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

**Preface** iii

**Table of Contents** v

<table>
<thead>
<tr>
<th>0 Mathematical Tools</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>0.2 Vector Calculus</td>
<td>1</td>
</tr>
<tr>
<td>0.3 Complex Numbers</td>
<td>4</td>
</tr>
<tr>
<td>0.4 Fourier Theory</td>
<td>7</td>
</tr>
<tr>
<td>0.5 Linear Algebra and Sylvester’s Theorem</td>
<td>11</td>
</tr>
<tr>
<td>Appendix 0.A Integral and Sum Table</td>
<td>14</td>
</tr>
<tr>
<td>Exercises</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1 Electromagnetic Phenomena</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Introduction</td>
<td>21</td>
</tr>
<tr>
<td>1.2 Gauss’s Law and Coulomb’s Law</td>
<td>22</td>
</tr>
<tr>
<td>1.3 The Lorentz Force, Biot-Savart Law, and Gauss’s Law for Magnetic Fields</td>
<td>23</td>
</tr>
<tr>
<td>1.4 Faraday’s Law</td>
<td>24</td>
</tr>
<tr>
<td>1.5 Ampere’s Law</td>
<td>25</td>
</tr>
<tr>
<td>1.6 Maxwell’s Adjustment to Ampere’s Law</td>
<td>26</td>
</tr>
<tr>
<td>1.7 Polarization of Materials</td>
<td>28</td>
</tr>
<tr>
<td>1.8 The Macroscopic Maxwell Equations</td>
<td>30</td>
</tr>
<tr>
<td>1.9 The Wave Equation</td>
<td>30</td>
</tr>
<tr>
<td>Exercises</td>
<td>33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2 Plane Waves and Refractive Index</th>
<th>37</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Introduction</td>
<td>37</td>
</tr>
<tr>
<td>2.2 Plane Wave Solutions to the Wave Equation</td>
<td>38</td>
</tr>
<tr>
<td>2.3 Index of Refraction in Dielectrics</td>
<td>41</td>
</tr>
<tr>
<td>2.4 The Lorentz Model of Dielectrics</td>
<td>43</td>
</tr>
<tr>
<td>2.5 Conductor Model of Refractive Index and Absorption</td>
<td>46</td>
</tr>
<tr>
<td>2.6 Poynting’s Theorem</td>
<td>47</td>
</tr>
<tr>
<td>2.7 Irradiance of a Plane Wave</td>
<td>49</td>
</tr>
<tr>
<td>Appendix 2.A Energy Density of Electric Fields</td>
<td>51</td>
</tr>
</tbody>
</table>
## CONTENTS

### 6 Multiple Parallel Interfaces 137

6.1 Introduction ......................................................... 137
6.2 Double Boundary Problem Solved Using Fresnel Coefficients . . . 138
6.3 Double Boundary Problem at Sub Critical Angles ................. 142
6.4 Beyond Critical Angle: Tunneling of Evanescent Waves ....... 145
6.5 Fabry-Perot ............................................................. 147
6.6 Setup of a Fabry-Perot Instrument ............................. 150
6.7 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument . 152
6.8 Multilayer Coatings ................................................ 157
6.9 Repeated Multilayer Stacks ..................................... 161
Exercises .................................................................. 163

### 7 Superposition of Quasi-Parallel Plane Waves 169

7.1 Introduction ............................................................. 169
7.2 Intensity ................................................................ 171
7.3 Group vs. Phase Velocity: Sum of Two Plane Waves ........ 172
7.4 Frequency Spectrum of Light .................................... 175
7.5 Group Delay of a Wave Packet ................................. 180
7.6 Quadratic Dispersion ............................................... 182
7.7 Generalized Context for Group Delay ......................... 185
Appendix 7.A Causality and Exchange of Energy with the Medium ... 189
Exercises .................................................................. 195

### 8 Coherence Theory 199

8.1 Introduction ............................................................. 199
8.2 Michelson Interferometer ......................................... 200
8.3 Temporal Coherence ................................................ 202
8.4 Fringe Visibility and Coherence Length ...................... 204
8.5 Fourier Spectroscopy .............................................. 206
8.6 Young's Two-Slit Setup and Spatial Coherence ............. 208
Appendix 8.A Spatial Coherence with a Continuous Source .... 213
Appendix 8.B The van Cittert-Zernike Theorem ................ 215
Exercises .................................................................. 218

### Review, Chapters 6–8 223

### 9 Light as Rays 229

9.1 Introduction ............................................................. 229
9.2 The Eikonal Equation .............................................. 231
9.3 Fermat's Principle .................................................. 233
9.4 Paraxial Rays and ABCD Matrices ............................ 238
9.5 Reflection and Refraction at Curved Surfaces ............... 240
9.6 Image Formation by Mirrors and Lenses ..................... 244
9.7 Image Formation by Complex Optical Systems .......... 246
9.8 Stability of Laser Cavities ....................................... 248

© 2004-2009 Peatross and Ware
## Contents

9.9 Aberrations and Ray Tracing .................................................. 251
Exercises .................................................................................. 256

10 Diffraction ............................................................................. 261
  10.1 Huygens' Principle ............................................................ 261
  10.2 Scalar Diffraction ............................................................... 263
  10.3 Babinet's Principle .............................................................. 265
  10.4 Fresnel Approximation ....................................................... 266
  10.5 Fraunhofer Approximation ............................................... 267
  10.6 Diffraction with Cylindrical Symmetry ............................... 269
  Appendix 10.A Significance of the Scalar Wave Approximation .... 271
  Appendix 10.B Fresnel-Kirchhoff Diffraction Formula .............. 271
  Appendix 10.C Green's Theorem .............................................. 275
Exercises .................................................................................. 277

11 Diffraction Applications ......................................................... 281
  11.1 Introduction ...................................................................... 281
  11.2 Diffraction of a Gaussian Field Profile ............................... 282
  11.3 Gaussian Laser Beams ....................................................... 284
  11.4 Fraunhofer Diffraction Through a Lens ............................ 285
  11.5 Resolution of a Telescope ................................................ 290
  11.6 The Array Theorem .......................................................... 293
  11.7 Diffraction Grating ............................................................ 295
  11.8 Spectrometers ................................................................. 297
  Appendix 11.A ABCD Law for Gaussian Beams ....................... 299
Exercises .................................................................................. 303

Review, Chapters 9–11 ................................................................ 313

12 Interferograms and Holography .............................................. 319
  12.1 Introduction ...................................................................... 319
  12.2 Interferograms .................................................................. 319
  12.3 Testing Optical Components ............................................ 321
  12.4 Generating Holograms ...................................................... 324
  12.5 Holographic Wavefront Reconstruction ............................ 325
Exercises .................................................................................. 329

13 Blackbody Radiation ............................................................. 331
  13.1 Introduction ...................................................................... 331
  13.2 Failure of the Equipartition Principle .................................. 333
  13.3 Planck's Formula ............................................................... 335
  13.4 Einstein's A and B Coefficients .......................................... 338
Exercises .................................................................................. 342

References ................................................................................ 345

© 2004–2009 Peatross and Ware
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td>347</td>
</tr>
<tr>
<td>Physical Constants</td>
<td>350</td>
</tr>
</tbody>
</table>

© 2004-2009 Peatross and Ware
Chapter 0

Mathematical Tools

0.1 Introduction

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. This chapter reviews a few of the needed mathematical skills. This is not a comprehensive review, but rather a few selected topics that we have found that students often struggle with. We assume that the student already has a basic understanding of differentiation, integration, and standard trigonometric and algebraic manipulation. Students can usually just begin with Chapter 1. The text will refer back to sections in this chapter when it may be helpful to brush-up on some particular technique that is useful to understanding the topic at hand.

The topics in this chapter are essentially in the order they will be encountered in the text. 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 reviews complex arithmetic, and students need to know this material by heart. Section 0.4 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.

0.2 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 \( \mathbf{r} = x\mathbf{\hat{x}} + y\mathbf{\hat{y}} + z\mathbf{\hat{z}} \), where \( \mathbf{\hat{x}},\mathbf{\hat{y}}, \) and \( \mathbf{\hat{z}} \) 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 $\mathbf{E}(\mathbf{r}) = q (\mathbf{r} - \mathbf{r}_0) / 4\pi\varepsilon_0 |\mathbf{r} - \mathbf{r}_0|^3$, 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

$$|\mathbf{r} - \mathbf{r}_0| = \left| (x - x_0) \mathbf{\hat{x}} + (y - y_0) \mathbf{\hat{y}} + (z - z_0) \mathbf{\hat{z}} \right| = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}$$

(0.1)

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:

$$\mathbf{k} \cdot \mathbf{r} = (k_x \mathbf{\hat{x}} + k_y \mathbf{\hat{y}} + k_z \mathbf{\hat{z}}) \cdot (x \mathbf{\hat{x}} + y \mathbf{\hat{y}} + z \mathbf{\hat{z}})$$

$$= k_x x + k_y y + k_z z$$

$$= |\mathbf{k}| |\mathbf{r}| \cos \phi$$

(0.2)

where $\phi$ 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:

$$\mathbf{E} \times \mathbf{B} = \begin{vmatrix} \mathbf{\hat{x}} & \mathbf{\hat{y}} & \mathbf{\hat{z}} \\ E_x & E_y & E_z \\ B_x & B_y & B_z \end{vmatrix}$$

$$= \left( E_y B_z - E_z B_y \right) \mathbf{\hat{x}} - \left( E_x B_z - E_z B_x \right) \mathbf{\hat{y}} + \left( E_x B_y - E_y B_x \right) \mathbf{\hat{z}}$$

(0.3)

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

We will use several multidimensional derivatives in our study. The vector first derivatives are the gradient, the divergence, and the curl. In Cartesian coordinates, the gradient of a scalar is given by

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

(0.4)

the divergence is given by

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

(0.5)

the curl is given by

$$\nabla \times \mathbf{E} = \begin{vmatrix} \mathbf{\hat{x}} & \mathbf{\hat{y}} & \mathbf{\hat{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix}$$

$$= \left( \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \right) \mathbf{\hat{x}} - \left( \frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} \right) \mathbf{\hat{y}} + \left( \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) \mathbf{\hat{z}}$$

(0.6)
We will need a number of vector second derivatives when we study the wave equations describing light. (Vector second derivatives result from various combinations of vector first derivatives.) For example, the scalar Laplacian is defined as the divergence of a gradient

\[ \nabla^2 f (x, y, z) \equiv \nabla \cdot \left[ \nabla f (x, y, z) \right] \]  
(0.7)

and in cartesian coordinates is given by

\[ \nabla^2 f (x, y, z) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \]  
(0.8)

We will also use the vector Laplacian defined by

\[ \nabla^2 \mathbf{E} \equiv \nabla (\nabla \cdot \mathbf{E}) - \nabla \times (\nabla \times \mathbf{E}) \]  
(0.9)

which takes on a simple form in cartesian coordinates

\[ \nabla^2 \mathbf{E} = \left( \frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} \right) \hat{x} + \left( \frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial y^2} + \frac{\partial^2 E_y}{\partial z^2} \right) \hat{y} \]

\[ + \left( \frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + \frac{\partial^2 E_z}{\partial z^2} \right) \hat{z} \]  
(0.10)

(All of the multidimensional derivatives take on more complicated forms 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

\[ \oint_S \mathbf{f} \cdot \hat{n} \, da = \int_V \nabla \cdot \mathbf{f} \, dv \]  
(0.11)

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 \( \hat{n} \) 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:

\[ \oint_S (\nabla \times \mathbf{f}) \cdot \hat{n} \, da = \oint_C \mathbf{f} \cdot d\ell \]  
(0.12)

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, \( \hat{n} \) 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 \( \hat{n} \). 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:

$$\int_V [f(\nabla \cdot g) + (g \cdot \nabla) f] \, dv = \oint_S f(g \cdot \hat{n}) \, da$$  \hspace{1cm} (0.13)

0.3 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 $\alpha$ represent parameters. The sine function is intrinsically present in this formula through the identity

$$\cos(x + \alpha) = \cos x \cos \alpha - \sin x \sin \alpha$$  \hspace{1cm} (0.14)

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

$$\sin(x + \alpha) = \sin x \cos \alpha + \sin \alpha \cos x$$  \hspace{1cm} (0.15)

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 (0.14) and (0.15) 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:

$$e^{i\phi} = \cos \phi + i \sin \phi$$  \hspace{1cm} (0.16)

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

$$f(x) = f(x_0) + \frac{1}{1!} (x - x_0) \frac{df}{dx} \bigg|_{x=x_0} + \frac{1}{2!} (x - x_0)^2 \frac{d^2f}{dx^2} \bigg|_{x=x_0} + \cdots$$  \hspace{1cm} (0.17)

By expanding each function appearing in (0.16) in a Taylor’s series about the origin we obtain

$$\cos \phi = 1 - \frac{\phi^2}{2!} + \frac{\phi^4}{4!} - \cdots$$

$$i \sin \phi = i\phi - \frac{i \phi^3}{3!} + \frac{i \phi^5}{5!} - \cdots$$  \hspace{1cm} (0.18)

$$e^{i\phi} = 1 + i\phi - \frac{\phi^2}{2!} - i \frac{\phi^3}{3!} + \frac{\phi^4}{4!} + i \frac{\phi^5}{5!} - \cdots$$

© 2004–2009 Peatross and Ware
The last line of (0.18) is seen to be the sum of the first two lines, from which Euler’s formula directly follows.

By inverting Euler’s formula (0.16) we can obtain the following representation of the cosine and sine functions:

\[
\begin{align*}
\cos \phi &= \frac{e^{i\phi} + e^{-i\phi}}{2}, \\
\sin \phi &= \frac{e^{i\phi} - e^{-i\phi}}{2i}
\end{align*}
\] (0.19)

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

\[
\begin{align*}
\sin i \gamma &= \frac{e^{-\gamma} - e^{\gamma}}{2i} = i \sinh \gamma \\
\cos i \gamma &= \frac{e^{-\gamma} + e^{\gamma}}{2} = \cosh \gamma
\end{align*}
\] (0.20)

There are several situations in optics where one is interested in a complex angle, \( \phi = \beta + i \gamma \) where \( \beta \) and \( \gamma \) 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 \( \phi \) introduces exponential decay or growth as is apparent upon examination of (0.19). 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 \( \phi \) can be found which satisfies (0.19). 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

\[
A \cos (x + \alpha) = \Re \{ \tilde{A} e^{ix} \}
\] (0.21)

where the phase \( \alpha \) is conveniently contained within the complex factor \( \tilde{A} \equiv A e^{i\alpha} \). The operation \( \Re \{ \} \) means to retain only the real part of the argument without regard for the imaginary part. As an example, we have \( \Re (1 + 2i) = 1 \). The expression (0.21) is a direct result of Euler’s equation (0.16).

It is conventional in the study of optics to omit the explicit writing of \( \Re \{ \} \). 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 (0.14)). This laziness is permissible because it is possible to perform linear operations on \( \Re \{ f \} \) such as addition, differentiation, or integration.
while procrastinating the taking of the real part until the end:

\[
\text{Re}\{f\} + \text{Re}\{g\} = \text{Re}\{f + g\}
\]

\[
\frac{d}{dx} \text{Re}\{f\} = \text{Re}\left\{ \frac{df}{dx} \right\}
\]

\[
\int \text{Re}\{f\} \, dx = \text{Re}\left\{ \int f \, dx \right\}
\]

As an example, note that \(\text{Re}\{1 + 2i\} + \text{Re}\{3 + 4i\} = \text{Re}\{(1 + 2i) + (3 + 4i)\} = 4\). 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

\[
\text{Re}\{f\} \times \text{Re}\{g\} \neq \text{Re}\{f \times g\}
\]

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

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 \(\rho e^{i\phi}\), where \(a, b, \rho,\) and \(\phi\) are real. From (0.16), the required connection between \((\rho, \phi)\) and \((a, b)\) is

\[
\rho e^{i\phi} = \rho \cos \phi + i \rho \sin \phi = a + ib
\]

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

\[
a = \rho \cos \phi
\]

\[
b = \rho \sin \phi
\]

These equations can be inverted to yield

\[
\rho = \sqrt{a^2 + b^2}
\]

\[
\phi = \tan^{-1} \frac{b}{a} \quad (a > 0)
\]

When \(a < 0\), we must adjust \(\phi\) by \(\pi\) since the arctangent has a range only from \(-\pi/2\) to \(\pi/2\).

The transformations in (0.25) and (0.26) 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, \(\rho\) is the hypotenuse of a right triangle having legs with lengths \(a\) and \(b\), and \(\phi\) 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 \(\rho e^{i(\phi + \pi)}\).
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:

\[
z^* = (a + ib)^* = a - ib
\]  

(0.27)

The complex conjugate is useful when computing the magnitude \( \rho \) as defined in (0.26):

\[
|z| = \sqrt{z^*z} = \sqrt{(a - ib)(a + ib)} = \sqrt{a^2 + b^2} = \rho
\]  

(0.28)

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

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

(0.29)

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 (0.16). As another example consider \( E \exp\{i(kz - \omega t)\} \)\(^*\) = \( E^* \exp\{-i(\kappa^*z - \omega t)\} \), assuming \( z, \omega \), and \( t \) are real, but \( E \) and \( \kappa \) 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:

\[
\text{Re}\{z\} = \frac{1}{2}(z + z^*)
\]  

(0.30)

Notice that the expression for \( \cos \phi \) in (0.19) 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.

### 0.4 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.

© 2004-2009 Peatross and Ware
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:

\[ f(t) = \sum_{n=0}^{\infty} a_n \cos(n\Delta\omega t) + b_n \sin(n\Delta\omega t) \]  

This is called a Fourier expansion. It is similar in idea to a Taylor's series (0.17), 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 (0.31) 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 (0.31) repeat with a maximum period of \( \frac{2\pi}{\Delta\omega} \). 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 (0.31) using (0.19) as follows:

\[
\begin{align*}
  f(t) &= \sum_{n=0}^{\infty} a_n \frac{e^{in\Delta\omega t} + e^{-in\Delta\omega t}}{2} + b_n \frac{e^{in\Delta\omega t} - e^{-in\Delta\omega t}}{2i} \\
  &= a_0 + \sum_{n=1}^{\infty} \frac{a_n - ib_n}{2} e^{in\Delta\omega t} + \sum_{n=1}^{\infty} \frac{a_n + ib_n}{2} e^{-in\Delta\omega t} \\
  &= a_0 + \sum_{n=1}^{\infty} c_n e^{-in\Delta\omega t} \\
  &= \sum_{n=-\infty}^{\infty} c_n e^{-in\Delta\omega t}
\end{align*}
\]  

Thus, we can rewrite (0.31) as

\[ f(t) = \sum_{n=-\infty}^{\infty} c_n e^{-in\Delta\omega t} \]  

where

\[
\begin{align*}
  c_{n<0} &= \frac{a_{-n} - ib_{-n}}{2} \\
  c_{n>0} &= \frac{a_n + ib_n}{2} \\
  c_0 &= a_0
\end{align*}
\]

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 (0.31), and the imaginary parts of the \( c_n \) coefficients are connected with the sine terms in (0.31).

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 (0.33) by \( e^{im\Delta\omega t} \), where
m is an integer, and integrate over the function period $2\pi/\Delta\omega$:

\[
\int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t)e^{im\Delta\omega t} dt = \sum_{n=-\infty}^{\infty} c_n \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} e^{i(m-n)\Delta\omega t} dt
\]

\[
= \sum_{n=-\infty}^{\infty} c_n \left[ \frac{e^{i(m-n)\Delta\omega}}{i(m-n)\Delta\omega} \right]_{-\pi/\Delta\omega}^{\pi/\Delta\omega}
\]

\[
= \sum_{n=-\infty}^{\infty} 2\pi c_n \frac{e^{i(m-n)\pi} - e^{-i(m-n)\pi}}{2i(m-n)\pi}
\]

\[
= \sum_{n=-\infty}^{\infty} \frac{2\pi c_n}{\Delta\omega} \sin [(m-n)\pi] (m-n)\pi
\]

The function $\sin [(m-n)\pi] / [(m-n)\pi]$ 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 (0.35). We now have

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

(0.36)

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 (0.36), and, if we know the coefficients $c_n$, we can generate the function $f(t)$ via (0.33). If we are feeling a bit silly, we can combine these into a single identity:

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

(0.37)

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 (0.37) we obtain

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

(0.38)

At this point, a brief review of the definition of an integral is helpful to better understand the next step that we administer to (0.38). Recall that an integral is
really a summation of rectangles under a curve with finely spaced steps:

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

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 (0.39) becomes

\[
\int_{-\infty}^{\infty} g(\omega) \, d\omega = \lim_{\Delta \omega \to 0} \sum_{n=-\infty}^{\infty} g(n\Delta \omega) \Delta \omega
\]  

This concludes our short review of calculus.

