°C. There is no pathway for thermal conduction to the chamber wall.

(a) If the gold is at a temperature of 100°C, what is the *initial* rate of temperature loss in °C/s? The emissivity for polished gold is $e = 0.02$. The specific heat of gold is 129 J/kg · °C and its density is 19.3 g/cm^3.

HINT: $Q = mc\Delta T$ and $Power = Q/\Delta t$.

(b) What is the *initial* rate of temperature loss if the ball is coated with flat black paint, which has emissivity $e = 0.95$?

HINT: You should consider the energy flowing both ways.

13.3 Planck's Formula

P13.3 Derive (or try to derive) the Stefan-Boltzmann law by integrating the

(a) Rayleigh-Jeans energy density

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Rayleigh-Jeans}}(\nu)\, d\nu$$

Please comment.

(b) Wien energy density

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Wien}}(\nu)\, d\nu$$

Please evaluate σ.

HINT: $\int_0^\infty x^3 e^{-ax} dx = \frac{6}{a^4}$.

(c) Planck energy density

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Planck}}(\nu)\, d\nu$$

Please evaluate σ. Compare results of (b) and (c).

HINT: $\int_0^\infty \frac{x^3 dx}{e^{ax}-1} = \frac{\pi^4}{15a^4}$.

P13.4 (a) Derive Wien's displacement law

$$\lambda_{\max} = \frac{0.00290 \text{ m} \cdot \text{K}}{T}$$

which gives the strongest wavelength present in the blackbody spectral distribution.

HINT: Transform the integral to wavelength instead of frequency:

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Planck}}(\nu)\, d\nu \Rightarrow u_{\text{field}} = \int_0^\infty \rho_{\text{Planck}}(\lambda)\, d\lambda$$

Then find what λ corresponds to the maximum of $\rho_{\text{Planck}}(\lambda)$. You may like to know that the solution to the transcendental equation $(5 - x)\, e^x = 5$ is $x = 4.965$.

(b) What is the strongest wavelength emitted by the Sun, which has a surface temperature of 5750 K (see P 13.1)?

(c) Is $\lambda_{\max}$ the same as $c/\nu_{\max}$, where $\nu_{\max}$ corresponds to the peak of $\rho_{\text{Planck}}(\nu)$? Why would we be interested mainly in $\lambda_{\max}$?

Bibliography

[1] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. (Wiley, 1999).

[2] M. Born and E. Wolf, *Principles of Optics*, seventh ed. (Cambridge University Press, 1999).

[3] G. R. Fowles, *Introduction to Modern Optics*, 2nd ed. (Dover, 1975).

[4] J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, 1968).

[5] R. D. Guenther, *Modern Optics* (Wiley, 1990).

[6] P. W. Milonni and J. H. Eberly, *Lasers* (Wiley, 1988).

[7] P. W. Milonni, *The Quantum Vacuum: an Introduction to Quantum Electrodynamics* (Academic Press, 1994).

[8] J. R. Reitz, F. J. Milford, and R. W. Christy, *Foundations of Electromagnetic Theory*, fourth ed. (Addison-Wesley, 1992).

Index

Physical Constants

Constant	Symbol	Value
Permittivity	ϵ_0	8.854×10^{-12} $\mathrm{C^2/N \cdot m^2}$
Permeability	μ_0	$4\pi \times 10^{-7}$ $\mathrm{T \cdot m/A}$
Speed of light in vacuum	c	2.9979×10^{8} m/s
Charge of an electron	q_e	1.602×10^{-19} C
Mass of an electron	m_e	9.108×10^{-31} kg
Boltzmann's constant	k_{B}	1.380×10^{-23} J/K
Plancks constant	h	6.626×10^{-34} $\mathrm{J \cdot s}$
	$\hbar$	1.054×10^{-34} $\mathrm{J \cdot s}$
Stefan-Boltzmann constant	σ	5.670×10^{-8} $\mathrm{W/m^2 \cdot K^4}$