We can use (0.40) in connection with (0.38) (where \(g(n\Delta \omega)\) represents everything in the square brackets). The result is the Fourier integral theorem:

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

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 (0.41) is often written with the following (potentially confusing) notation:

\[
f(\omega) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{i\omega t} \, dt
\]

\[
f(t) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\omega) e^{-i\omega t} \, d\omega
\]  

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 (0.42) is called the transform and which is called the inverse transform. In other words, the signs in the exponents of (0.42) may be interchanged. The convention varies in published works. Also, the factor \(2\pi\) 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

$$f(t) = \int_{-\infty}^{\infty} f(t') \delta(t' - t) \, dt' \quad (0.43)$$

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 (0.41) can be written as

$$f(t) = \int_{-\infty}^{\infty} f(t') \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega(t' - t)} \, d\omega \right] \, dt' \quad (0.44)$$

A comparison of (0.43) and (0.44) reveals the delta function to be a uniform superposition of all frequency components:

$$\delta(t' - t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega(t' - t)} \, d\omega \quad (0.45)$$

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.

### 0.5 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 $2 \times 2$ matrix is given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \frac{1}{AD - BC} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \quad (0.46)$$
This can be proven by direct substitution:

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}^{-1} = \frac{1}{AD-BC} \begin{bmatrix}
A & B \\
C & D
\end{bmatrix} = \frac{1}{AD-BC} \begin{bmatrix}
AD-BC & 0 \\
0 & AD-BC
\end{bmatrix}
\]

(0.47)

The next result is Sylvester’s Theorem, which is useful when a 2 \times 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 \times 2 matrix is one, i.e.

\[
\begin{vmatrix}
A & B \\
C & D
\end{vmatrix} = AD-BC = 1
\]

(0.48)

then the following holds:

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

(0.49)

where

\[
\cos \theta = \frac{1}{2} (A + D)
\]

(0.50)

Example 0.1

Prove the Sylvester’s theorem by induction.

Solution: 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 \( N+1 \):

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

Now we inject condition (0.48) on the determinant \((AD - BC) = 1\) into the right-hand side

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

and rearrange the result to give

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

and then

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

(© 2004-2009 Peatross and Ware)
In each matrix element, the expression
\[(A + D)\sin N\theta = 2\cos \theta \sin N\theta = \sin (N + 1)\theta + \sin (N - 1)\theta \quad (0.51)\]
occurs, which we have rearranged using \(\cos \theta = \frac{1}{2} (A + D)\) while twice invoking (0.15). The result is
\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}^{N+1} = \frac{1}{\sin \theta}
\begin{bmatrix}
A\sin (N + 1)\theta - \sin N\theta & B\sin (N + 1)\theta \\
C\sin (N + 1)\theta & D\sin (N + 1)\theta - \sin N\theta
\end{bmatrix}
\]
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.

\[
\int_{-\infty}^{\infty} e^{-ax^2+bx+c} \, dx = \sqrt{\frac{\pi}{-a}} e^{\frac{b^2}{4a}} + c \quad \text{Re} \{a\} > 0 \quad (0.52)
\]

\[
\int_{0}^{\infty} \frac{e^{iax}}{1 + x^2/b^2} \, dx = \frac{\pi |b|}{2} e^{-|ab|} \quad b > 0 \quad (0.53)
\]

\[
\int_{0}^{2\pi} e^{\pm ia\cos(\theta-\theta')} \, d\theta = 2\pi J_0(a) \quad (0.54)
\]

\[
\int_{0}^{a} J_0(bx) \, dx = \frac{a}{b} J_1(ab) \quad (0.55)
\]

\[
\int_{0}^{\infty} e^{-ax^2} J_0(bx) \, dx = \frac{e^{-b^2/4a}}{2a} \quad (0.56)
\]

\[
\int_{0}^{\infty} \frac{\sin^2(ax)}{(ax)^2} \, dx = \frac{\pi}{2a} \quad (0.57)
\]

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

\[
\sum_{n=1}^{N} ar^n = a \frac{1 - r^N}{1 - r} \quad (0.59)
\]

\[
\sum_{n=1}^{\infty} ar^n = \frac{a}{1 - r} \quad (r < 1) \quad (0.60)
\]
Exercises

Exercises for 0.2 Vector Calculus

P0.1 Let \( \mathbf{r} = (\hat{x} + 2\hat{y} - 3\hat{z}) \) m and \( \mathbf{r}_0 = (-\hat{x} + 3\hat{y} + 2\hat{z}) \) 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°.

P0.2 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 (0.2). Call it the \( xy \)-plane. In this coordinate system, the two vectors can be written as \( \mathbf{r}_k = k \cos \theta \hat{x} + k \sin \theta \hat{y} \) and \( \mathbf{r} = r \cos \alpha \hat{x} + r \sin \alpha \hat{y} \), where \( \theta \) and \( \alpha \) are the respective angles that the two vectors make with the \( x \)-axis. The dot product gives \( \mathbf{k} \cdot \mathbf{r} = kr \cos (\theta - \alpha) \), which shows that \( \theta - \alpha \) is the angle between the vectors.

P0.3 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.

P0.4 Verify the “BAC-CAB” rule: \( \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B} (\mathbf{A} \cdot \mathbf{C}) - \mathbf{C} (\mathbf{A} \cdot \mathbf{B}) \).

P0.5 Prove the following identity:

\[
\nabla \cdot \frac{1}{|\mathbf{r} - \mathbf{r}'|} = -\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3},
\]

where \( \nabla \) operates only on \( \mathbf{r} \), treating \( \mathbf{r}' \) as a constant vector.

P0.6 Prove that \( \nabla \cdot \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \) is zero, except at \( \mathbf{r} = \mathbf{r}' \) where a singularity situation occurs.

P0.7 Verify \( \nabla \cdot (\nabla \times \mathbf{f}) = 0 \) for any vector function \( \mathbf{f} \).

P0.8 Verify \( \nabla \times (\nabla \times \mathbf{f}) = \nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f} \)

Solution: From (0.6), we have

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

After rearranging, we get
\[
\nabla \times (\nabla \times f) = \left[ \frac{\partial^2 f_x}{\partial x^2} + \frac{\partial^2 f_y}{\partial y^2} + \frac{\partial^2 f_z}{\partial z^2} \right] \hat{x} + \left[ \frac{\partial^2 f_x}{\partial x \partial y} + \frac{\partial^2 f_y}{\partial y \partial z} + \frac{\partial^2 f_z}{\partial z \partial x} \right] \hat{y} + \left[ \frac{\partial^2 f_x}{\partial x \partial z} + \frac{\partial^2 f_y}{\partial y \partial x} + \frac{\partial^2 f_z}{\partial z \partial y} \right] \hat{z}
\]

where we have added and subtracted \( \frac{\partial^2 f_x}{\partial x \partial y} + \frac{\partial^2 f_y}{\partial y \partial z} + \frac{\partial^2 f_z}{\partial z \partial x} \). After some factorization, we obtain
\[
\nabla \times (\nabla \times f) = \nabla (\nabla \cdot f) - \nabla^2 f
\]

where on the final line we invoked (0.4), (0.5), and (0.8).

\[\text{P0.9} \quad \text{Verify } \nabla \times (f \times g) = f (\nabla \cdot g) - g (\nabla \cdot f) + (g \cdot \nabla) f - (f \cdot \nabla) g.\]

\[\text{P0.10} \quad \text{Verify } \nabla \cdot (f \times g) = g \cdot (\nabla \times f) - f \cdot (\nabla \times g).\]

\[\text{P0.11} \quad \text{Verify } \nabla \cdot (gf) = f \cdot \nabla g + g \nabla \cdot f.\]

\[\text{P0.12} \quad \text{Verify } \nabla \times (gf) = (\nabla g) \times f + g \nabla \times f.\]

\[\text{P0.13} \quad \text{Verify the divergence theorem (0.11) for } f(x, y, z) = y^2 \hat{x} + xy \hat{y} + x^2 z \hat{z}. \text{ Take as the volume a cube contained by the six planes } |x| = \pm 1, |y| = \pm 1, \text{ and } |z| = \pm 1.\]

\[
\int \int \int_S f \cdot n \, da = \int_{-1}^{1} \int_{-1}^{1} \int_{-1}^{1} dxdy \left| x^2 z \right|_{z=1} - \int_{-1}^{1} \int_{-1}^{1} dxdy \left| x^2 z \right|_{z=-1} + \int_{-1}^{1} \int_{-1}^{1} dxdy \left| xy \right|_{y=1} - \int_{-1}^{1} \int_{-1}^{1} dxdy \left| xy \right|_{y=-1}
\]
\[
= \int_{-1}^{1} \int_{-1}^{1} dxdy x^2 + 2 \int_{-1}^{1} \int_{-1}^{1} dx \left| x^3 \right|_{x=-1} + 2 \int_{-1}^{1} \int_{-1}^{1} dy \left| x^2 \right|_{x=-1} = \int_{-1}^{1} \int_{-1}^{1} dxdy x^2 + \int_{-1}^{1} dx \left| x^3 \right|_{x=-1} + 2 \int_{-1}^{1} dy \left| x^2 \right|_{x=-1} = \frac{8}{3}.
\]

\[
\int \nabla \cdot f dv = \frac{1}{8} \int \int \int dxdydz \left[ x + x^2 \right] = \frac{1}{8} \int \frac{1}{2} dx \left[ x + x^2 \right] = \frac{1}{8} \left[ \frac{x^2}{2} + \frac{x^3}{3} \right]_{-1} = \frac{8}{3}.
\]
P0.14 Verify Stokes’ theorem (0.12) for the function given in P 0.13. Take the surface to be a square in the $xy$-plane contained by $|x| = \pm 1$ and $|y| = \pm 1$.

P0.15 Use the divergence theorem to show that the function in P 0.6 is $4\pi$ times the three-dimensional delta function.

Solution: We have by the divergence theorem

$$\oint_{S} \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \cdot \hat{n} da = \int_{V} \nabla \cdot \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} dv$$

From P 0.6, 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

$$\oint_{S_2} \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \cdot \hat{n} da = -2\pi \int_{0}^{\pi} \int_{0}^{\pi/2} r'^2 \sin \phi d\phi da = -4\pi$$

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

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

This implies

$$\int_{V} \nabla \cdot \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} dv = \begin{cases} 4\pi & \text{if } V \text{ contains } \mathbf{r}' \\ 0 & \text{otherwise} \end{cases}$$

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, $\nabla \cdot \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} = 4\pi \delta^3(\mathbf{r} - \mathbf{r}')$. The delta function is defined in (0.43)

Exercises for 0.3 Complex Numbers

P0.16 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.
P0.17 Show that $-3 + 4i$ can be written as $5\exp\{-i \tan^{-1} 4/3 + i\pi\}$.

P0.18 Show $(a - ib)/(a + ib) = \exp\{-2i \tan^{-1} b/a\}$ regardless of the sign of $a$, assuming $a$ and $b$ are real.

P0.19 Invert (0.16) to get both formulas in (0.19).

P0.20 Show $\text{Re}\{A\times B\} = (AB + A^* B)/4$ regardless of the sign of $a$, assuming $a$ and $b$ are real.

P0.21 If $E = |E| e^{i\alpha_E}$ and $B = |B| e^{i\alpha_B}$, and if $k$, $z$, $\omega$, and $t$ are all real, prove

$\text{Re}\left\{ E e^{i(kz - \omega t)} \right\} \text{Re}\left\{ B e^{i(kz - \omega t)} \right\} = \frac{1}{4} (E^* B + EB^*) + \frac{1}{2} |E||B| \cos[2(kz - \omega t) + \alpha_E + \alpha_B].$

P0.22 (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 $\phi$ in (a) is $\pi/2 - i \ln(2 + \sqrt{3})$.

P0.23 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) $3\cos(5t) + 10\sin(5t + 0.4)$

Exercises for 0.4 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.
The following operation is referred to as the convolution of the functions \( g(t) \) and \( h(t) \):

\[
g(t) \otimes h(t) \equiv \int_{-\infty}^{\infty} g(t) h(t - \tau) \, dt
\]

A convolution measures the overlap of \( g(t) \) and a reversed \( h(t) \) as a function of the offset \( \tau \).

(a) Prove the convolution theorem:

\[
\mathcal{F} \{ g(t) \otimes h(t) \} = \sqrt{2\pi} \mathcal{F} \{ g(t) \} \mathcal{F} \{ h(t) \}
\]

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

\[
\mathcal{F} \{ g(t) h(t) \} = \frac{1}{\sqrt{2\pi}} \mathcal{F} \{ g(t) \} \otimes \mathcal{F} \{ h(t) \}
\]

Solution: Part (a)

\[
\mathcal{F} \left\{ \int_{-\infty}^{\infty} g(t) h(t - \tau) \, dt \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} g(t) h(\tau - t) \, dt \right\} e^{i\omega \tau} \, d\tau \\
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} g(t) h(\tau + t') \, dt' \right\} e^{i\omega(\tau + t')} \, d\tau' \\
= \sqrt{2\pi} \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(t) e^{i\omega t} \, dt \right) \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t') e^{i\omega t'} \, dt' \right) \\
= \sqrt{2\pi} G(\omega) H(\omega)
\]

Prove the autocorrelation theorem:

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

Prove Parseval’s theorem:

\[
\int_{-\infty}^{\infty} |f(\omega)|^2 \, d\omega = \int_{-\infty}^{\infty} |f(t)|^2 \, dt
\]

(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 (0.45).
(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 1800s James Maxwell assembled the various known relationships of electricity and magnetism into a concise\(^1\) set of equations:

\[
\begin{align*}
\nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0} \quad \text{(Gauss's Law from Coulomb's Law)} \\
\nabla \cdot \mathbf{B} &= 0 \quad \text{(Gauss's Law for magnetism from Biot-Savart)} \\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \quad \text{(Faraday's Law)} \\
\n\nabla \times \frac{\mathbf{B}}{\mu_0} &= \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J} \quad \text{(Ampere's Law revised by Maxwell)}
\end{align*}
\]

Where \( \mathbf{E} \) and \( \mathbf{B} \) represent electric and magnetic fields, respectively. The charge density \( \rho \) describes the charge per volume distributed through space. The current density \( \mathbf{J} \) describes the motion of charge density (in units of \( \rho \) times velocity). The constant \( \epsilon_0 \) is called the permittivity, and the constant \( \mu_0 \) is called the permeability.

After introducing a key revision 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. Once the connection was made, it became clear that Maxwell's equations form the theoretical foundation for optics, and this is where we begin our study of light.

In this chapter, we review the physical principles associated with each of Maxwell's equations and illustrate the connection between electromagnetic phe-

\(^1\)In Maxwell's original notation, this set of equations was hardly concise. Lacking the convenience of modern vector notation, he wrote them as twenty equations in twenty variables. (Try fitting that on a T-shirt!)
nomena and light. While many of the details discussed in this chapter (e.g. static fields and magnetic effects) are not directly used in later chapters, they are included to better appreciate the basic physics that Maxwell’s equations describe. It may be helpful to study the vector calculus review in section 0.2 before beginning this chapter.

1.2 Gauss’s Law and Coulomb’s Law

The force on a point charge \( q \) located at \( \mathbf{r} \) exerted by another point charge \( q' \) located at \( \mathbf{r}' \) is

\[
\mathbf{F} = q \mathbf{E}(\mathbf{r})
\]  

(1.5)

where

\[
\mathbf{E}(\mathbf{r}) = \frac{q'}{4\pi\varepsilon_0} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}
\]  

(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 the second charge \( q' \) is actually present). The permittivity \( \varepsilon_0 \) amounts to a proportionality constant.

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

\[
\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int_V \rho(\mathbf{r}') \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \, dV
\]  

(1.7)

This three-dimensional integral gives the net electric field produced by the charge density \( \rho \) 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

\[
\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}
\]  

(1.8)

(see Example 1.1 for details). No new physical phenomenon is introduced by writing Gauss’s law. It is simply a mathematical interpretation of Coulomb’s law.

Example 1.1

Derive Gauss’s law from (1.7).
Solution: To derive Gauss’s law, we take the divergence of (1.7):

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int_V \rho(\mathbf{r}') \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \, d\mathbf{r}'$$

(1.9)

The subscript on $\nabla_r$ 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 $\rho(\mathbf{r}')$. In modern mathematical language, the vector expression in the integral is a three-dimensional delta function:

$$\nabla_r \cdot \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \equiv 4\pi \delta^3 (\mathbf{r}' - \mathbf{r}) \equiv 4\pi \delta (x' - x) \delta (y' - y) \delta (z' - z)$$

(1.10)

A derivation of this formula and a description of its properties are addressed in problem P 0.15. The delta function allows the integral in (1.9) to be performed, and the relation becomes simply

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\varepsilon_0}$$

(1.11)

which is the differential form of Gauss’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 (0.11) to the left-hand side:

$$\oint_S \mathbf{E}(\mathbf{r}) \cdot \hat{n} \, d\mathbf{a} = \frac{1}{\varepsilon_0} \int_V \rho(\mathbf{r}) \, d\mathbf{v}$$

(1.12)

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 The Lorentz Force, Biot-Savart Law, and Gauss’s Law for Magnetic Fields

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

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

(1.13)

where

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

(1.14)
The first equation is known as the Lorentz force for a magnetic field, and the latter equation is referred to as the Biot-Savart law. The permeability $\mu_0$ 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 0.5, we can rewrite (1.14) as

$$\mathbf{B}(\mathbf{r}) = -\frac{\mu_0}{4\pi} \int_{V} \mathbf{J}(\mathbf{r}') \times \nabla_{\mathbf{r}} \frac{1}{|\mathbf{r} - \mathbf{r'}|} \, d\mathbf{v}'$$

$$= \frac{\mu_0}{4\pi} \nabla \times \int_{V} \mathbf{J}(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r'}|} \, d\mathbf{v}'$$

(1.15)

Since the divergence of a curl is identically zero, taking the divergence of (1.15) gives (see P 0.7)

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

(1.16)

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, (1.16) is known as Gauss’s law for magnetic fields.

In integral form, Gauss’s law for magnetic fields looks like that for electric fields (1.12), 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 observed in nature, there is no point in introducing them.

### 1.4 Faraday’s Law

Michael Faraday discovered and characterized the relationship between changing magnetic fluxes and induced electric fields. This effect, called induction, can be observed by moving a magnet around near a loop of wire and noting that the changing magnetic field induces an electric field (usually detected by the current it causes in a wire loop). 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

$$\oint_{C} \mathbf{E} \cdot d\ell = -\frac{\partial}{\partial t} \int_{S} \mathbf{B} \cdot \hat{n} \, da$$

(1.17)

This relation is known as Faraday’s law, the third in our list of Maxwell’s equations. In (1.17), Faraday’s law is written in integral form. The right side describes a change in the magnetic flux through a surface and the left side describes an electric field produced on the boundary of this surface.
To obtain the differential form of Faraday’s law, we apply Stokes’ theorem to the left-hand side and obtain

$$\int_S \nabla \times \mathbf{E} \cdot \hat{n} \, da = -\frac{\partial}{\partial t} \int_S \mathbf{B} \cdot \hat{n} \, da$$  \hspace{1cm} (1.18)

or

$$\int_S \left( \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} \right) \cdot \hat{n} \, da = 0$$  \hspace{1cm} (1.19)

Since this equation is true regardless of what surface we choose, it implies

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$  \hspace{1cm} (1.20)

which is the differential form of Faraday’s law.

### 1.5 Ampere’s Law

The Biot-Savart law (1.14) can also be used to obtain another of Maxwell’s equations: Ampere’s law. Ampere’s law is obtained by inverting the Biot-Savart law (1.14) 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 process for inverting the Biot-Savart law is given in Example 1.2. The result is

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$$  \hspace{1cm} (1.21)

which is the differential form of Ampere’s law.

#### Example 1.2

Obtain Ampere’s law from the Biot-Savart law under the assumption \(\nabla \cdot \mathbf{J} \equiv 0\).

**Solution:** We take the curl of (1.14):

$$\nabla \times \mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_V \nabla_{\mathbf{r}} \times \left[ \mathbf{J}(\mathbf{r}') \times \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \right] \, d\mathbf{v}'$$  \hspace{1cm} (1.22)

We next apply the differential vector rule from P 0.9 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

$$\nabla \times \mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_V \left[ \mathbf{J}(\mathbf{r}') \left( \nabla_{\mathbf{r}} \cdot \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \right) - \mathbf{J}(\mathbf{r}') \left( \nabla_{\mathbf{r}} \cdot \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \right) \right] \, d\mathbf{v}'$$  \hspace{1cm} (1.23)

According to (1.10), 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

© 2004-2009 Peatross and Ware
be changed to operate on the primed variables without affecting the final result (i.e. \( \nabla_r \rightarrow -\nabla_{r'} \)). In addition, we take advantage of the vector integral theorem (0.13) to arrive at

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

(1.24)

The last term in (1.24) 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 \( \mathbf{B}(r) \) reduces to

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

(1.25)

The latter term in (1.25) vanishes if \( \nabla \cdot \mathbf{J} \equiv 0 \), yielding Ampere’s law (1.21).

The integral form of Ampere’s law can be obtained by integrating both sides of (1.21) over an open surface \( S \), bounded by contour \( C \). Stokes’ theorem (0.12) is applied to the left-hand side of Ampere’s law to get

\[
\oint_C \mathbf{B}(r) \cdot d\ell = \mu_0 \int_S \mathbf{J}(r) \cdot \mathbf{n} \ da \equiv \mu_0 I
\]

(1.26)

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. 1.5). 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.

It is important to note that during the derivation of Ampere’s law (1.21) one must make the approximation

\[
\nabla \cdot \mathbf{J} \equiv 0 \quad \text{(steady-state approximation)}
\]

(1.27)

(see Example 1.2). This approximation is valid only if the current density \( \mathbf{J} \) does not vary rapidly in time, which is not true in general (especially for optical phenomena). Thus, we need to do some additional work to get Ampere’s law into a form that is true in non-steady-state situations.

### 1.6 Maxwell’s Adjustment to Ampere’s Law

Maxwell was the first to realize that Ampere’s law is incomplete as written in (1.21) when the current density \( \mathbf{J} \) varies dynamically in time. To arrive at the correct equation, we need to understand what the expression \( \nabla \cdot \mathbf{J} \) (i.e. the term that we neglected) represents. Consider a volume of space enclosed by a surface \( S \) through which current is flowing. The total current exiting the volume is

\[
I = \oint_S \mathbf{J} \cdot \mathbf{n} \ da
\]

(1.28)
where \( \hat{n} \) is the outward normal to the surface. 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:

\[
I = -\frac{\partial}{\partial t} \int_V \rho \, dV \tag{1.29}
\]

Upon equating these two expressions for current, as well as applying the divergence theorem (0.11) to the former, we get

\[
\int_V \nabla \cdot J \, dV = -\int_V \frac{\partial \rho}{\partial t} \, dV \tag{1.30}
\]

or

\[
\int_V \left( \nabla \cdot J + \frac{\partial \rho}{\partial t} \right) \, dV = 0 \tag{1.31}
\]

Since (1.31) is true regardless of which volume \( V \) we choose, it implies the following continuity equation:

\[
\nabla \cdot J = -\frac{\partial \rho}{\partial t} \tag{1.32}
\]

The continuity equation (1.32) is simply a statement of the conservation of charge. It requires that the charge inside a volume must decrease in time if we are to have a net current flowing out of the volume. This is not a concern in the steady-state situation (where the form of Ampere’s law derived in the previous section applies) since a steady current has equal amounts of charge flowing both into and out of any particular volume (i.e. \( \frac{\partial \rho}{\partial t} = 0 \)).

Maxwell’s main contribution (aside from organizing other people’s formulas) was the injection of the continuity equation (1.32) into the derivation of Ampere’s law to make it applicable to dynamical situations. If we insert the continuity equation (1.32) into the derivation of Ampere’s law (see Example 1.3), we find the generalized form of Ampere’s law:

\[
\nabla \times \frac{\mathbf{B}}{\mu_0} = J + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{1.33}
\]

The final term is known as the displacement current (density), which exists even in the absence of any actual charge density \( \rho \). It indicates that a changing electric field behaves like a current in the sense that it produces magnetic fields. The similarity between Faraday’s law and the corrected Ampere’s law is apparent, and no doubt played a part in motivating Maxwell’s work.

### Example 1.3

Use the continuity equation and the Biot-Savert law to derive the corrected form of Ampere’s law.
Solution: The derivation proceeds as in Example 1.2 until we reach (1.25). Inserting the continuity equation into (1.25) yields:

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

Then substitution of (1.7) into this formula gives

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

In summary, in the previous section we saw that the basic physics in Ampere’s law is present in the Biot-Savart law (the two laws are connected through mathematics). Adding the requirement of charge conservation yields the corrected form, which can be used with rapidly varying fields.

### 1.7 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 \( \rho \). 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 \( \mathbf{J}_p \) to distinguish it from free currents. The total current (ignoring magnetic effects) is then

$$\mathbf{J} = \mathbf{J}_{\text{free}} + \mathbf{J}_p \quad (1.34)$$

The polarization current \( \mathbf{J}_p \) 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. 1.6) change their orientation as a function of time in some coordinated fashion, an effective current density results. (In a static \( \mathbf{P}(\mathbf{r}) \) no charges are moving, so there is no current.) Since the time-derivative of dipole moments renders charge times velocity, a distribution of “sloshing” dipoles gives a current density equal to

$$\mathbf{J}_p = \frac{\partial \mathbf{P}}{\partial t} \quad (1.35)$$

With this, Maxwell’s equation (1.33) becomes

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \mathbf{J}_{\text{free}} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} \quad (1.36)$$
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 $P(r)$. The degree of polarization can vary spatially 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

$$\nabla \cdot J_p = -\frac{\partial \rho_p}{\partial t}$$  \hspace{1cm} (1.37)$$

where $\rho_p$ is the charge density created by variations in the polarization $P(r)$. Substitution of (1.35) into this equation yields an expression for the resulting charge density $\rho_p$:

$$\rho_p = -\nabla \cdot P$$  \hspace{1cm} (1.38)

To further appreciate local charge buildup due to variation medium polarization, consider the divergence theorem (0.11) applied to $P(r)$ in a neutral medium:

$$-\oint_S P(r) \cdot \hat{n} \, da = -\int_V \nabla \cdot P(r) \, dv$$  \hspace{1cm} (1.39)$$

The left-hand side of (1.39) 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. 1.6. Keep in mind that $P(r)$ is a continuous function so that Fig. 1.6 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 $\nabla \cdot P$ is zero, there are equal numbers of positive and negative charges touching $S$ from within. When $\nabla \cdot P$ 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 (1.39) is equal to the excess charge inside the volume, $-\nabla \cdot P$ may be interpreted as a charge density (it certainly has the right units—charge per volume), in agreement with (1.38). Again, the negative sign occurs since when $P$ points out of the surface $S$, negative charges are left inside.

The total charge density thus can be written as

$$\rho = \rho_{\text{free}} + \rho_p$$  \hspace{1cm} (1.40)$$

With (1.38), Gauss’s law (1.8) becomes

$$\nabla \cdot E = \frac{\rho_{\text{free}} - \nabla P}{\varepsilon_0}$$  \hspace{1cm} (1.41)$$

For typical optics problems (involving neutral materials), we have $\rho_{\text{free}} = 0$. 

© 2004-2009 Peatross and Ware
1.8 The Macroscopic Maxwell Equations

In summary, in electrically neutral non-magnetic materials, Maxwell’s equations are
\[ \nabla \cdot \mathbf{E} = -\frac{\nabla \cdot \mathbf{P}}{\varepsilon_0} \quad \text{(Coulomb's law \Rightarrow Gauss's law)} \tag{1.42} \]
\[ \nabla \cdot \mathbf{B} = 0 \quad \text{(Biot-Savart law \Rightarrow Gauss's law for magnetism)} \tag{1.43} \]
\[ \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad \text{(Faraday's law)} \tag{1.44} \]
\[ \nabla \times \mathbf{B} = \frac{\varepsilon_0}{\mu_0} \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} + \mathbf{J}_{\text{free}} \quad \text{(Ampere's law; fixed by Maxwell)} \tag{1.45} \]

Notice that the assumption of an electrically neutral material dismisses the possibility of a free charge density \( \rho_{\text{free}} \), but 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.

It is common to see the macroscopic Maxwell equations written in terms of two auxiliary fields: \( \mathbf{H} \) and \( \mathbf{D} \). For the benefit of those who have used these auxiliary fields before, we take a moment to review their definitions and explain why we don't use them in this book. The field \( \mathbf{H} \) is useful in magnetic materials. In these materials, the combination \( \mathbf{B} / \mu_0 \) in Ampere’s law is replaced by \( \mathbf{H} \equiv \mathbf{B} / \mu_0 - \mathbf{M} \), where \( \mathbf{M} \) is the material’s magnetization. Since we only consider nonmagnetic materials \( (\mathbf{M} = 0) \), there is little point in using \( \mathbf{H} \). The field \( \mathbf{D} \), called the displacement, is defined as \( \mathbf{D} \equiv \varepsilon_0 \mathbf{E} + \mathbf{P} \). This combination of \( \mathbf{E} \) and \( \mathbf{P} \) occurs in Coulomb’s law and Ampere’s law. For the purposes of this book, it is conceptually more clear to retain the polarization \( \mathbf{P} \) as a separate field in these two equations. For instance, in isotropic media \( \nabla \cdot \mathbf{P} \) is zero in (1.42) while \( \partial \mathbf{P}/\partial t \) is non-zero in (1.45).

1.9 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 \( 1/\sqrt{\varepsilon_0 \mu_0} \) gives the correct speed of light \( c = 3.00 \times 10^8 \) 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 \( \mathbf{E} \) and \( \mathbf{B} \)) into the familiar wave equation (second order differential equations for either \( \mathbf{E} \) or \( \mathbf{B} \), 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 (1.44), from which we obtain

\[
\nabla \times (\nabla \times E) + \frac{\partial}{\partial t} (\nabla \times B) = 0 \quad (1.46)
\]

Then we eliminate \( B \) by substitution for \( \nabla \times B \) from (1.45) to obtain the wave equation for the electric field:

\[
\nabla \times (\nabla \times E) + \mu_0 \varepsilon_0 \frac{\partial^2 E}{\partial t^2} = -\mu_0 \frac{\partial J_{\text{free}}}{\partial t} - \mu_0 \frac{\partial^2 P}{\partial t^2} \quad (1.47)
\]

We can put the wave equation into a more familiar form using the differential vector identity (see P 0.8):

\[
\nabla \times (\nabla \times E) = \nabla (\nabla \cdot E) - \nabla^2 E \quad (1.48)
\]

Using this identity with (1.47), and applying Coulomb's law (1.42) to the \( \nabla (\nabla \cdot E) \) term, we get

\[
\nabla^2 E - \mu_0 \varepsilon_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial J_{\text{free}}}{\partial t} + \mu_0 \frac{\partial^2 P}{\partial t^2} - \frac{1}{\varepsilon_0} \nabla (\nabla \cdot P) \quad (1.49)
\]

The left-hand side of (1.49) 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 (1.49) 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 (1.49) describes dipole oscillations, which behave similar to currents. In a non-conducting optical material such as glass, the free current is zero, but \( \partial^2 P / \partial t^2 \) 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 (1.49) is important in anisotropic media such as crystals. 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 (1.49) are zero. However, usually at least one of the terms must be retained when considering propagation in a medium other than vacuum.

After solving the wave equation (1.49) for \( E \), one may obtain \( B \) through an application of Faraday's law (1.44). Even though the magnetic field \( B \) satisfies a similar wave equation, decoupled from \( E \) (see P 1.7), the two waves are not independent. The fields for \( E \) and \( B \) must be chosen to be consistent with each other through Maxwell's equations.

In vacuum all of the terms on the right-hand side in (1.49) are zero, in which case the equation reduces to

\[
\nabla^2 E - \mu_0 \varepsilon_0 \frac{\partial^2 E}{\partial t^2} = 0 \quad \text{(vacuum)} \quad (1.50)
\]
The solutions to the vacuum wave equation (1.50) propagate with speed

\[ c \equiv \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = 2.9979 \times 10^8 \text{ m/s} \quad \text{(vacuum)} \quad (1.51) \]

and any function \( E \) is a valid solution as long as it carries the dependence on the argument \( \hat{u} \cdot r - ct \), where \( \hat{u} \) is a unit vector specifying the direction of propagation. The argument \( \hat{u} \cdot r - ct \) preserves the shape of the waveform as it propagates in the \( \hat{u} \) direction; features occurring at a given position recur 'downstream' at a distance \( ct \) after a time \( t \). By checking this solution in (1.50), 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.
Exercises

Exercises for 1.1 Introduction

P1.1 Suppose that an electric field is given by \( \mathbf{E}(r, t) = E_0 \cos \left( k \cdot r - \omega t + \phi \right) \), where \( k \perp E_0 \) and \( \phi \) is a constant phase. Show that \( \mathbf{B}(r, t) = \frac{k \times E_0}{\omega} \cos \left( k \cdot r - \omega t + \phi \right) \) is consistent with (1.3).

Exercises for 1.2 Gauss's Law and Coulomb's Law

P1.2 Consider an infinitely long hollow cylinder (inner radius \( a \), outer radius \( b \)) which carries a volume charge density \( \rho = \frac{k}{s^2} \) 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. 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.

Exercises for 1.5 Ampere's Law

P1.3 A conducting cylinder with the same geometry as P 1.2 carries a volume current density \( \mathbf{J} = \frac{k}{s} \mathbf{\hat{z}} \) 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.

Exercises for 1.6 Maxwell's Adjustment to Ampere's Law

P1.4 (a) Use Gauss's law to find the electric field in a gap introduced in a current-carrying wire, as shown in Fig. 1.8. 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 electric field is essentially zero except in the gap.

(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 $\varepsilon_0 \partial E / \partial t$ by the cross-sectional area to obtain a “current”. The current in the wire is related to the charge $Q$ through $I = \partial Q / \partial t$.

**Exercises for 1.8 The Macroscopic Maxwell Equations**

**P1.5** Memorize the Maxwell equations (1.1)-(1.4) and the macroscopic Maxwell equations (1.42)-(1.45). 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. Also write the assumptions made in writing the macroscopic Maxwell equations in this way.

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

**Exercises for 1.9 The Wave Equation**

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

**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 $E (\hat{u} \cdot \mathbf{r} - ct)$ satisfies the vacuum wave equation (1.50), where $E$ is an arbitrary functional form.

**P1.10** (a) Show that $E (\mathbf{r}, t) = E_0 \cos \left( k (\hat{u} \cdot \mathbf{r} - ct) + \phi \right)$ is a solution to the vacuum wave equation (1.50), where $\hat{u}$ is an arbitrary unit vector and $k$ is a constant with units of inverse length.

(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 $\hat{u}$ direction. HINT: Set the cosine argument to a constant, solve for $\mathbf{r}$, and differentiate.

(d) By analysis, determine the wavelength $\lambda$ in terms of $k$. HINT: Find the distance between identical wave fronts by changing the cosine argument by $2\pi$ at a given instant in time.

(e) Use (1.42) to show that $E_0$ and $\hat{u}$ must be perpendicular to each other in vacuum.
L1.11 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 1.9 A schematic of the setup for lab 1.11.

Figure 1.10 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 $\theta$ deflects the beam by $2\theta$.

Figure 1.10 Geometry for lab 1.11.

P1.12 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.

(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 $1.5 \times 10^{11}$ m. To simplify the geometry, just assume that Earth move directly toward or away from Jupiter over the entire 40 orbits (see Fig. 1.11).
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.

Figure 1.11 fig:Roemer

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

P1.13 Suppose we have an electric field given by \( \mathbf{E} = (7x^2 y^3 \mathbf{x} + 2z^4 \mathbf{y}) \cos(\omega t) \)

(a) Use Gauss’s law (1.1) to find the charge density \( \rho(x, y, z, t) \).

(b) Use Faraday’s law (1.3) to find \( \frac{\partial \mathbf{B}(x, y, z, t)}{\partial t} \).

(c) Determine if \( \mathbf{E} \) is a solution to the vacuum wave equation, (1.50).

P1.14 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. \( \chi_2, \chi_3, \ldots \)) 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:

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

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 focus on sinusoidal solutions of Maxwell's equations, called plane waves, and their interaction with matter. Restricting our attention to plane wave solutions 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 \( E \) and \( B \). 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 electric field of a plane wave induces oscillating dipoles in a medium, and these oscillating dipoles feed back on the electric field to make the wave propagate at a different speed than it would in vacuum. We describe this effect using the index of refraction. Since materials respond differently to different frequencies of light, plane waves of different frequencies travel at different speeds in materials. Thus, when we have a waveform composed of many frequencies, it invariably changes its shape when it travels as its frequency components change relationship with one another. This phenomenon (called dispersion) is one of the primary reasons why physicists and engineers choose to work with sinusoidal waves.

When describing plane waves, 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. We will introduce complex electric field waves in section 2.2, and the index of refraction in section 2.3. When the electric field is represented using complex notation, the index of refraction 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. Complex notation will be used extensively in this chapter (and throughout the rest of the book), so we recommend that the student
2.2 Plane Wave Solutions to the Wave Equation

Consider the wave equation for an electric field waveform propagating in vacuum (1.50):

$$\nabla^2 E - \mu_0\varepsilon_0 \frac{\partial^2 E}{\partial t^2} = 0 \quad (2.1)$$

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

$$E(\mathbf{r}, t) = E_0 \cos \left( \mathbf{k} \cdot \mathbf{r} - \omega t + \phi \right) \quad (2.2)$$

Here $\phi$ represents an arbitrary (constant) phase term. The vector $\mathbf{k}$, called the wave vector, may be written as

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

where $k$ has units of inverse length, $\hat{\mathbf{u}}$ is a unit vector defining the direction of propagation, and $\lambda_{\text{vac}}$ 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

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

Table 2.1 The electromagnetic spectrum.

<table>
<thead>
<tr>
<th>AM Radio</th>
<th>Frequency $\nu = \omega / 2\pi$</th>
<th>Wavelength $\lambda_{\text{vac}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$10^9$ Hz</td>
<td>300 m</td>
</tr>
<tr>
<td>FM Radio</td>
<td>$10^8$ Hz</td>
<td>3 m</td>
</tr>
<tr>
<td>Radar</td>
<td>$10^{10}$ Hz</td>
<td>0.03 m</td>
</tr>
<tr>
<td>Microwave</td>
<td>$10^9 - 10^{12}$ Hz</td>
<td>$0.3 , \text{m} \cdot 3 \times 10^{-4}$ m</td>
</tr>
<tr>
<td>Infrared</td>
<td>$10^{12} - 4 \times 10^{14}$ Hz</td>
<td>$3 \times 10^{-4} - 7 \times 10^{-7}$ m</td>
</tr>
<tr>
<td>Light (red)</td>
<td>$4.6 \times 10^{14}$ Hz</td>
<td>$6.5 \times 10^{-7}$ m</td>
</tr>
<tr>
<td>Light (yellow)</td>
<td>$5.5 \times 10^{14}$ Hz</td>
<td>$5.5 \times 10^{-7}$ m</td>
</tr>
<tr>
<td>Light (blue)</td>
<td>$6.7 \times 10^{14}$ Hz</td>
<td>$4.5 \times 10^{-7}$ m</td>
</tr>
<tr>
<td>Ultraviolet</td>
<td>$10^{15} - 10^{17}$ Hz</td>
<td>$4 \times 10^{-7} - 3 \times 10^{-9}$ m</td>
</tr>
<tr>
<td>X-rays</td>
<td>$10^{17} - 10^{20}$ Hz</td>
<td>$3 \times 10^{-9} - 3 \times 10^{-12}$ m</td>
</tr>
<tr>
<td>Gamma rays</td>
<td>$10^{20} - 10^{23}$ Hz $\rightarrow$</td>
<td>$3 \times 10^{-12} - 3 \times 10^{-15}$ m $\rightarrow$</td>
</tr>
</tbody>
</table>

thoroughly study the review on complex notation found in section 0.3 at this point.

In sections 2.4 and 2.5 we consider a very successful physical model developed by Hendrik Lorentz for describing the index of refraction in both dielectrics and conductors. 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.
2.2 Plane Wave Solutions to the Wave Equation

Notice that $k$ and $\omega$ are not independent of each other, are related through the *vacuum dispersion relation*

$$k = \frac{\omega}{c} \quad \text{(vacuum)} \quad (2.5)$$

Typical values for $\lambda_{\text{vac}}$ 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

$$B(r, t) = B_0 \cos \left( k \cdot r - \omega t + \phi \right), \quad (2.6)$$

but it is important to note that $B_0$, $k$, $\omega$, and $\phi$ are not independently chosen in (2.6). In order to satisfy Faraday’s law (1.3), the arguments of the cosine in (2.2) and (2.6) must be identical. In addition, Faraday’s law requires (see P 1.1)

$$B_0 = \frac{k \times E_0}{\omega} \quad (2.7)$$

In vacuum, the electric and magnetic fields travel in phase. They are directed perpendicular to each other as defined by the cross product in (2.7). 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 (2.5).

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 2.12) where the electric field is sufficiently strong to cause electrons to oscillate with velocities near the speed of light. Therefore, the magnetic field can be ignored in most optics problems. 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 via Faraday’s law.
The depiction of the electric field (2.2) and the associated magnetic field (2.6) in Fig. 2.1 shows the fields drawn like transverse waves on a string. However, they are actually large planar sheets containing uniform fields (different fields in different planes) that move in the direction of $\mathbf{k}$. The name plane wave is given since the argument in (2.2) 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 and magnetic field moving in the $\mathbf{k}$ direction.

At this point, we rewrite our plane wave solution using complex number notation. Although this change in notation will not make the task at hand any easier (and may even appear to complicate things), we introduce it here in preparation for later sections, where it will save considerable labor. (For a review of complex notation, see section 0.3.)

Using complex notation we rewrite (2.2) as

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

(2.8)

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

$$\tilde{E}_0 \equiv E_0 e^{i\phi}$$

(2.9)

The next step we take is to become intentionally sloppy. Physicists throughout the world have conspired to avoid writing $\text{Re} \{ \}$ 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 $\phi$ inside of the field amplitude as in (2.8), 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. Putting this all together, our plane wave solution in complex notation is written simply as

$$E(r, t) = E_0 e^{i(k \cdot r - \omega t)}$$

(2.10)

It is possible to construct any electromagnetic disturbance from a linear superposition of such waves.

**Example 2.1**

Verify that the complex plane wave (2.10) is a solution to the wave equation (2.1).
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. $J_{\text{free}} = 0$). In this case, we expect $E$ and $P$ to be parallel to each other so $\nabla \cdot P = 0$ from (1.42). The general wave equation (1.49) for the electric field reduces in this case to

$$\nabla^2 E - \varepsilon_0 \mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P}{\partial t^2}$$  \hspace{1cm} (2.13)

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

$$E = E_0 e^{i(k \cdot r - \omega t)}$$

$$P = P_0 e^{i(k \cdot r - \omega t)}$$  \hspace{1cm} (2.14)

By writing this, we are making the (reasonable) assumption that if an electric field stimulates a medium at frequency $\omega$, then the polarization in the medium also oscillates at frequency $\omega$. This assumption is typically rather good except when extreme electric fields are used (see P 1.14). Recall that by our prior agreement, the complex amplitudes of $E_0$ and $P_0$ carry phase information. Thus, while $E$ and $P$ in (2.14) oscillate at the same frequency, they can be out of phase with each other. This phase mismatch is most pronounced for materials that absorb energy at the plane wave frequency.

Substitution of the trial solutions (2.14) into (2.13) yields

$$-k^2 E_0 e^{i(k \cdot r - \omega t)} + \varepsilon_0 \mu_0 \omega^2 E_0 e^{i(k \cdot r - \omega t)} = -\mu_0 \omega^2 P_0 e^{i(k \cdot r - \omega t)}$$  \hspace{1cm} (2.15)

At this point, we need to make an explicit connection between $E_0$ and $P_0$. 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:

\[ P_0(\omega) = \varepsilon_0 \chi(\omega) E_0(\omega) \]  

We have introduced a dimensionless proportionality factor \( \chi(\omega) \) called the susceptibility, which depends on the frequency of the field. We account for the possibility that \( E \) and \( P \) oscillate out of phase by allowing \( \chi(\omega) \) be a complex number in (2.16).

By inserting (2.16) into (2.15) and canceling the field terms, we obtain the dispersion relation in dielectrics:

\[ k^2 = \varepsilon_0 \mu_0 \left[ 1 + \chi(\omega) \right] \omega^2 \]  

or

\[ k = \frac{\omega}{c} \sqrt{1 + \chi(\omega)} \]  

where we have used \( c \equiv 1/\sqrt{\varepsilon_0 \mu_0} \). When absorption is small we can neglect the imaginary part of \( \chi(\omega) \). By direct comparison with vacuum case (2.5), we see that the speed of the phase fronts for sinusoidal wave is

\[ v = c/n(\omega) \]  

where

\[ n(\omega) = \sqrt{1 + \chi(\omega)} \]  

(negligible absorption)

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 index of refraction, and hence the wave speed \( v \), are a function of frequency. This reflects the fact that all materials have some frequency dependence in their response to light.

In cases where absorption plays a role, \( \chi(\omega) \) cannot be approximated as being real, and we must use the complex index of refraction, defined by

\[ N \equiv (n + i\kappa) = \sqrt{1 + \chi(\omega)} \]  

where \( n \) and \( \kappa \) are respectively the real and imaginary parts of the index. (Note that \( \kappa \) is not \( k \).) According to (2.18), the magnitude of the wave vector also becomes complex according to

\[ k = \frac{N\omega}{c} = \frac{(n + i\kappa)\omega}{c} \]  

The complex index \( N \) takes account of absorption as well as the usual oscillatory behavior of the wave. We see this by explicitly placing (2.22) into (2.14):

\[ E(\mathbf{r}, t) = E_0 e^{-\frac{\kappa}{c} \mathbf{u} \cdot \mathbf{r}} e^{i\left(\frac{\omega}{c} \mathbf{u} \cdot \mathbf{r} - \omega t\right)} \]  

\(^1\)Electrodynamics books often use the electric displacement \( D \equiv \varepsilon_0 E + P \). The permittivity \( \varepsilon \) encapsulates the constitutive relation that connects \( P \) with \( E \). In a linear medium we have \( \varepsilon \equiv \varepsilon_0 (1 + \chi) \), so that the index of refraction is given by \( n = \sqrt{\varepsilon/\varepsilon_0} \).
As before, here \( \hat{u} \) is a \textit{real} unit vector specifying the direction of \( \mathbf{k} \). Again, when looking at (2.23), by special agreement in advance, we should just think of the real part, namely

\[
\mathbf{E}(\mathbf{r}, t) = \tilde{\mathbf{E}}_0 e^{-\frac{\omega}{c} \hat{u} \cdot \mathbf{r}} \cos \left( \frac{n \omega}{c} \hat{u} \cdot \mathbf{r} - \omega t \right)
\]

(2.24)

where the phase \( \phi \) was formerly held in the complex vector \( \tilde{\mathbf{E}}_0 \) (where the tilde had been suppressed). Figure 2.2 shows a graph of (2.24). The imaginary part of the index \( \kappa \) causes the wave to decay as it travels. The real part of the index \( n \) is associated with the oscillations of the wave. While the amplitude of the oscillations decays as the wave moves along, the speed of the wave is still controlled by \( n(\omega) \) as in (2.19). Note that the complex index of refraction only makes sense when we use complex notation to represent the plane waves.

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

\[
\text{Re}\{\mathbf{k}\} \equiv \frac{2\pi}{\lambda} \hat{u},
\]

(2.25)

and

\[
\omega = \frac{2\pi c}{\lambda n},
\]

(2.26)

where

\[
\lambda \equiv \lambda_{\text{vac}} / n.
\]

(2.27)

While the frequency \( \omega \) is the same, whether in a material or in vacuum, the wavelength \( \lambda \) in the material is different as from the wavelength in vacuum, as indicated by (2.27).

### 2.4 The Lorentz Model of Dielectrics

To compute the index of refraction in either a dielectric or a conducting material, we require a model that describes 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.

Hendrik 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 and a damping force that dissipates energy and gives rise to absorption.

\[\text{For the sake of simplicity in writing (2.24) we assumed linearly polarized light. That is, all vector components of } \mathbf{E}_0 \text{ were assumed to have the same complex phase } \phi. \text{ The expression would be somewhat more complicated, for example, in the case of circularly polarized light (described in chapter 4).}\]
The Lorentz model determines the susceptibility $\chi(\omega)$ (the connection between the electric field $E_0$ and the polarization $P_0$) and hence the index of refraction. The model assumes 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

$$P = N q_e r_{\text{micro}}$$ \hspace{1cm} (2.28)

Recall that polarization has units of dipoles per volume. Each dipole has strength $q_e r_{\text{micro}}$, where $r_{\text{micro}}$ is a microscopic displacement of the electron from equilibrium.

At the time of Lorentz, atoms were thought to be clouds of positive charge wherein point-like electrons sat at rest unless stimulated by an applied electric field. In our modern quantum-mechanical viewpoint, $r_{\text{micro}}$ corresponds to an average displacement of the electronic cloud, which surrounds the nucleus (see Fig. 2.3). The displacement $r_{\text{micro}}$ of the electron charge in an individual atom depends on the local strength of the applied electric field $E$ at 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:

$$m_e \ddot{r}_{\text{micro}} = q_e E - m_e \gamma \dot{r}_{\text{micro}} - k_{\text{Hooke}} r_{\text{micro}}$$ \hspace{1cm} (2.29)

The electric field pulls on the electron with force $q_e E$. A dragging force $-m_e \gamma \dot{r}_{\text{micro}}$ 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, $-k_{\text{Hooke}} r_{\text{micro}}$ 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, so this term is essentially the familiar Hooke’s law. With some rearranging, (2.29) can be written as

$$\ddot{r}_{\text{micro}} + \gamma \dot{r}_{\text{micro}} + \omega_0^2 r_{\text{micro}} = \frac{q_e}{m_e} E$$ \hspace{1cm} (2.30)

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 (2.14) into (2.30) and obtain (2.31) into (2.30) and obtain

$$\ddot{r}_{\text{micro}} + \gamma \dot{r}_{\text{micro}} + \omega_0^2 r_{\text{micro}} = \frac{q_e}{m_e} E_0 e^{i(k r - \omega t)}$$ \hspace{1cm} (2.31)

Figure 2.3 A distorted electronic cloud becomes a dipole.

---

$^3$The electron also experiences a force due to the magnetic field of the light, $F = q_e v_{\text{micro}} \times B$, but this force is tiny for typical optical fields.
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 (2.31) is (see P 2.1)

\[
\mathbf{r}_{\text{micro}} = \left( \frac{q_e}{m_e} \right) \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \frac{1}{\omega_0^2 - i\omega \gamma - \omega^2}
\]

(2.32)

The electron position \( \mathbf{r}_{\text{micro}} \) oscillates (not surprisingly) with the same frequency \( \omega \) 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 \( \gamma \) (the imaginary part in the denominator) causes the two to be out of phase somewhat. The complex algebra in (2.32) accomplishes what would otherwise be cumbersome and require trigonometric manipulations.

We are now able to write the polarization in terms of the electric field. By substituting (2.32) into (2.28) and rearranging, we obtain

\[
\mathbf{P} = \varepsilon_0 \left( \frac{\omega_p^2}{\omega_0^2 - i\omega \gamma - \omega^2} \right) \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}
\]

(2.33)

where the plasma frequency \( \omega_p \) is

\[
\omega_p = \sqrt{\frac{N q_e^2}{\varepsilon_0 m_e}}
\]

(2.34)

A comparison of (2.33) with (2.16) in view of (2.14) reveals the (complex) susceptibility:

\[
\chi(\omega) = \frac{\omega_p^2}{\omega_0^2 - i\omega \gamma - \omega^2}
\]

(2.35)

The index of refraction is then found by substituting the susceptibility (2.35) into (2.21). The real and imaginary parts of the index are solved by equating separately the real and imaginary parts of (2.21), namely

\[
(n + i\kappa)^2 = 1 + \chi(\omega) = 1 + \frac{\omega_p^2}{\omega_0^2 - i\omega \gamma - \omega^2}
\]

(2.36)

A graph of \( n \) and \( \kappa \) is given in Fig. 2.4.

Most materials actually have more than one species of active electron, and different active electrons behave differently. The generalization of (2.36) in this case is

\[
(n + i\kappa)^2 = 1 + \chi(\omega) = 1 + \sum_j \frac{f_j \omega_{p j}^2}{\omega_0^2 - i\omega \gamma_j - \omega^2}
\]

(2.37)

where \( f_j \) is the aptly named oscillator strength for the \( j^{th} \) species of active electron. Each species also has its own plasma frequency \( \omega_{p j} \), natural frequency \( \omega_{0 j} \), and damping coefficient \( \gamma_j \).
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 Schrödinger 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 the oscillator strength, which before the development of quantum mechanics had to be inserted _ad hoc_ to make the model agree with experiments.

### 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 a conducting medium, electrons don’t experience a restoring force and are free to move outside of atoms, but they are still subject to a damping force due to collisions that removes energy and gives rise to absorption. Such collisions give rise to resistance in a conductor. In this model, we will ignore polarization (i.e. \( P = 0 \)), but take the current density \( J_{\text{free}} \) to be non-zero. The wave equation then becomes

\[
\nabla^2 E - \varepsilon_0 \mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial}{\partial t} J_{\text{free}} \quad (2.38)
\]

In a manner similar to (2.14), we assume sinusoidal solutions:

\[
E = E_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}
\]

\[
J_{\text{free}} = J_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \quad (2.39)
\]

We assume that the current is made up of individual electrons traveling with velocity \( \mathbf{v}_{\text{micro}} \):

\[
J_{\text{free}} = N q_e \mathbf{v}_{\text{micro}} \quad (2.40)
\]

Again, \( N \) is the number density of free electrons (in units of number per volume). Recall that current density \( J_{\text{free}} \) 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.

We have already solved Newton’s equation of motion (2.29) for the active electrons in a dielectric. We can consider the electrons in a conductor as the special case of this solution where the restoring “spring” constant \( k_{\text{hook}} \) is zero (and hence \( \omega_0 = 0 \)). The solution for \( \mathbf{r}_{\text{micro}} \) in a conductor is given by (2.32) with \( \omega_0 = 0 \):

\[
\mathbf{r}_{\text{micro}} = \left( \frac{q_e}{m_e} \right) \frac{E_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}}{-i \omega \gamma - \omega^2} \quad (2.41)
\]

We then find the electron velocity by taking a time derivative of (2.41):

\[
\mathbf{v}_{\text{micro}} = \left( \frac{q_e}{m_e} \right) \frac{E_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}}{\gamma - i \omega} \quad (2.42)
\]
With (2.41), we can find an expression for the current density (2.40) in terms of the electric field:

\[
J_{\text{free}} = \left( \frac{N q_e^2}{m_e} \right) \mathbf{E}_0 e^{i(k \cdot \mathbf{r} - \omega t)} \frac{1}{\gamma - i\omega}
\]  

Note that in the DC case (i.e. \( \omega = 0 \)), this expression reduces to Ohm’s law \( J = \sigma \mathbf{E} \), where \( \sigma = N q_e^2 / m_e \gamma \) is the conductivity. Although this formula relates the dragging term \( \gamma \) to the DC conductivity \( \sigma \), the connection matches poorly with experimental observations made for visible frequencies. This is because the collision rate actually varies somewhat with frequency.

We substitute (2.43) together with the electric field in (2.39) into the wave equation (2.38) and obtain the dispersion relation in this model:

\[
-k^2 \mathbf{E}_0 e^{i(k \cdot \mathbf{r} - \omega t)} + \frac{\omega^2}{c^2} \mathbf{E}_0 e^{i(k \cdot \mathbf{r} - \omega t)} = -i\omega \left( \frac{\mu_0 N q_e^2}{m_e} \right) \mathbf{E}_0 e^{i(k \cdot \mathbf{r} - \omega t)} \frac{1}{\gamma - i\omega}
\]  

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

\[
k^2 = \frac{\omega^2}{c^2} \left( 1 - \frac{\omega_p^2}{i\gamma \omega + \omega^2} \right)
\]  

(2.45)

Comparing this expression with (2.22), we find that the complex index of refraction for the conductor model is given by

\[
(n + i\kappa)^2 = 1 - \frac{\omega_p^2}{i\gamma \omega + \omega^2}
\]  

(2.46)

A graph of \( n \) and \( \kappa \) in the conductor model is given in Fig. 2.5.

The form of the complex refractive index for the conductor model is quite similar to the index found for the dielectric 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...
Chapter 2 Plane Waves and Refractive Index

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: (1.44) and (1.45). 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

$$\frac{\mathbf{B}}{\mu_0} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \frac{\mathbf{B}}{\mu_0}) + \varepsilon_0 \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \frac{\mathbf{B}}{\mu_0} \cdot \frac{\partial \mathbf{B}}{\partial t} = - \mathbf{E} \cdot \left( \mathbf{J}_{\text{free}} + \frac{\partial \mathbf{P}}{\partial t} \right)$$  (2.47)

The first two terms can be simplified using the vector identity $P \cdot 0.10$. The next two terms are the time derivatives of $\varepsilon_0 E^2/2$ and $B^2/2\mu_0$, respectively. The relation (2.47) then becomes

$$\nabla \cdot \left( \mathbf{E} \times \frac{\mathbf{B}}{\mu_0} \right) - \frac{\partial}{\partial t} \left( \frac{\varepsilon_0 E^2}{2} + \frac{B^2}{2\mu_0} \right) = - \mathbf{E} \cdot \left( \mathbf{J}_{\text{free}} + \frac{\partial \mathbf{P}}{\partial t} \right)$$  (2.48)

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:

$$\nabla \cdot \mathbf{S} + \frac{\partial u_{\text{field}}}{\partial t} = - \frac{\partial u_{\text{medium}}}{\partial t}$$  (2.49)

where

$$\mathbf{S} \equiv \mathbf{E} \times \frac{\mathbf{B}}{\mu_0}$$  (2.50)

$$u_{\text{field}} \equiv \frac{\varepsilon_0 E^2}{2} + \frac{B^2}{2\mu_0}$$  (2.51)

and

$$\frac{\partial u_{\text{medium}}}{\partial t} \equiv \mathbf{E} \cdot \left( \mathbf{J}_{\text{free}} + \frac{\partial \mathbf{P}}{\partial t} \right).$$  (2.52)

$\mathbf{S}$ is called the Poynting vector and has units of power per area, called irradiance. The quantity $u_{\text{field}}$ 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 (2.67) and (2.74).) The term $\partial u_{\text{medium}}/\partial t$ is the power per volume delivered to the medium. Equation (2.52) is reminiscent of the familiar circuit power law, $\text{Power} = \text{Voltage} \times \text{Current}$. 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. $\mathbf{S}$ describes
2.7 Irradiance of a Plane Wave

the flow of energy. To see this more clearly, consider Poynting’s theorem (2.49) integrated over a volume \( V \) (enclosed by surface \( S \)). If we also apply the divergence theorem (0.11) to the term involving \( \nabla \cdot \mathbf{S} \) we obtain

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

(2.53)

Notice that the volume integral over energy densities \( u_{\text{field}} \) and \( u_{\text{medium}} \) 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 (2.53) 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 \( \mathbf{S} \) 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 (2.10). The magnetic field that accompanies this electric field can be found from Maxwell’s equation (1.44), and it turns out to be

\[
\mathbf{B}(r, t) = \frac{k \times \mathbf{E}_0}{\omega} e^{i(k r - \omega t)}
\]  

(2.54)

When \( k \) is complex, \( \mathbf{B} \) is out of phase with \( \mathbf{E} \), and this occurs when absorption takes place. When there is no absorption, then \( k \) is real, and \( \mathbf{B} \) and \( \mathbf{E} \) carry the same complex phase.

Before computing the Poynting vector (2.50), 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 (0.23)). We could rewrite \( \mathbf{B} \) and \( \mathbf{E} \) like in (2.23), 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 (0.30)). The real field associated with (2.10) is

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

(2.55)

and the real field associated with (2.54) is

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

(2.56)

By writing (2.55) and (2.56), we have merely exercised our previous agreement that only the real parts of (2.36) and (2.54) are to be retained.
Now we are ready to calculate the Poynting vector. The algebra is a little messy in general, so we restrict to the case of an isotropic medium to simplify.

**Example 2.2**

Use (2.55) and (2.54) to calculate the Poynting vector (2.50) associated with the plane wave. Assume that an isotropic medium with \( V \cdot \mathbf{E}(r, t) = 0 \) so that \( \mathbf{u} \cdot \mathbf{E}_0 = 0 \)

**Solution:**

\[
\mathbf{S} = \mathbf{E} \times \frac{\mathbf{B}}{\mu_0} = \frac{1}{2} \left[ \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} + \mathbf{E}_0^* e^{-i(\mathbf{k}^* \cdot \mathbf{r} - \omega t)} \right] \times \frac{1}{2\mu_0} \left[ i\mathbf{k} \times \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} + i\mathbf{k}^* \times \mathbf{E}_0^* e^{-i(\mathbf{k}^* \cdot \mathbf{r} - \omega t)} \right]
\]

\[
= \frac{1}{4\mu_0} \left[ \frac{\mathbf{E}_0 \times [\mathbf{k} \times \mathbf{E}_0]}{\omega} e^{2i(\mathbf{k} \cdot \mathbf{r} - \omega t)} + \frac{\mathbf{E}_0^* \times [\mathbf{k}^* \times \mathbf{E}_0^*]}{\omega} e^{-2i(\mathbf{k}^* \cdot \mathbf{r} - \omega t)} + \frac{\mathbf{E}_0 \times [\mathbf{k} \times \mathbf{E}_0]}{\omega} e^{i(\mathbf{k} \cdot \mathbf{r})} - \frac{\mathbf{E}_0^* \times [\mathbf{k}^* \times \mathbf{E}_0^*]}{\omega} e^{-i(\mathbf{k}^* \cdot \mathbf{r})} \right]
\]

\[
= \frac{1}{4\mu_0} \left[ \frac{k}{\omega} \mathbf{E}_0 \times \left( \mathbf{u} \times \mathbf{E}_0 \right) e^{2i(\mathbf{k} \cdot \mathbf{r} - \omega t)} + \frac{k}{\omega} \mathbf{E}_0^* \times \left( \mathbf{u} \times \mathbf{E}_0^* \right) e^{-2i(\mathbf{k}^* \cdot \mathbf{r})} + C.C. \right]
\]

The letters “C.C.” stand for the complex conjugate of what precedes. The direction of \( \mathbf{k} \) is specified with the real unit vector \( \mathbf{u} \). We have also used (2.22) to rewrite \( i(\mathbf{k} - \mathbf{k}^*) \) as \(-2(\kappa \omega / c)\mathbf{u}\).

From the assumption that we have an isotropic medium (not a crystal) we have \( \mathbf{u} \cdot \mathbf{E}_0 = 0 \). We can use this fact together with the BAC-CAB rule P 0.4 to replace the above expression with

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

The final expression (2.58) shows that in an isotropic medium the flow of energy is in the direction of \( \mathbf{u} \) (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 (2.58) vanishes since it rapidly 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

\[
\langle \mathbf{S} \rangle_t = \frac{\mathbf{u}}{4\mu_0} \frac{k + k^*}{\omega} \left( \mathbf{E}_0 \cdot \mathbf{E}_0^* \right) e^{-2i\frac{\omega}{c} \mathbf{u} \cdot \mathbf{r}}
\]

\[
= \frac{\mathbf{u} \mu_0 c}{2} \left( \left| E_{0x} \right|^2 + \left| E_{0y} \right|^2 + \left| E_{0z} \right|^2 \right) e^{-2i\frac{\omega}{c} \mathbf{u} \cdot \mathbf{r}}
\]

We have used (2.22) to rewrite \( k + k^* \) as \( 2(\kappa \omega / c) \). We have also used (1.51) to rewrite \( 1/\mu_0 c \) as \( \varepsilon_0 c \).
The expression (2.59) is called irradiance (with the direction \( \hat{u} \) included). However, we often speak of the intensity of a field \( I \), which amounts to the same thing, but without regard for the direction \( \hat{u} \). 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

\[
I = \frac{n\epsilon_0 c}{2} E_0 \cdot E_0^* = \frac{n\epsilon_0 c}{2} \left( |E_{0x}|^2 + |E_{0y}|^2 + |E_{0z}|^2 \right)
\]  

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

**Appendix 2.A Energy Density of Electric Fields**

In this appendix we prove that the term \( \epsilon_0 E^2/2 \) in (2.51) corresponds to the energy density of an electric field. The electric potential \( \phi(r) \) (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

\[
E(r) = -\nabla \phi(r)
\]  

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 \( \rho(r) \)) located within the potential:

\[
U = \frac{1}{2} \int_V \phi(r) \rho(r) \, dV
\]  

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 \( \rho(r) \) into (2.62) gives

\[
U = \frac{\epsilon_0}{2} \int_V \phi(r) \nabla \cdot E(r) \, dV
\]  

Next, we use the vector identity in P 0.11 and get

\[
U = \frac{\epsilon_0}{2} \int_V \nabla \cdot [\phi(r) E(r)] \, dV - \frac{\epsilon_0}{2} \int_V E(r) \cdot \nabla \phi(r) \, dV
\]  

© 2004-2009 Peatross and Ware
An application of the Divergence theorem (0.11) on the first integral and a substitution of (2.61) into the second integral yields

\[
U = \frac{\varepsilon_0}{2} \oint_S \phi(\mathbf{r}) \mathbf{E}(\mathbf{r}) \cdot \hat{n} \, da + \frac{\varepsilon_0}{2} \int_V \mathbf{E}(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r}) \, dv \tag{2.65}
\]

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

\[
U = \int_V u_E(\mathbf{r}) \, dv \tag{2.66}
\]

where

\[
u_E(\mathbf{r}) \equiv \frac{\varepsilon_0 E^2}{2} \tag{2.67}
\]

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\( \times \)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

\[
\mathbf{B}(\mathbf{r}) = \nabla \times \mathbf{A}(\mathbf{r}) \tag{2.68}
\]

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:

\[
U = \frac{1}{2} \int_V \mathbf{J}(\mathbf{r}) \cdot \mathbf{A}(\mathbf{r}) \, dv \tag{2.69}
\]

As in (2.62), 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 (1.21) into (2.69), which yields

\[
U = \frac{1}{2\mu_0} \int_V [\nabla \times \mathbf{B}(\mathbf{r})] \cdot \mathbf{A}(\mathbf{r}) \, dv \tag{2.70}
\]
Next we employ the vector identity \( P 0.10 \) from which the previous expression becomes

\[
U = \frac{1}{2\mu_0} \int_V \mathbf{B}(\mathbf{r}) \cdot [\nabla \times \mathbf{A}(\mathbf{r})] \, d\mathbf{v} - \frac{1}{2\mu_0} \int_V \nabla \cdot [\mathbf{A}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})] \, d\mathbf{v}
\]  
\( (2.71) \)

Upon substituting (2.68) into the first equation and applying the Divergence theorem (0.11) on the second integral, this expression for total energy becomes

\[
U = \frac{1}{2\mu_0} \int_V \mathbf{B}(\mathbf{r}) \cdot \mathbf{B}(\mathbf{r}) \, d\mathbf{v} - \frac{1}{2\mu_0} \oint_S [\mathbf{A}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})] \cdot \mathbf{n} \, d\mathbf{a}
\]  
\( (2.72) \)

As was done in connection with (2.65), 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 \( d\mathbf{a} \sim R^2 \). The total energy (2.72) then reduces to

\[
U = \int_V u_B(\mathbf{r}) \, d\mathbf{v}
\]  
\( (2.73) \)

where

\[
u_B(\mathbf{r}) \equiv \frac{B^2}{2\mu_0}
\]  
\( (2.74) \)

is the energy density for a magnetic field.

### 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 \( \lambda_{\text{vac}} = 555 \text{ nm} \) and drops to near zero for wavelengths longer than \( \lambda_{\text{vac}} = 700 \text{ nm} \) or shorter than \( \lambda_{\text{vac}} = 400 \text{ nm} \). 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 \((1/683) \text{ W}\) of light with wavelength \( \lambda_{\text{vac}} = 555 \text{ nm} \), 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.
### Table 2.2 Radiometric quantities and units.

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

### Table 2.3 Photometric quantities and units.

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

Exercises for 2.3 Index of Refraction in Dielectrics

P2.1 Verify that (2.32) is a solution to (2.31).

P2.2 Derive the Sellmeier equation

\[ n^2 = 1 + \frac{A \lambda_0^2}{\lambda^2 - \lambda_0^2\text{vac}} \]

from (2.36) for a gas with negligible absorption (i.e. \( \gamma \approx 0 \), valid far from resonance \( \omega_0 \)), where \( \lambda_0\text{vac} \) corresponds to frequency \( \omega_0 \) 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 \( \omega_0 = 6 \times 10^{15} \text{ rad/sec} \) (in the UV), and damping \( \gamma = \omega_0/5 \) (quite broad). Assume \( E_0 \) is \( 10^4 \text{ V/m} \).

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 \( r_{\text{micm}} \) \( (2.32) \)

(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 \( \kappa \) 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 \( \lambda \).

(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 \( \kappa \) as a function of \( \omega \) frequency for a dielectric (i.e. obtain graphs such as the ones in Fig. ??(a)). Use these parameters to keep things
simple: \( \omega_p = 1, \omega_0 = 10, \) and \( \gamma = 1; \) plot your function from \( \omega = 0 \) to \( \omega = 20. \)

(b) Plot \( n \) and \( \kappa \) 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 \( \omega_p = 1 \) for all three resonances, and plot the results from \( \omega = 0 \) to \( \omega = 30. \) Comment on your plots.

**Exercises for 2.5 Conductor Model of Refractive Index and Absorption**

**P2.5**
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 \( \lambda_{\text{vac}} = 633 \text{ nm}. \) What is the speed of the wave crests in the silver (written as a number times \( c)\)? Are you surprised?

**P2.6**
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 \( \omega_p. \)

**P2.7**
Use the result from P 2.6.
(a) If the index of refraction of the ionosphere is \( n = 0.9 \) for an FM station at \( v = \omega/2\pi = 100 \text{ MHz}, \) 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.

**P2.8**
Use a computer graphing program to plot \( n \) and \( \kappa \) as a function of frequency for a conductor (obtain plots such as the ones in Fig. ??(b)). Use these parameters to keep things simple: \( \omega_p = 1 \) and \( \gamma = 0.02. \) Plot your function from \( \omega = 0.6 \) to \( \omega = 2. \)

**Exercises for 2.7 Irradiance of a Plane Wave**

**P2.9**
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 (2.59) directly from the real field (2.24). For simplicity, you may ignore absorption (i.e. \( \kappa \equiv 0. \))

HINT: The time-average of \( \cos^2 \left( k \cdot r - \omega t + \phi \right) \) is \( 1/2. \)
P2.10  (a) Find the intensity (in $\text{W/cm}^2$) produced by a short laser pulse (linearly polarized) with duration $\Delta t = 2.5 \times 10^{-14}$ s and energy $E = 100 \text{ mJ}$, focused in vacuum to a round spot with radius $r = 5 \mu\text{m}$.

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

HINT: The SI units of electric field are $\text{N/C} = \text{V/m}$.

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

P2.11  (a) What is the intensity (in $\text{W/cm}^2$) on the retina when looking directly at the sun? Assume that the eye's pupil has a radius $r_{\text{pupil}} = 1 \text{ mm}$. Take the Sun's irradiance at the earth's surface to be $1.4 \text{ kW/m}^2$, and neglect refractive index (i.e. set $n = 1$). HINT: The Earth-Sun distance is $d_o = 1.5 \times 10^8 \text{ km}$ and the pupil-retina distance is $d_i = 22 \text{ mm}$. The radius of the Sun $r_{\text{Sun}} = 7.0 \times 10^5 \text{ km}$ is de-magnified on the retina according to the ratio $d_i/d_o$.

(b) 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 $r_{\text{waist}} = 0.5 \text{ mm}$ positioned $d_o = 2 \text{ m}$ in front of the eye, and that the entire beam enters the pupil. Compare with part (a).

P2.12  Show that the magnetic field of an intense laser with $\lambda = 1 \mu\text{m}$ becomes important for a free electron oscillating in the field at intensities above $10^{18} \text{ W/cm}^2$. 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 $10^{18} \text{ W/cm}^2$. 

© 2004-2009 Peatross and Ware
Chapter 3

Reflection and Refraction

3.1 Introduction

As we know from everyday experience, when light arrives at an interface between materials it is partially reflected and partially transmitted. 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. We will derive expressions to quantify 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 $P$, 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 and 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 the refractive index. Each plane wave is thus characterized by a real wave vector $\mathbf{k}$. We will write each plane wave in the form $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \exp [i (\mathbf{k} \cdot \mathbf{r} - \omega t)]$, 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
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 $k_i$, $k_r$, and $k_t$ 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. $k_i$ specifies an incident plane wave making an angle $\theta_i$ with the normal to the interface. $k_r$ specifies a reflected plane wave making an angle $\theta_r$ with the interface normal. These two waves exist only to the left of the interface. $k_t$ specifies a transmitted plane wave making an angle $\theta_t$ 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 $k_i$, $k_r$, and $k_t$ (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 $k_i$, the electric field vector $E_i$ can be decomposed into arbitrary components as long as they are perpendicular to $k_i$. 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. $E_i^{(p)}$ denotes this component, represented by an arrow in the plane of the page. The remaining electric field vector component, denoted by $E_i^{(s)}$, is directed normal to the plane of incidence. The superscript

![Figure 3.1 Incident, reflected, and transmitted plane wave fields at a material interface.](image)
s stands for *senkrecht*, a German word meaning perpendicular. In Fig. 3.1, $E_i^{(s)}$ is represented by the tail of an arrow pointing into the page, or the $x$-direction, by our convention. The other fields $E_r$ and $E_t$ are similarly split into $s$ and $p$ components as indicated in Fig. 3.1. All field components are considered to be positive when they point in the direction of their respective arrows.\footnote{Many textbooks draw the arrow for $E_i^{(p)}$ in the direction opposite of ours. However, that choice leads to an awkward situation at normal incidence (i.e. $\theta_i = \theta_t = 0$) where the arrows for the incident and reflected fields are parallel for the $s$-component but anti parallel for the $p$-component.}

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

\[
\begin{align*}
\mathbf{k}_i &= k_i (\hat{y} \sin \theta_i + \hat{z} \cos \theta_i) \\
\mathbf{k}_r &= k_r (\hat{y} \sin \theta_r - \hat{z} \cos \theta_r) \\
\mathbf{k}_t &= k_t (\hat{y} \sin \theta_t + \hat{z} \cos \theta_t)
\end{align*}
\]

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 $\hat{x}$, $\hat{y}$, and $\hat{z}$:

\[
\begin{align*}
\mathbf{E}_i &= \left[ E_i^{(p)} (\hat{y} \cos \theta_i - \hat{z} \sin \theta_i) + \hat{x} E_i^{(s)} \right] e^{i[k_i(y \sin \theta_i - z \cos \theta_i) - \omega_i t]} \\
\mathbf{E}_r &= \left[ E_r^{(p)} (\hat{y} \cos \theta_r + \hat{z} \sin \theta_r) + \hat{x} E_r^{(s)} \right] e^{i[k_r(y \sin \theta_r + z \cos \theta_r) - \omega_r t]} \\
\mathbf{E}_t &= \left[ E_t^{(p)} (\hat{y} \cos \theta_t - \hat{z} \sin \theta_t) + \hat{x} E_t^{(s)} \right] e^{i[k_t(y \sin \theta_t + z \cos \theta_t) - \omega_t t]}
\end{align*}
\]

Each field has the form (2.8), and we have utilized the $k$-vectors (3.1) in the exponents of (3.2).

Now we are ready to connect the fields on one side of the boundary to the fields on the other side. This is done using boundary conditions. As explained in appendix 3.A, Maxwell’s equations require that the component of $\mathbf{E}$ that are parallel to the interface must be identical on either side of the interface. In our coordinate system, the $\hat{x}$ and $\hat{y}$ components are parallel to the interface, and $z = 0$ defines the interface. This means that at $z = 0$ the $\hat{x}$ and $\hat{y}$ components of the combined incident and reflected fields must match the corresponding components of the transmitted field:

\[
\begin{align*}
\left[ E_i^{(p)} (\hat{y} \cos \theta_i + \hat{x} E_i^{(s)}) \right] e^{i(k_i(y \sin \theta_i - \omega_i t))} + \left[ E_r^{(p)} (\hat{y} \cos \theta_r + \hat{x} E_r^{(s)}) \right] e^{i(k_r(y \sin \theta_r - \omega_r t))} \\
= \left[ E_t^{(p)} (\hat{y} \cos \theta_t + \hat{x} E_t^{(s)}) \right] e^{i(k_t(y \sin \theta_t - \omega_t t))}
\end{align*}
\]

Since this equation must hold for all conceivable values of $t$ and $y$, we are compelled to set all the phase factors in the complex exponentials equal to each other. The time portion of the phase factors requires the frequency of all waves to be the same:

\[
\omega_i = \omega_r = \omega_t \equiv \omega
\]

(We could have guessed that all frequencies would be the same; otherwise wave fronts would be annihilated or created at the interface.) Equating the spatial
terms in the exponents of (3.3) also requires
\[ k_1 \sin \theta_i = k_t \sin \theta_r = k_t \sin \theta_t \quad (3.5) \]

Now recall from (2.22) the relations \( k_i = k_t = n_i \omega / c \) and \( k_t = n_t \omega / c \). With these relations, (3.5) yields the law of reflection
\[ \theta_r = \theta_i \quad (3.6) \]
and Snell’s law
\[ n_1 \sin \theta_i = n_t \sin \theta_t \quad (3.7) \]
The three angles \( \theta_i, \theta_r, \) and \( \theta_t \) 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 \( \theta_i \) and \( \theta_t \) are different.

Because the exponents are all identical, (3.3) reduces to two relatively simple equations (one for each dimension, \( \hat{x} \) and \( \hat{y} \)):
\[ E_i^{(s)} + E_t^{(s)} = E_t^{(s)} \quad (3.8) \]
and
\[ \left( E_i^{(p)} + E_t^{(p)} \right) \cos \theta_1 = E_t^{(p)} \cos \theta_t \quad (3.9) \]

We have derived these equations from the boundary condition (3.52) on the parallel component of the electric field. This set of equations has four unknowns \((E_i^{(p)}, E_i^{(s)}, E_t^{(p)}, \) and \( E_t^{(s)}))\) assuming that the incident fields are given, so we require two further equations to solve the system. These are obtained using the separate boundary condition on the on the parallel component of magnetic fields given in (3.56) (also discussed in appendix 3.A).

From Maxwell’s equation (1.44), we have for a plane wave
\[ \mathbf{B} = \frac{\mathbf{k} \times \mathbf{E}}{\omega} = \frac{n}{c} \hat{u} \times \mathbf{E} \quad (3.10) \]
where \( \hat{u} = \mathbf{k} / k \) is a unit vector in the direction of \( \mathbf{k} \). We have also utilized (2.22).

This expression is useful to obtain expressions for \( \mathbf{B}_i, \mathbf{B}_r, \) and \( \mathbf{B}_t \) 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
\[ \mathbf{B}_i = \frac{n_i}{c} \left[ -\hat{x} E_i^{(p)} + E_i^{(s)} \left( -\hat{z} \sin \theta_i + \hat{y} \cos \theta_i \right) \right] e^{i[k_i(y \sin \theta_i + z \cos \theta_i) - \omega_i t]} \]
\[ \mathbf{B}_r = \frac{n_i}{c} \left[ -\hat{x} E_r^{(p)} + E_r^{(s)} \left( -\hat{z} \sin \theta_r - \hat{y} \cos \theta_r \right) \right] e^{i[k_i(y \sin \theta_r - z \cos \theta_i) - \omega_r t]} \]
\[ \mathbf{B}_t = \frac{n_t}{c} \left[ -\hat{x} E_t^{(p)} + E_t^{(s)} \left( -\hat{z} \sin \theta_t + \hat{y} \cos \theta_t \right) \right] e^{i[k_t(y \sin \theta_t + z \cos \theta_t) - \omega_t t]} \quad (3.11) \]

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 \( \hat{x} \) and \( \hat{y} \) 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 \( \theta_r = \theta_i \) and \( n_i = n_r \), the boundary condition gives

\[
\frac{n_i}{c} \left[ -\hat{x}E_i^{(p)} + E_i^{(s)} \hat{y} \cos \theta_i \right] + \frac{n_i}{c} \left[ \hat{x}E_t^{(p)} - E_t^{(s)} \hat{y} \cos \theta_i \right] = \frac{n_t}{c} \left[ -\hat{x}E_t^{(p)} + E_t^{(s)} \hat{y} \cos \theta_t \right]
\]  

(3.12)

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

\[
n_i \left( E_i^{(p)} - E_i^{(p)} \right) = n_t E_t^{(p)}
\]

(3.13)

and

\[
n_i \left( E_i^{(s)} - E_i^{(s)} \right) \cos \theta_i = n_t E_t^{(s)} \cos \theta_t
\]

(3.14)

These two equations (wherein the permeability \( \mu_0 \) 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 \( E_i \), these equations can be used to predict \( E_r \) and \( E_t \). 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 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).
Augustin Fresnel (1788–1829, French) Fresnel was a major proponent of the wave theory of light. He studied polarization, and invented the Fresnel prism for generating circularly polarized light. He also invented the Fresnel lens, originally used in light houses. Today fresnel lenses are used in many applications such as overhead projectors.

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)

\[ E_i^{(s)} + E_r^{(s)} = E_i^{(s)} \]  \hspace{1cm} [3.8]

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

\[ E_i^{(s)} - E_r^{(s)} = \frac{\sin \theta_1 \cos \theta_1}{\sin \theta_i \cos \theta_i} E_i^{(s)} \]  \hspace{1cm} [3.15]

If we add these two equations, we get

\[ 2E_i^{(s)} = \left[ 1 + \frac{\sin \theta_1 \cos \theta_1}{\sin \theta_i \cos \theta_i} \right] E_i^{(s)} \]  \hspace{1cm} [3.16]

and after dividing by \( E_i^{(s)} \) and doing a little algebra, we obtain

\[ \frac{E_t^{(s)}}{E_i^{(s)}} = \frac{2 \sin \theta_1 \cos \theta_1}{\sin \theta_i \cos \theta_1 + \sin \theta_i \cos \theta_1}. \]

To get the ratio of reflected to incident, we subtract (3.15) from (3.8) to obtain

\[ 2E_r^{(s)} = \left[ 1 - \frac{\sin \theta_1 \cos \theta_1}{\sin \theta_i \cos \theta_i} \right] E_i^{(s)} \]  \hspace{1cm} [3.17]

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

\[ \frac{E_t^{(s)}}{E_i^{(s)}} = \frac{\sin \theta_i \cos \theta_1 - \sin \theta_i \cos \theta_1}{\sin \theta_i \cos \theta_1 + \sin \theta_i \cos \theta_1} \]

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

\[ r_s = \frac{E_r^{(s)}}{E_i^{(s)}} = \frac{\sin \theta_1 \cos \theta_1 - \sin \theta_1 \cos \theta_1}{\sin \theta_i \cos \theta_1 + \sin \theta_1 \cos \theta_1} = \frac{\sin (\theta_1 - \theta_1)}{\sin (\theta_1 + \theta_1)} = \frac{n_t \cos \theta_1 - n_t \cos \theta_1}{n_t \cos \theta_1 + n_t \cos \theta_1} \]  \hspace{1cm} [3.18]

\[ t_s = \frac{E_t^{(s)}}{E_i^{(s)}} = \frac{2 \sin \theta_1 \cos \theta_1}{\sin \theta_i \cos \theta_1 + \sin \theta_1 \cos \theta_1} = \frac{2 \sin \theta_1 \cos \theta_1}{\sin (\theta_1 + \theta_1)} = \frac{2 n_t \cos \theta_1}{n_t \cos \theta_1 + n_t \cos \theta_1} \]  \hspace{1cm} [3.19]

\[ r_p = \frac{E_r^{(p)}}{E_i^{(p)}} = \frac{\cos \theta_1 \sin \theta_1 - \cos \theta_1 \sin \theta_1}{\cos \theta_i \sin \theta_1 + \cos \theta_i \sin \theta_1} = \frac{\tan (\theta_1 - \theta_1)}{\tan (\theta_1 + \theta_1)} = \frac{n_t \cos \theta_1 - n_t \cos \theta_1}{n_t \cos \theta_1 + n_t \cos \theta_1} \]  \hspace{1cm} [3.20]

\[ t_p = \frac{E_t^{(p)}}{E_i^{(p)}} = \frac{2 \cos \theta_1 \sin \theta_1}{\cos \theta_i \sin \theta_1 + \cos \theta_i \sin \theta_1} = \frac{2 \cos \theta_1 \sin \theta_1}{\sin (\theta_1 + \theta_1) \cos (\theta_1 - \theta_1)} = \frac{2 n_t \cos \theta_1}{n_t \cos \theta_1 + n_t \cos \theta_1} \]  \hspace{1cm} [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 $\pi$ 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 the electric field, we can write the fraction of the light reflected from the surface (called reflectance) in terms of the Fresnel coefficients as

$$R_s \equiv |r_s|^2 \quad \text{and} \quad R_p \equiv |r_p|^2$$

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

$$I_r^{(\text{total})} = I_r^{(s)} + I_r^{(p)} = R_s I_i^{(s)} + R_p I_i^{(p)}$$

where the incident intensity is given by (2.60):

$$I_i^{(\text{total})} = I_i^{(s)} + I_i^{(p)} = \frac{1}{2} n_i \epsilon_0 c \left| E_i^{(s)} \right|^2 + \left| E_i^{(p)} \right|^2$$

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

$$P_r^{(\text{total})} = P_r^{(s)} + P_r^{(p)} = R_s P_i^{(s)} + R_p P_i^{(p)}$$

Using this expression and requiring that energy be conserved (i.e. $P_i^{(\text{total})} = P_r^{(\text{total})} + P_t^{(\text{total})}$), we find the fraction of the power that transmits:

$$P_t^{(\text{total})} = (P_i^{(s)} + P_i^{(p)}) - (P_r^{(s)} + P_r^{(p)})$$

$$= (1 - R_s) P_i^{(s)} + (1 - R_p) P_i^{(p)}$$

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

$$T_s \equiv 1 - R_s \quad \text{and} \quad T_p \equiv 1 - R_p$$
Figure 3.3 The reflectance and transmittance plotted versus $\theta_i$ for the case of an air-glass interface with $n_i = 1$ and $n_t = 1.5$.

Figure 3.4 Light refracting into a surface.

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

You might be surprised at first to learn that

$$T_s \neq |t_s|^2 \quad \text{and} \quad T_p \neq |t_p|^2 \quad (3.28)$$

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

$$I_{t}^\text{total} = I_{t}^{(s)} + I_{t}^{(p)} = \frac{1}{2} n_t \varepsilon_0 c \left[ |E_{t}^{(s)}|^2 + |E_{t}^{(p)}|^2 \right] \quad (3.29)$$

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 $\cos \theta_t / \cos \theta_i$. 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:

$$T_s = \frac{n_t \cos \theta_t}{n_i \cos \theta_i} |t_s|^2 \quad \text{(valid when no total internal reflection)} \quad (3.30)$$

$$T_p = \frac{n_t \cos \theta_t}{n_i \cos \theta_i} |t_p|^2$$

Note that (3.30) is valid only if a real angle $\theta_t$ 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 $R_p + T_p = 1$, where $R_p$ is given by (3.22) and $T_p$ is given by (3.30).

**Solution:** From (3.20) we have

$$R_p = \frac{|\cos \theta_i \sin \theta_i - \cos \theta_i \sin \theta_i|^2}{\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i}$$

$$= \frac{\cos^2 \theta_i \sin^2 \theta_i - 2 \cos \theta_i \sin \theta_i \cos \theta_i \sin \theta_i + \cos^2 \theta_i \sin^2 \theta_i}{(\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i)^2}$$
From (3.21) and (3.30) we have

\[
T_p = \frac{n_i \cos \theta_i}{n_i \cos \theta_i} \left[ \frac{2 \cos \theta_i \sin \theta_i}{\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i} \right]^2
\]

\[
= \frac{\sin \theta_i \cos \theta_i}{\sin \theta_i \cos \theta_i} \left[ \frac{4 \cos^2 \theta_i \sin^2 \theta_i}{(\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i)^2} \right]
\]

\[
= \frac{4 \cos \theta_i \sin \theta_i \cos \theta_i}{(\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i)^2}
\]

Then

\[
R_p + T_p = \frac{\cos^2 \theta_i \sin^2 \theta_i + 2 \cos \theta_i \sin \theta_i \cos \theta_i \sin \theta_i + \cos^2 \theta_i \sin^2 \theta_i}{(\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i)^2}
\]

\[
= \frac{(\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i)^2}{(\cos \theta_i \sin \theta_i + \cos \theta_i \sin \theta_i)^2}
\]

\[
= 1
\]

### 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 \( \tan(\theta_i + \theta_t) \) in the denominator. Since the tangent “blows up” at \( \pi/2 \), the reflection coefficient goes to zero when

\[
\theta_i + \theta_t = \frac{\pi}{2} \quad \text{(requirement for zero \( p \)-polarized reflection)} \quad (3.31)
\]

By inspecting Fig. 3.1, we see that this condition occurs when the reflected and transmitted wave vectors, \( \mathbf{k}_r \) and \( \mathbf{k}_t \), are perpendicular to each other. If we insert (3.31) into Snell’s law (3.7), we can solve for the incident angle \( \theta_i \) that gives rise to this special circumstance:

\[
n_i \sin \theta_i = n_t \sin \left( \frac{\pi}{2} - \theta_i \right) = n_t \cos \theta_i
\]

\[
\theta_i = \tan^{-1} \left( \frac{n_t}{n_i} \right) \quad (3.32)
\]

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

\[
\theta_i = \tan^{-1} \left( \frac{n_t}{n_i} \right) \quad (3.33)
\]

We have replaced the specific \( \theta_i \) with \( \theta_s \) in honor of Sir David Brewster (1781-1868) who first discovered the phenomenon. The angle \( \theta_s \) is called Brewster’s angle. At Brewster’s angle, no \( p \)-polarized light reflects (see L 3.4). Physically, the \( p \)-polarized light cannot reflect because \( \mathbf{k}_r \) and \( \mathbf{k}_t \) 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:

\[ \theta_t = \sin^{-1} \left( \frac{n_i}{n_t} \sin \theta_i \right) \]  

(3.34)

The angle \( \theta_t \) is real only if the argument of the inverse sine is less than or equal to one. If \( n_i > n_t \), we can find a critical angle at which the argument begins to exceed one:

\[ \theta_c = \sin^{-1} \left( \frac{n_t}{n_i} \right) \]  

(3.35)

When \( \theta_i > \theta_c \), then there is total internal reflection and we can directly show that \( R_s = 1 \) and \( R_p = 1 \) (see P 3.7). To demonstrate this, one computes the Fresnel coefficients (3.18) and (3.20) while employing the following substitutions:

\[ \sin \theta_t = \frac{n_i}{n_t} \sin \theta_i \quad (\theta_i > \theta_c) \]  

(Snell’s law)  

(3.36)

and

\[ \cos \theta_t = i \sqrt{\frac{n_i^2}{n_t^2} \sin^2 \theta_i - 1} \quad (\theta_i > \theta_c) \]  

(3.37)

(see P 0.22).

In this case, \( \theta_t \) 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 \( \theta_t \); we need only the values for \( \sin \theta_t \) and \( \cos \theta_t \), as specified in (3.36) and (3.37). Even though \( \sin \theta_t \) is greater than one and \( \cos \theta_t \) 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

\[ r_s = \frac{n_i \cos \theta_i - i \sqrt{n_i^2 \sin^2 \theta_i - 1}}{n_i \cos \theta_i + i \sqrt{n_i^2 \sin^2 \theta_i - 1}} \quad (\theta_i > \theta_c) \]  

(3.38)

and

\[ r_p = -\frac{\cos \theta_i - i \frac{n_i}{n_t} \sqrt{n_i^2 \sin^2 \theta_i - 1}}{\cos \theta_i + i \frac{n_i}{n_t} \sqrt{n_i^2 \sin^2 \theta_i - 1}} \quad (\theta_i > \theta_c) \]  

(3.39)

These Fresnel coefficients can be manipulated (see P 3.7) into the forms

\[ r_s = \exp \left\{ -2i \tan^{-1} \left( \frac{n_i}{n_t \cos \theta_i} \sqrt{\frac{n_i^2}{n_t^2} \sin^2 \theta_i - 1} \right) \right\} \quad (\theta_i > \theta_c) \]  

(3.40)
and

\[ r_p = -\exp\left\{-2i \tan^{-1} \left[ \frac{n_i}{n_t} \sqrt{n_i^2 \sin^2 \theta_i - 1} \right] \right\} \quad (\theta_i > \theta_c) \quad (3.41) \]

Each coefficient has a different phase (note \(n_i/n_t\) vs. \(n_t/n_i\) 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. The transmittance is zero as dictated by (3.25). We emphasize that one should not employ (3.29) in the case of total internal reflection, as the imaginary \(\theta_t\) makes the geometric factor in this equation invalid.

Even with zero transmittance, 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 \(t_p \neq 0\). While this situation may seem like a contradiction at first, on closer examination we can see that it actually is an accurate description of what happens. The coefficients \(t_s\) and \(t_p\) characterize an 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):

\[
E_t = \left[ t_p E_t^{(p)} \right] \left[ \hat{y} \cos \theta_i - \hat{z} \sin \theta_i \right] e^{i (k_i n_i \sin \theta_i + z \cos \theta_i) - \omega t}

= \left[ t_p E_t^{(p)} \right] \left[ \hat{y} \sqrt{n_i^2 \sin^2 \theta_i - 1} \sin \theta_i \right] + \hat{x} t_s E_t^{(s)} e^{-k_z \sqrt{n_i^2 \sin^2 \theta_i - 1}} e^{i (k_i n_i \sin \theta_i - \omega t)}
\]

(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 3.8).

### 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 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—just replace the real index \(n\) with the complex index \(\mathcal{N}\). 
However, we do need to be a bit careful when applying them.

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

\[
\sin \theta_t = \frac{\sin \theta_i}{\mathcal{N}}
\]

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

\[
\cos \theta_t = \sqrt{1 - \sin^2 \theta_i} = \frac{1}{\mathcal{N}} \sqrt{\mathcal{N}^2 - \sin^2 \theta_i}
\]

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

\[
\begin{align*}
rs &= \frac{\cos \theta_i - \sqrt{\mathcal{N}^2 - \sin^2 \theta_i}}{\cos \theta_i + \sqrt{\mathcal{N}^2 - \sin^2 \theta_i}} \\
rd &= \frac{\sqrt{\mathcal{N}^2 - \sin^2 \theta_i} - \mathcal{N} \cos \theta_i}{\sqrt{\mathcal{N}^2 - \sin^2 \theta_i} + \mathcal{N} \cos \theta_i}
\end{align*}
\]

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

\[
r_s = |r_s| e^{i\phi_s}
\]
and

\[ r_p = |r_p| e^{i \phi_p} \]  

(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 it after a specific value for \( N \equiv n + i \kappa \) 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. \( \phi_p \) and \( \phi_s \) can be very different). This is true in general, even when the reflectivity is high (i.e. \( |r_s| \) and \( |r_p| \) 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 incident angle \( \theta_i \). Rather, the reflection of \( p \)-polarized light can go through a minimum at some angle \( \theta_i \), 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 \( \kappa \) and as a function of \( \theta_i \) 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 \( \mu_0 \) is the same on both sides). To derive the boundary conditions, we 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.

![Figure 3.7 Interface of two materials.](image-url)
Chapter 3 Reflection and Refraction

First we examine the implications of Faraday’s law (1.17):

$$\oint_{C} E \cdot d\ell = -\frac{\partial}{\partial t} \int_{S} B \cdot \hat{n} \, da$$

(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

$$\oint_{C} E \cdot d\ell = E_{1\parallel} \ell_{1} - E_{2\perp} \ell_{2} - E_{2\parallel} \ell_{2} + E_{1\perp} \ell_{1} = \left( E_{1\parallel} - E_{2\parallel} \right) d$$

(3.50)

Here, $E_{1\parallel}$ 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, $E_{2\parallel}$ 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 $\ell_{1}$ and $\ell_{2}$ go to zero. In this situation, the right-hand side of Faraday’s law goes to zero

$$\int_{S} B \cdot \hat{n} \, da \to 0$$

(3.51)

and we are left with

$$E_{1\parallel} = E_{2\parallel}$$

(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 using the integral form of Ampere’s law:

$$\oint_{C} B \cdot d\ell = \mu_{0} \int_{S} \left( J_{\text{free}} + \frac{\partial P}{\partial t} + \epsilon_{0} \frac{\partial E}{\partial t} \right) \cdot \hat{n} \, da$$

(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

$$\oint_{C} B \cdot d\ell = B_{1\parallel} \ell_{1} - B_{2\perp} \ell_{2} - B_{2\parallel} \ell_{2} + B_{1\perp} \ell_{1} = \left( B_{1\parallel} - B_{2\parallel} \right) d$$

(3.54)

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

---

2 This form can be obtained from (1.26) by integration over the surface $S$ in Fig. 3.7 and applying Stokes’ theorem (0.12) to the magnetic field term.
Again, we can continue to shrink the loop down until it has zero surface area by letting the lengths \( \ell_1 \) and \( \ell_2 \) go to zero. In this situation, the right-hand side of (3.53) goes to zero (not considering the possibility of surface currents):

\[
\int_S \left( J_{\text{free}} + \frac{\partial P}{\partial t} + \epsilon_0 \frac{\partial E}{\partial t} \right) \cdot \hat{n} \, d a \rightarrow 0 \quad (3.55)
\]

and we are left with

\[
B_{1\parallel} = B_{2\parallel} \quad (3.56)
\]

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

Exercises for 3.3 The Fresnel Coefficients

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

P3.2 Verify that each the alternative forms given in (3.18)–(3.21) are equivalent (given Snell’s law). Show that at normal incidence (i.e. \( \theta_i = \theta_t = 0 \)) the Fresnel coefficients reduce to

\[
\lim_{\theta_i \to 0} r_s = \lim_{\theta_i \to 0} r_p = -\frac{n_t - n_i}{n_t + n_i}
\]

and

\[
\lim_{\theta_i \to 0} t_s = \lim_{\theta_i \to 0} t_p = \frac{2n_i}{n_t + n_i}
\]

P3.3 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 \( n_i = 1 \) for air and \( n_t = 1.6 \) 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)

Exercises for 3.4 Reflectance and Transmittance

L3.4 (a) In the laboratory, measure the reflectance for both \( s \) and \( p \) polarized light from a flat glass surface at about ten points. 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 3.4.](image-url)
Exercises 75

(b) 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 $\theta_i$). Take the index of refraction for glass to be $n_t = 1.54$ and the index for air to be one. Plot this theoretical calculation as a smooth line on a graph. Plot your experimental data from (a) as points on this same graph (not points connected by lines).

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).

Exercises for 3.5 Brewster's Angle

P3.6 Find Brewster's angle for glass $n = 1.5$.

Exercises for 3.6 Total Internal Reflection

P3.7 Derive (3.40) and (3.41) and show that $R_s = 1$ and $R_p = 1$.

HINT:

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

where $a$ is positive and real and $b$ is real.

P3.8 Compute $t_s$ and $t_p$ in the case of total internal reflection.

P3.9 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 $\theta_i$). 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.

P3.10 Light ($\lambda_{\text{vac}} = 500$ nm) reflects internally from a glass surface ($n = 1.5$) surrounded by air. The incident angle is $\theta_i = 45^\circ$. 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?

Exercises for 3.7 Reflection from Metallic or other Absorptive Surfaces

P3.11 Using a computer graphing program that understands complex numbers (e.g. Matlab), plot $|r_s|$, $|r_p|$ versus $\theta_i$ for silver ($n = 0.2$ and $\kappa = 3.4$). Make a separate plot of the phases $\phi_s$ and $\phi_p$ 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.
P3.12 Find Brewster’s angle for silver \((n = 0.2 \text{ and } \kappa = 3.4)\) by calculating \(R_p\) and finding its minimum. You will want to use a computer program to do this (Matlab, Maple, Mathematica, etc.).

P3.13 The complex index for silver is given by \(n = 0.2 \text{ and } \kappa = 3.4\). Find \(r_s\) and \(r_p\) when \(\theta_1 = 80^\circ\) 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.)

\(\text{Figure 3.9 Geometry for P 3.13}\)
Chapter 4

Polarization

4.1 Introduction

When the direction of the electric field of light oscillates in a regular, predictable fashion, we say that the light is polarized. In this chapter, we develop a formalism for describing polarized light and for describing the effect of devices that modify polarization. First we introduce the vocabulary used to describe the various types of polarization that plane waves exhibit. Then in section 4.3 we introduce a convenient way for keeping track of polarization using a two-dimensional Jones vector, and in the next section we show how this formalism is used to describe a general form of polarization referred to as elliptically polarized light.

In section 4.5 we discuss a method for describing devices that can change polarization, and show how their effect on a light field can be represented by $2 \times 2$ Jones matrices operating on the polarization vector. In the next section we derive a general Jones matrix for a linear polarizer oriented at an arbitrary angle with respect to the coordinate system. We then apply this analysis to describe wave plates, which are devices that introduce a relative phase delay of one field component with respect to the other. A wave plate can be used used to convert, for example, linearly polarized light into circularly polarized light.

Beginning in section 4.8, 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 $2 \times 2$ matrix formulation for handling polarization. As we saw previously, 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). In section 4.9 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, in most natural sources of light (e.g. sunlight or the light from an incandescent lamp) the direction of the electric field varies rapidly and randomly.
Such sources are commonly referred to as *unpolarized*. It is common to have a mixture of unpolarized and polarized light, called *partially polarized* light. The Jones vector formalism used in this chapter is inappropriate for describing the unpolarized portions of the light. In appendix 4.A we describe a more general formalism for dealing with light having an arbitrary degree of polarization.

### 4.2 Linear, Circular, and Elliptical Polarization

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

$$ E(\mathbf{r}, t) = E_0 e^{i(k \mathbf{r} - \omega t)} \quad (4.1) $$

The wave vector $k$ 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 (2.22)-(2.27)). In an isotropic medium we know that $k$ and $E_0$ are perpendicular, but even after the direction of $k$ is specified, we are still free to have $E_0$ point anywhere in two dimensions perpendicular to $k$. If we orient our coordinate system with the $z$-axis in the direction of $k$, we can write (4.1) as

$$ E(z, t) = \left( E_x \hat{x} + E_y \hat{y} \right) e^{i(kz - \omega t)} \quad (4.2) $$

As always, only the real part of (4.2) is physically relevant. The complex amplitudes of $E_x$ and $E_y$ keep track of the phase of the oscillating field components. In general the complex phases of $E_x$ and $E_y$ can differ, so that the wave in one of the dimensions lags or leads the wave in the other dimension.

The relationship between $E_x$ and $E_y$ describes the polarization of the light. For example, if the $y$-component of the field $E_y$ 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 $E_x$ and $E_{0y}$ have the same complex phase (or differ by an integer times $\pi$). For our purposes, we will take the $x$-dimension to be horizontal and the $y$-dimension to be vertical unless otherwise noted.

As an example, suppose $E_y = i E_x$, where $E_x$ 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

$$ E(z, t) = \text{Re} \left[ E_x e^{i(kz - \omega t)} \right] \hat{x} + \text{Re} \left[ e^{i \pi/2} E_x e^{i(kz - \omega t)} \right] \hat{y} $$

$$ = E_x \cos(kz - \omega t) \hat{x} + E_x \cos(kz - \omega t + \pi/2) \hat{y} \quad \text{(left circular)} \quad (4.3) $$

$$ = E_x \left[ \cos(kz - \omega t) \hat{x} - \sin(kz - \omega t) \hat{y} \right] $$

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 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
4.3 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

\[ E(z, t) = \left( |E_x| e^{i\phi_x} \hat{x} + |E_y| e^{i\phi_y} \hat{y} \right) e^{i(kz-\omega t)} \quad (4.4) \]

and then factor (4.4) as follows:

\[ E(z, t) = E_{\text{eff}} \left( A \hat{x} + B e^{i\delta} \hat{y} \right) e^{i(kz-\omega t)} \quad (4.5) \]

where

\[ E_{\text{eff}} \equiv \sqrt{|E_x|^2 + |E_y|^2} e^{i\phi_x} \quad (4.6) \]

\[ A \equiv \frac{|E_x|}{\sqrt{|E_x|^2 + |E_y|^2}} \quad (4.7) \]

\[ B \equiv \frac{|E_y|}{\sqrt{|E_x|^2 + |E_y|^2}} \quad (4.8) \]

\[ \delta \equiv \phi_y - \phi_x \quad (4.9) \]

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 \( E_x \) happens to be zero, then its phase \( e^{i\phi_x} \) is indeterminant. In this case we let \( E_{\text{eff}} = |E_y| e^{i\phi_y} \), \( B = 1 \), and \( \delta = 0 \). (If \( E_y \) is zero, then \( e^{i\delta} \) 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 \( E_{\text{eff}} \) 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 (2.60) we have

\[ I = \langle S \rangle_t = \frac{1}{2} nce_a E \cdot E^* = \frac{1}{2} nce_a |E_{\text{eff}}|^2 \quad (4.10) \]

The phase of \( E_{\text{eff}} \) 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 \( A \hat{x} + B e^{i\delta} \hat{y} \), 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 \((A \hat{x} + B e^{i\delta} \hat{y}) \cdot (A \hat{x} + B e^{i\delta} \hat{y})^* = 1 \) (the asterisk represents the complex conjugate). When writing a Jones vector we dispense with the \( \hat{x} \) and \( \hat{y} \) notation and organize the components into a column vector (for later use in matrix algebra) as follows:

\[ \begin{bmatrix} A \\ B e^{i\delta} \end{bmatrix} \quad (4.11) \]

This vector can describe the polarization state of any plane wave field. Table 4.1 lists some Jones vectors representing various polarization states.

<table>
<thead>
<tr>
<th>Vector</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \begin{bmatrix} 1 \ 0 \end{bmatrix} ]</td>
<td>linearly polarized along the ( x ) dimension</td>
</tr>
<tr>
<td>[ \begin{bmatrix} 0 \ 1 \end{bmatrix} ]</td>
<td>linearly polarized along the ( y ) dimension</td>
</tr>
<tr>
<td>[ \begin{bmatrix} \cos \alpha \ \sin \alpha \end{bmatrix} ]</td>
<td>linearly polarized at an angle ( \alpha ) from the ( x )-axis</td>
</tr>
<tr>
<td>[ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \ -i \end{bmatrix} ]</td>
<td>right circularly polarized</td>
</tr>
<tr>
<td>[ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \ i \end{bmatrix} ]</td>
<td>left circularly polarized</td>
</tr>
</tbody>
</table>

Table 4.1 Jones Vectors for various polarization states
4.4 Elliptically Polarized Light

In general, the Jones vector (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

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

(4.12)

with respect to the $x$-axis (see P 4.8). 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 $\delta$. 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 $\alpha$ and when it spirals through $\alpha \pm \pi/2$. The strength of the electric field at $\alpha$ is given by (see P 4.8)

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

(4.13)

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

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

(4.14)

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

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

(4.15)

Figure 4.2 The electric field of elliptically polarized light traces an ellipse in the plane perpendicular to its propagation direction. The two plots are for different values of $A$, $B$, and $\delta$. The angle $\alpha$ can describe the major axis (left figure) or the minor axis (right figure), depending on the values of these parameters.
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:

\[ e = \frac{E_{\text{min}}}{E_{\text{max}}} \]  

(4.16)

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):

\[ 0 < \delta < \pi \quad \text{(left-handed helicity)} \]  

(4.17)

\[ \pi < \delta < 2\pi \quad \text{(right-handed helicity)} \]  

(4.18)

### 4.5 Linear Polarizers and 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 \( J_{\text{free}} \) 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 \( 2 \times 2 \) 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. 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 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
4.5 Linear Polarizers and Jones Matrices

Figure 4.3 Light transmitting through a Polaroid sheet. The conducting polymer chains run vertically in this drawing, and light polarized along the chains is absorbed. Light polarized perpendicular to the polymer chains passes through the polarizer.

represented with the following Jones matrices:

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\] (polarizer with transmission along x-axis) \hspace{1cm} (4.19)

\[
\begin{bmatrix}
0 & 0 \\
0 & 1
\end{bmatrix}
\] (polarizer with transmission along y-axis) \hspace{1cm} (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.

Example 4.1

Use the Jones matrix (4.19) to calculate the effect of a horizontal polarizer on light that is initially horizontally polarized, vertically polarized, and arbitrarily polarized.

Solution: First we consider a horizontally polarized plane wave traversing a polarizer with its transmission axis oriented also horizontally (x-dimension):

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix} = \begin{bmatrix}
1 \\
0
\end{bmatrix}
\] (horizontal polarizer on horizontally polarized field) \hspace{1cm} (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:

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 \\
1
\end{bmatrix} = \begin{bmatrix}
0 \\
0
\end{bmatrix}
\] (horizontal polarizer on vertical linear polarization) \hspace{1cm} (4.22)
As expected, the polarizer extinguishes the light. Finally, when a horizontally oriented polarizer operates on light with an arbitrary Jones vector (4.11), we have
\[
\begin{bmatrix}
1 & 0 \\
0 & 0 
\end{bmatrix}
\begin{bmatrix}
A \\
Re^{i\delta}
\end{bmatrix} =
\begin{bmatrix}
A \\
0
\end{bmatrix}
\] (horizontal polarizer on arbitrary polarization) \hspace{1cm} (4.23)

Only the horizontal component of polarization is transmitted through the polarizer.

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. In the next few sections we will derive Jones matrices for a number of optical elements that can modify polarization: polarizers at arbitrary angle, wave plates at arbitrary angle, and reflection or transmissions at an interface. Table 4.2 shows Jones matrices for each of these devices. Before deriving these specific Jones matrices, however, we take a moment to more fully appreciate why the Jones matrix formulation is useful.

The real power of the formalism becomes clear as we consider situations where light encounters multiple polarization elements in sequence. In these situations, we use a product of Jones matrices to represent the effect of the compound systems. We can represent this situation by
\[
\begin{bmatrix}
A' \\
B'e^{i\delta'}
\end{bmatrix} = \mathbf{J}_{\text{system}} \begin{bmatrix}
A \\
B'e^{i\delta}
\end{bmatrix}
\] \hspace{1cm} (4.24)

where the unprimed Jones vector represents light going into the system and the primed Jones vector represents light emerging from the system. The matrix \( \mathbf{J}_{\text{system}} \) is a Jones matrix formed by the series polarization devices. If there are \( N \) devices in the system, the compound matrix is calculated as
\[
\mathbf{J}_{\text{system}} = \mathbf{J}_N \mathbf{J}_{N-1} \cdots \mathbf{J}_2 \mathbf{J}_1 \] \hspace{1cm} (4.25)

where \( \mathbf{J}_n \) is the matrix for the \( n^{\text{th}} \) optical element encountered in the system. Notice that the matrices operate on the Jones vector in the order that the light encounters the devices. Therefore, the matrix for the first device (\( \mathbf{J}_1 \)) is written on the right, and so on until the last device encountered, which is written on the left, farthest from the Jones vector.

When part of the light is absorbed by passing through one or more polarizers in a system, the Jones vector of the exiting light is no longer normalized to magnitude one. Since the components of a Jones vector represent the electric field, we find the factor by which the intensity of the light decreases by dotted the vector with its complex conjugate. In accordance with (4.10), the intensity of the exiting light is
\[
I = \frac{1}{2} n c e_0 |E_{\text{eff}}|^2 \left( \left| A' \hat{x} + B' e^{i\delta'} \hat{y} \right| \cdot \left| A' \hat{x} + B' e^{i\delta'} \hat{y} \right|^* \right)
\]
\[
= \frac{1}{2} n c e_0 |E_{\text{eff}}|^2 \left( |A'|^2 + |B'|^2 \right)
\] \hspace{1cm} (4.26)
4.6 Jones Matrix for Polarizers at Arbitrary Angles

In this section we derive a Jones matrix to describe a plane wave with arbitrary polarization passing through a polarizer with its transmission axis aligned at an arbitrary angle $\theta$ with the $x$-axis. We derive this Jones matrix in a general context so that we can take advantage of present work when we discuss wave plates. To help keep things on a more conceptual level, we revert back to using electric field components directly. We will make the connection with Jones calculus at a later point. The electric field of our plane wave is

$$E(z, t) = (E_x \hat{x} + E_y \hat{y}) e^{i(kz - \omega t)} \quad (4.27)$$

The transmission axis of the polarizer is specified by the unit vector $\hat{e}_1$ and the

<table>
<thead>
<tr>
<th>Optical Element</th>
<th>Jones Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear polarizer, transmission axis oriented at $\theta$ with respect to the $x$-axis</td>
<td>$\begin{bmatrix} \cos^2 \theta &amp; \sin \theta \cos \theta \ \sin \theta \cos \theta &amp; \sin^2 \theta \end{bmatrix}$</td>
</tr>
<tr>
<td>Half wave plate, fast axis oriented at $\theta$ with respect to $x$-axis</td>
<td>$\begin{bmatrix} \cos 2\theta &amp; \sin 2\theta \ \sin 2\theta &amp; -\cos 2\theta \end{bmatrix}$</td>
</tr>
<tr>
<td>Quarter wave plate, fast axis oriented at $\theta$ with respect to $x$-axis</td>
<td>$\begin{bmatrix} \cos^2 \theta + i\sin^2 \theta &amp; \sin \theta \cos \theta - i\sin \theta \cos \theta \ \sin \theta \cos \theta - i\sin \theta \cos \theta &amp; \sin^2 \theta + i\cos^2 \theta \end{bmatrix}$</td>
</tr>
<tr>
<td>Right circular polarizer</td>
<td>$\frac{1}{2} \begin{bmatrix} 1 &amp; i \ -i &amp; 1 \end{bmatrix}$</td>
</tr>
<tr>
<td>Left circular polarizer</td>
<td>$\frac{1}{2} \begin{bmatrix} 1 &amp; -i \ i &amp; 1 \end{bmatrix}$</td>
</tr>
<tr>
<td>Reflection from an interface</td>
<td>$\frac{1}{2} \begin{bmatrix} -r_p &amp; 0 \ 0 &amp; r_s \end{bmatrix}$</td>
</tr>
<tr>
<td>Transmission through an interface</td>
<td>$\frac{1}{2} \begin{bmatrix} t_p &amp; 0 \ 0 &amp; t_s \end{bmatrix}$</td>
</tr>
</tbody>
</table>

Table 4.2 Summary of Jones Matrices.

Notice that the intensity is attenuated by the factor $|A'|^2 + |B'|^2$ after propagating through the system. Recall that $E_{\text{eff}}$ 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 $\delta'$ 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 (4.26) ensure that we get the correct value for intensity.
absorption axis of the polarizer be specified by \( \hat{e}_2 \) (orthogonal to the transmission axis). The vector \( \hat{e}_1 \) is oriented at an angle \( \theta \) from the \( x \)-axis as depicted in Fig. 4.4. We need to write the electric field components in terms of the new basis specified by the unit vectors \( \hat{e}_1 \) and \( \hat{e}_2 \) as shown in Fig. 4.5. The \( x \)-\( y \) unit vectors are connected to the new coordinate system via (see Fig. 4.4(b)):

\[
\hat{x} = \cos \theta \hat{e}_1 - \sin \theta \hat{e}_2 \\
\hat{y} = \sin \theta \hat{e}_1 + \cos \theta \hat{e}_2
\]  

(4.28)

By direct substitution of (4.28) into (4.27), the electric field can be written as

\[
E(z, t) = (E_1 \hat{e}_1 + E_2 \hat{e}_2) e^{i(kz-\omega t)}
\]  

(4.29)

where

\[
E_1 \equiv E_x \cos \theta + E_y \sin \theta \\
E_2 \equiv -E_x \sin \theta + E_y \cos \theta
\]  

(4.30)

Figure 4.5 Electric field components written in either the \( \hat{x}-\hat{y} \) basis or the \( \hat{e}_1-\hat{e}_2 \) basis.
Now we introduce the effect of the polarizer on the field: $E_1$ is transmitted unaffected, while $E_2$ is extinguished. To account for the effect of the device, we multiply $E_2$ by a parameter $\xi$. In the case of the polarizer, $\xi$ is simply zero, but when we consider wave plates we can have other values for $\xi$. After traversing the polarizer, the field becomes

$$E_{\text{after}}(z, t) = (E_1 \hat{e}_1 + \xi E_2 \hat{e}_2) e^{i(kz-\omega t)}$$ (4.31)

We now have the field after the polarizer, but it would be nice to rewrite it in terms of the original $x$–$y$ basis. By inverting (4.28), or by inspection of Fig. 4.4, if preferred, we see that

$$\hat{e}_1 = \cos \theta \hat{x} + \sin \theta \hat{y}$$
$$\hat{e}_2 = -\sin \theta \hat{x} + \cos \theta \hat{y}$$ (4.32)

Substitution of these relationships into (4.31) together with the definitions (4.30) for $E_1$ and $E_2$ yields

$$E_{\text{after}}(z, t) = \left[ \left( E_x \cos \theta + E_y \sin \theta \right) \left( \cos \theta \hat{x} + \sin \theta \hat{y} \right) + \xi \left( -E_x \sin \theta + E_y \cos \theta \right) \left( -\sin \theta \hat{x} + \cos \theta \hat{y} \right) \right] e^{i(kz-\omega t)}$$

$$= \left[ E_x (\cos^2 \theta + \xi \sin^2 \theta) + E_y (\sin \theta \cos \theta - \xi \sin \theta \cos \theta) \right] \hat{x} e^{i(kz-\omega t)}$$
$$+ \left[ E_x (\sin \theta \cos \theta - \xi \sin \theta \cos \theta) + E_y (\sin^2 \theta + \xi \cos^2 \theta) \right] \hat{y} e^{i(kz-\omega t)}$$ (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 (4.27)).

To get to the Jones matrix for the polarizer, we note that (4.33) is a linear mixture of $E_x$ and $E_y$ which can be represented with matrix algebra. If we represent the electric field 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

$$E_{\text{after}}(z, t) = \left[ \begin{array}{cc} \cos^2 \theta + \xi \sin^2 \theta & \sin \theta \cos \theta - \xi \sin \theta \cos \theta \\ \sin \theta \cos \theta - \xi \sin \theta \cos \theta & \sin^2 \theta + \xi \cos^2 \theta \end{array} \right] \left[ \begin{array}{c} E_x \\ E_y \end{array} \right] e^{i(kz-\omega t)}$$ (4.34)

The matrix here is a proper Jones matrix, but the vector it operates on is not a properly normalized Jones vector. However, we can make it into a proper Jones vector by simply by factoring as specified in (4.5). We can now write down the Jones matrix for a polarizer by simply inserting $\xi = 0$ into the matrix:

$$\left[ \begin{array}{cc} \cos^2 \theta & \sin \theta \cos \theta \\ \sin \theta \cos \theta & \sin^2 \theta \end{array} \right] \quad \text{(polarizer with transmission axis at angle $\theta$)}$$ (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).
4.7 Jones Matrices for Wave Plates

Another device that influencing polarization is called a wave plate (or retarder). Wave plates are usually made from a anisotropic 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 90° apart in the plane of the window. If the light is polarized along the fast axis, it experiences an index of refraction $n_{\text{fast}}$. This index is less than an index $n_{\text{slow}}$ 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 fast component gets ahead, and this introduces a relative phase between the two polarization components. The wave vectors associated with the two electric field components within the wave plate are given by

$$k_{\text{slow}} = \frac{2\pi n_{\text{slow}}}{\lambda_{\text{vac}}} \quad \text{and} \quad k_{\text{fast}} = \frac{2\pi n_{\text{fast}}}{\lambda_{\text{vac}}} \quad (4.36)$$

As light passes through a wave plate of thickness $d$, the phase difference that accumulates between the fast and the slow polarization components is

$$k_{\text{slow}}d - k_{\text{fast}}d = \frac{2\pi d}{\lambda_{\text{vac}}} (n_{\text{slow}} - n_{\text{fast}}) \quad (4.37)$$

By adjusting the thickness of the wave plate, we can introduce any desired phase difference between the two components.

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

$$k_{\text{slow}}d - k_{\text{fast}}d = \frac{\pi}{2} + 2\pi m \quad \text{(quarter-wave plate)} \quad (4.38)$$

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

$$k_{\text{slow}}d - k_{\text{fast}}d = \pi + 2\pi m \quad \text{(half-wave plate)} \quad (4.39)$$

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 $\hat{e}_1$ correspond to the fast axis, and let $\hat{e}_2$ correspond to the slow axis. We proceed as before. However, instead of setting $\xi$ equal to zero in (4.34), we must choose values for...
\[ \xi \] appropriate for each wave plate. Since nothing is absorbed, \( \xi \) should have a magnitude equal to one. The important feature is the phase of \( \xi \). As seen in (4.37), the field component along the slow axis accumulates excess phase relative to the component along the fast axis, and we let \( \xi \) account for this. In the case of the quarter-wave plate, the appropriate factor from (4.38) is

\[ \xi = e^{i\pi/2} = i \quad \text{(quarter-wave plate)} \] (4.40)

This describes a relative phase delay for the light emerging with polarization along the slow axis. Substituting (4.40) into (4.33) yields the Jones matrix for a quarter wave plate:

\[
\begin{bmatrix}
\cos^2 \theta + i \sin^2 \theta & \sin \theta \cos \theta - i \sin \theta \cos \theta \\
\sin \theta \cos \theta - i \sin \theta \cos \theta & \sin^2 \theta + i \cos^2 \theta
\end{bmatrix}
\quad \text{(quarter-wave plate)}
\] (4.41)

For the half-wave plate, the appropriate relative phase delay for the slow axis is

\[ \xi = e^{i\pi} = -1 \quad \text{(half-wave plate)} \] (4.42)

and the Jones matrix becomes:

\[
\begin{bmatrix}
\cos^2 \theta - \sin^2 \theta & 2 \sin \theta \cos \theta \\
2 \sin \theta \cos \theta & \sin^2 \theta - \cos^2 \theta
\end{bmatrix}
= \begin{bmatrix}
\cos 2\theta & \sin 2\theta \\
\sin 2\theta & -\cos 2\theta
\end{bmatrix}
\quad \text{(half-wave plate)}
\] (4.43)

Remember that \( \theta \) refers to the angle that the fast axis makes with respect to the \( x \)-axis.

Before moving on, consider the following two examples that illustrate how wave plates are often used:

---

**Example 4.2**
Calculate the Jones matrix for a quarter wave plate at \( \theta = 45^\circ \), and calculate its effect on horizontally polarized light.

**Solution:** At \( \theta = 45^\circ \), the Jones matrix for the quarter-wave plate (4.41) reduces to

\[
\frac{e^{i\pi/4}}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \quad \text{(quarter-wave plate, fast axis at } \theta = 45^\circ) \quad (4.44)
\]

The overall phase 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.

Now we calculate the effect of the quarter wave plates (oriented at \( \theta = 45^\circ \)) operating on horizontally polarized light:

\[
\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix} \quad (4.45)
\]

Notice that a quarter-wave plate (properly oriented) turns linearly polarized light into right-circularly polarized light (see Table 4.1).

**Example 4.3**

Calculate the effect of a half wave plate at an arbitrary \( \theta \) on horizontally polarized light.

**Solution:** Carrying out the multiplication, we obtain

\[
\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos 2\theta \\ \sin 2\theta \end{bmatrix} \quad (4.46)
\]

The resulting Jones vector describes linearly polarized light an angle of \( 2\theta \) from the \( x \)-axis.

This example illustrates that a half wave plate rotates the polarization angle of linearly polarized light to another angle (the amount of rotation depending on the value of \( \theta \)) while preserving the linear polarization.

### 4.8 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

$$
\begin{bmatrix}
-r_p & 0 \\
0 & r_s
\end{bmatrix}
$$

(Jones matrix for reflection) (4.47)

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 $\pi$ 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 (4.47) 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
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).

If a beam of light encounters a series of mirrors, the final polarization is determined by multiplying the sequence of appropriate Jones matrices (4.47) 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 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 $\theta$ in (4.30). This rotation can be accomplished by multiplying the following matrix onto the incident Jones vector:

$$
\begin{bmatrix}
cos\theta & -\sin\theta \\
\sin\theta & \cos\theta
\end{bmatrix}
\quad \text{(rotation of coordinates through an angle $\theta$)} 
$$

This is a rotation about the $z$-axis, and the angle of rotation $\theta$ 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 its representation.
4.9 Ellipsometry

Measuring the polarization of light reflected from a surface can yield information regarding the optical constants of that surface (i.e. \( n \) and \( \kappa \)). 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 \( \kappa \) at a range of frequencies, and this would require a different quarter-wave plate thickness \( d \) for each wavelength used (see (4.38)). 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 \( \kappa \) can be pinned down (allowing a 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 \( \alpha \) to give both \( s \) and \( p \)-components). The Jones matrix for the test surface reflection is given by (4.47), and the Jones matrix for the analyzing polarizer oriented at angle \( \theta \) is given by (4.35). The Jones vector for the light arriving at the detector is then

\[
\begin{bmatrix}
\cos^2 \theta & \sin \theta \cos \theta \\
\sin \theta \cos \theta & \sin^2 \theta
\end{bmatrix}
\begin{bmatrix}
-r_p & 0 \\
0 & r_s
\end{bmatrix}
\begin{bmatrix}
\cos \alpha \\
\sin \alpha
\end{bmatrix}
\]

\[
= \begin{bmatrix}
-r_p \cos \alpha \cos^2 \theta + r_s \sin \alpha \sin \theta \cos \theta \\
-r_p \cos \alpha \sin \theta \cos \theta + r_s \sin \alpha \sin^2 \theta
\end{bmatrix}
\]

In an ellipsometer, the angle \( \theta \) 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 \( \theta \) and \( \alpha \) are continuously varied, it is possible to extract the values of \( n \) and \( \kappa \) (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:

$$I = I_{\text{pol}} + I_{\text{un}}$$

(4.51)

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:

$$P \equiv \frac{I_{\text{pol}}}{I_{\text{pol}} + I_{\text{un}}}$$

(4.52)

The degree of polarization takes on values between zero and one. Thus, if the light is completely unpolarized (such that $I_{\text{pol}} = 0$), then the degree of polarization is zero. On the other hand, if the beam is fully polarized (such that $I_{\text{un}} = 0$), 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

$$S_0 \equiv \frac{I}{I_{\text{in}}}$$

(4.53)

is a comparison of the beam's intensity (or power) with a benchmark intensity, $I_{\text{in}}$, 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

$$S_1 \equiv \frac{2I_{\text{hor}}}{I_{\text{in}}} - S_0$$

Here, $I_{\text{hor}}$ 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 $I_{\text{in}}$ 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

$$S_2 \equiv \frac{2I_{45^\circ}}{I_{\text{in}}} - S_0$$

Similar to the previous case, $I_{45^\circ}$ represents the amount of light detected if an ideal linear polarizer is placed with its axis at $45^\circ$ 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 $45^\circ$ or $135^\circ$.

Finally, $S_3$ characterizes the extent to which the beam is either right or left circularly polarized:

$$S_3 \equiv \frac{2I_{\text{r-cir}}}{I_{\text{in}}} - S_0$$

Here, $I_{\text{r-cir}}$ 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 quarter wave plate, followed by a linear polarizer with the transmission axis aligned $45^\circ$ from the wave-plate fast axis, followed by another quarter wave plate at $-45^\circ$ from the polarizer (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 $I = I_{\text{in}}$, then we have $I_{\text{hor}} = I_{\text{in}}$, $I_{45^\circ} = I_{\text{in}}/2$, and $I_{\text{r-cir}} = I_{\text{in}}/2$. 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 $I = I_{\text{in}}/3$ but is purely left circularly polarized. Then we have $I_{\text{hor}} = I_{\text{in}}/6$, $I_{45^\circ} = I_{\text{in}}/6$, and $I_{\text{r-cir}} = 0$. 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, $I_{\text{hor}} = I_{45^\circ} = I_{\text{r-cir}} = I/2$ and $S_1 = S_2 = S_3 = 0$. 

© 2004-2009 Peatross and Ware
Example 4.4

Find the Stokes parameters for perfectly polarized light, represented by an arbitrary Jones vector

\[
\begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix}
\]

where \(A, B, \text{ and } \delta\) are all real. (Recall that depending on the values \(A, B, \text{ and } \delta\), the polarization can follow any ellipse.)

Solution: The intensity of this polarized beam is \(I_{\text{pol}} = A^2 + B^2\), according to Eq. (4.26), where we absorb the factor \(\frac{1}{2}\epsilon_0 c|E_{\text{eff}}|^2\) into \(A\) and \(B\) for convenience. The Jones vector for the light that passes through a horizontal polarizer is

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix} = \begin{bmatrix}
A \\
0
\end{bmatrix}
\]

which gives a measured intensity of \(I_{\text{hor}} = A^2\). Similarly, the Jones vector when the beam is passed through a polarizer oriented at 45° is

\[
\frac{1}{2} \begin{bmatrix}
1 & 1 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix} = \frac{1}{2} \begin{bmatrix}
A + Be^{i\delta} \\
A + Be^{i\delta}
\end{bmatrix}
\]

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

\[
\frac{1}{2} \begin{bmatrix}
1 & i \\
-i & 1
\end{bmatrix}
\begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix} = \frac{1}{2} \begin{bmatrix}
A + iB e^{i\delta} \\
B \cos \delta + i(B \sin \delta - A)
\end{bmatrix}
\]

giving an intensity of

\[
I_{\text{rc}} = \frac{A^2 - 2AB \sin \delta + B^2 \sin^2 \delta + B^2 \cos^2 \delta}{2} = \frac{A^2 + B^2 - 2AB \sin \delta}{2}
\]

Thus, the Stokes parameters become

\[
S_0 = \frac{A^2 + B^2}{I_{\text{in}}}
\]

\[
S_1 = \frac{2A^2}{I_{\text{in}}} - \frac{A^2 + B^2}{I_{\text{in}}} = \frac{A^2 - B^2}{I_{\text{in}}}
\]

\[
S_2 = \frac{A^2 + B^2 + 2AB \cos \delta}{I_{\text{in}}} - \frac{A^2 + B^2}{I_{\text{in}}} = \frac{2AB \cos \delta}{I_{\text{in}}}
\]

\[
S_3 = \frac{A^2 + B^2 - 2AB \sin \delta}{I_{\text{in}}} - \frac{A^2 + B^2}{I_{\text{in}}} = \frac{-2AB \sin \delta}{I_{\text{in}}}
\]
It is clear from the linear dependence of \( S_0, S_1, S_2, \) and \( S_3 \) on intensity (see Eqs. (4.53)–(4.56)) 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 \( S_i = S_i^{(\text{pol})} + S_i^{(\text{un})}, \ i = 0, 1, 2, 3. \)

This is certainly true for \( S_0 = \frac{I_\text{pol} + I_\text{un}}{I_\text{in}} \) (4.57) 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. (4.54)–(4.56) and therefore cancels out.

A completely general form of the Stokes vector may then be written as (see Example 4.4)

\[
\begin{bmatrix}
S_0 \\
S_1 \\
S_2 \\
S_3
\end{bmatrix}
= \frac{1}{I_\text{in}} \begin{bmatrix}
I_\text{pol} + I_\text{un} \\
A^2 - B^2 \\
2AB\cos\delta \\
2AB\sin\delta
\end{bmatrix}
\] (4.58)

where the Jones vector

\[
\begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix}
\]
describes the polarized portion of the light, which has intensity

\( I_\text{pol} = A^2 + B^2 \) (4.59)

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

\[
\sqrt{S_1^2 + S_2^2 + S_3^2} = \sqrt{\left(\frac{A^2 - B^2}{I_\text{in}}\right)^2 + \left(\frac{2AB\cos\delta}{I_\text{in}}\right)^2 + \left(\frac{2AB\sin\delta}{I_\text{in}}\right)^2}
= \frac{1}{I_\text{in}} \sqrt{(A^2 - B^2)^2 + 4A^2B^2(\cos^2\delta + \sin^2\delta)}
= \frac{A^2 + B^2}{I_\text{in}}
= \frac{I_\text{pol}}{I_\text{in}}
\] (4.60)

Substituting (4.57) and (4.60) into the expression for the degree of polarization (4.52) yields

\( P \equiv \frac{1}{S_0} \sqrt{S_1^2 + S_2^2 + S_3^2} \) (4.61)

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 (4.61).

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 4.5**

Determine the Mueller matrix that represents a linear polarizer with transmission axis at arbitrary angle \( \theta \).

**Solution:** We know that the 50% of the unpolarized light transmits through the polarizer, ending up with Jones vector

\[
\begin{bmatrix}
A'_1 \\
B'_1
\end{bmatrix} = \frac{I_{\text{un}}}{2} \begin{bmatrix}
\cos \theta \\
\sin \theta
\end{bmatrix}
\]

(see table 4.1). We also know that the Jones matrix (4.26) acts on the polarized portion of the light, represented by arbitrary Jones vector

\[
\begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix}
\]

This gives a transmitted Jones vector of

\[
\begin{bmatrix}
A'_2 \\
B'_2 e^{i\delta'}
\end{bmatrix} = \begin{bmatrix}
\cos^2 \theta & \cos \theta \sin \theta \\
\cos \theta \sin \theta & \sin^2 \theta
\end{bmatrix} \begin{bmatrix}
A \\
B e^{i\delta}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
A \cos \theta + B \sin \theta e^{i\delta} \\
B \sin \theta \cos \theta
\end{bmatrix}
\]

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

\[
|A'|^2 = |A'_1|^2 + |A'_2|^2 = \left( \frac{I_{\text{un}}}{2} + A^2 \cos^2 \theta + B^2 \sin^2 \theta + 2AB \cos \theta \sin \theta \cos \delta \right) \cos^2 \theta
\]

\[
= \left( \frac{I_{\text{un}} + A^2 + B^2}{2} + \left( A^2 - B^2 \right) \frac{\cos 2\theta}{2} + 2AB \cos \delta \frac{\sin 2\theta}{2} \right) \cos^2 \theta
\]

\[
= \left[ \frac{S_0}{2} + \frac{\cos 2\theta}{2} S_1 + \frac{\sin 2\theta}{2} S_2 \right] \cos^2 \theta
\]

Similarly,

\[
|B'|^2 = |B'_1|^2 + |B'_2|^2 = \left[ \frac{S_0}{2} + \frac{\cos 2\theta}{2} S_1 + \frac{\sin 2\theta}{2} S_2 \right] \sin^2 \theta
\]
This gives

\[ S'_0 = |A'|^2 + |B'|^2 \]
\[ = \frac{S_0}{2} + \frac{\cos 2\theta}{2} S_1 + \frac{\sin 2\theta}{2} S_2 \]
\[ S'_1 = |A'|^2 - |B'|^2 \]
\[ = \left[ \frac{S_0}{2} + \frac{\cos 2\theta}{2} S_1 + \frac{\sin 2\theta}{2} S_2 \right] (\cos^2 \theta - \sin^2 \theta) \]
\[ = \frac{S_0 \cos 2\theta}{2} + \frac{\cos^2 2\theta}{2} S_1 + \frac{\sin 4\theta}{4} S_2 \]

and since \( \delta' = 0 \) we have

\[ S'_2 = 2 |A'| |B'| \cos \delta' \]
\[ = 2 \left[ \frac{S_0}{2} + \frac{\cos 2\theta}{2} S_1 + \frac{\sin 2\theta}{2} S_2 \right] \cos \theta \sin \theta \]
\[ = \frac{S_0 \sin 2\theta}{2} + \frac{\sin 4\theta}{2} S_1 + \frac{\sin^2 2\theta}{2} S_2 \]
\[ S'_3 = 2 |A'| |B'| \sin \delta' \]
\[ = 0 \]

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

Exercises for 4.3 Jones Vectors for Representing Polarization

P4.1 Show that \((A\hat{x} + Be^{i\delta}\hat{y}) \cdot (A\hat{x} + Be^{i\delta}\hat{y})^* = 1\), 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)

\[E(z, t) = |E_{\text{eff}}| \left[ \hat{x}A\cos(kz - \omega t + \phi) + \hat{y}B\cos(kz - \omega t + \phi + \delta) \right] \]

where \(\phi\) is the phase of \(E_{\text{eff}}\). Freeze time at, say, \(t = \phi/\omega\). Determine the field at \(z = 0\) and at \(z = \lambda/4\) (a quarter cycle), say. If \(E(0, t) \times E(\lambda/4, t)\) 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](image)

HINT: Linearly polarized light contains equal amounts of right and left circularly polarized light. Consider

\[
\frac{1}{2} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} + \frac{e^{i\phi}}{2} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}
\]
where $\phi$ 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 $\alpha$ is the angle through which the polarization is rotated, beginning with horizontally polarized light. The overall phase $\delta$ is unimportant.

**Exercises for 4.5 Linear Polarizers and Jones Matrices**

**P4.5**

(a) Suppose that linearly polarized light is oriented at an angle $\alpha$ 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?

**Exercises for 4.6 Jones Matrix for Polarizers at Arbitrary Angles**

**P4.6**

Horizontally polarized light ($\alpha = 0$) is sent through two polarizers, the first oriented at $\theta_1 = 45^\circ$ and the second at $\theta_2 = 90^\circ$. 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 $\alpha$ with respect to the horizontal or $x$-axis. What fraction of the original intensity emerges from a polarizer oriented with its transmission at angle $\theta$ 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 (4.26). Then find the polarizer angle (call it $\alpha$) 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 $\pi$. You can add $\pi$ for another valid result (which corresponds to the second ellipse axis).
**Exercises for 4.7 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. $\theta_i \geq 80^\circ$). 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](image)

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 $\theta$) 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 (4.41). As you solve the matrix equation, it is helpful to note that the inverse of (4.41) 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 $n_{\text{slow}} = 1.55335$.

**Exercises for 4.8 Polarization Effects of Reflection and Transmission**

**P4.11**  Light is linearly polarized at $\alpha = 45^\circ$ with a Jones vector according to table 4.1. The light is reflected from a vertical silver mirror with angle of incidence $\theta_i = 80^\circ$, as described in (P 3.13). 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.

© 2004-2009 Peatross and Ware
Exercises 103

Figure 4.10 Geometry for P 4.11

Answer: \[
\begin{pmatrix}
0.668 \\
0.702e^{1.13i}
\end{pmatrix}.
\]

P4.12 Calculate the angle \( \theta \) 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: \( \theta \approx 50^\circ \) and \( \theta \approx 53^\circ \).

Exercises for 4.A Partially Polarized Light

P4.13 (a) One way to construct a right-circular polarizer is using a quarter wave plate with fast axis at \( 45^\circ \), followed by a linear polarizer oriented vertically, and finally a quarter wave plate with fast axis at \( -45^\circ \). Calculate the Jones matrix for this system.

Answer: \[
\frac{1}{2}
\begin{bmatrix}
1 & i \\
-i & 1
\end{bmatrix}
\]

(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

To this point, we have considered only isotropic media with $\mathbf{P} = \epsilon_0 \chi(\omega) \mathbf{E}$. The fact that the susceptibility $\chi(\omega)$ is the same in all directions for these materials leads to an index of refraction that is independent of the direction of travel and polarization. When we introduced wave plates in chapter 4, we saw that in some materials it is possible for light to experience a different index of refraction depending on the polarization of the electric field $\mathbf{E}$. This difference in the index of refraction occurs in anisotropic media, where the direction and strength of the induced dipoles depends in a non-trivial way on the direction and strength of the electric field. This behavior is observed in crystals where the lattice structure has a low degree of symmetry.\(^1\) The unique properties of anisotropic materials make them important elements in many optical systems.

In section 5.2 we discuss how to connect $\mathbf{E}$ and $\mathbf{P}$ in anisotropic media using the susceptibility tensor. In section 5.3 we apply Maxwell’s equations to a plane wave traveling in a crystal. The analysis leads to Fresnel’s equation (section 5.4), which connects the components of the $\mathbf{k}$-vector with the components of the susceptibility tensor. In section 5.6 we apply Fresnel’s equation to a uniaxial crystal (e.g. quartz, sapphire) where $\chi_x = \chi_y \neq \chi_z$. In section 5.8 we examine the flow of energy in a uniaxial crystal and show that the Poynting vector and the $\mathbf{k}$-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 $\mathbf{k}$-vector always obeys Snell’s law, but in a crystal the $\mathbf{k}$-vector points in a different direction than the Poynting vector, and it is the

\(^1\)Not all crystals are anisotropic. For instance, crystals with a cubic lattice structure (such as NaCl) are highly symmetric and respond to electric fields the same in any direction.
Poynting vector that delivers the energy seen by an observer.

### 5.2 Constitutive Relation in Crystals

In this section, we explore the connection between the polarization $\mathbf{P}$ of the medium and the electric field $\mathbf{E}$ in an anisotropic crystal. In this type of crystal, the lattice can cause directional asymmetries so that the $\mathbf{P}$ in the crystal does not necessarily respond in the same direction as the electric field $\mathbf{E}$ (i.e. $\mathbf{P} \neq \epsilon_0 \chi \mathbf{E}$). However, at low intensities the response of materials is still linear (or proportional) to the strength of the electric field. The linear constitutive relation which connects $\mathbf{P}$ to $\mathbf{E}$ in a crystal can be expressed in its most general form as

$$
\begin{bmatrix}
P_x \\
P_y \\
P_z 
\end{bmatrix} = \epsilon_0 \begin{bmatrix}
\chi_{xx} & \chi_{xy} & \chi_{xz} \\
\chi_{yx} & \chi_{yy} & \chi_{yz} \\
\chi_{zx} & \chi_{zy} & \chi_{zz}
\end{bmatrix} \begin{bmatrix}
E_x \\
E_y \\
E_z 
\end{bmatrix}
$$

(5.1)

The matrix in (5.1) is called the susceptibility tensor. To visualize the behavior of electrons in such a material we imagine the electron bound as though by tiny springs with different strengths in different dimensions to represent the anisotropy (see Fig. 5.1). When an external electric field is applied, the electron experiences a force that moves it from its equilibrium position. The “springs” (actually the electric force from ions bound in the crystal lattice) exert a restoring force, but the restoring force is not equal in all directions—the electron tends to move more along the dimension of the weaker spring. The displaced electron creates a microscopic dipole, but the asymmetric restoring force causes $\mathbf{P}$ to be in a direction different than $\mathbf{E}$.

To understand the geometrical interpretation of the many coefficients $\chi_{ij}$, assume, for example, that the electric field is directed along the $x$-axis (i.e. $E_y = E_z = 0$) as depicted in Fig. 5.1. In this case, the three equations encapsulated in (5.1) reduce to

$$
\begin{align*}
P_x &= \epsilon_0 \chi_{xx} E_x \\
P_y &= \epsilon_0 \chi_{yx} E_x \\
P_z &= \epsilon_0 \chi_{zx} E_x
\end{align*}
$$

The coefficient $\chi_{xx}$ connects the strength of $\mathbf{P}$ in the $x$ direction with the strength of $\mathbf{E}$ in that same direction. The coefficients $\chi_{yx}$ indicate the amount of $\mathbf{P}$ in the $y$-dimension produced by the electric field component in the $x$-dimension. The other coefficients with mixed subscripts in (5.1) likewise describe the contribution to $\mathbf{P}$ in one dimension made by an electric field component in another dimension.

As you might imagine, working with nine susceptibility coefficients can get complicated. Fortunately, we can greatly reduce the complexity of the description by a judicious choice of coordinate system. For ordinary crystals, it turns out that if we assume that there is no absorption in the crystal, conservation of energy
requires that the susceptibility tensor in (5.1) be real and symmetric ($\chi_{ij} = \chi_{ji}$).\(^2\)

In Appendix 5.A we show that these requirements allow us to find a coordinate system for which off-diagonal elements of the tensor vanish. 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. In that frame the constitutive relation simplifies to

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} \chi_{xx} & 0 & 0 \\ 0 & \chi_{yy} & 0 \\ 0 & 0 & \chi_{zz} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}$$

(5.2)

or

$$\mathbf{P} = \hat{x}\epsilon_0\chi_{xx}E_x + \hat{y}\epsilon_0\chi_{yy}E_y + \hat{z}\epsilon_0\chi_{zz}E_z$$

(5.3)

The assumption of no absorption requires that the diagonal elements of the matrix in (5.2) be real.

**Example 5.1**

Show that the assumption that a medium is non-absorbing implies that the susceptibility tensor is symmetric.

**Solution:** We assume that $\mathbf{P}$ is due to a single species of electron, so that we have $\mathbf{P} = N \mathbf{p}$. Here $N$ is the number of microscopic dipoles per volume and $\mathbf{p} = q_e \mathbf{r}$, where $q_e$ is the charge on the electron and $\mathbf{r}$ is the microscopic displacement of the electron. The force on this electron due to the electric field is given by $\mathbf{F} = \mathbf{E}q_e$. With these definitions, we can use (5.1) to write a connection between the force due to a static $\mathbf{E}$ and the electron displacement:

$$N q_e \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{\epsilon_0}{q_e} \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{bmatrix} \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix}$$

(5.4)

The column vector on the left represents the components of the displacement $\mathbf{r}$. We next invert (5.4) to find the force of the electric field on an electron as a function of its displacement:\(^3\)

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \begin{bmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{yx} & k_{yy} & k_{yz} \\ k_{zx} & k_{zy} & k_{zz} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

(5.5)

\(^2\)By “ordinary” we mean that the crystal does not exhibit optical activity. Optically active crystals have a complex susceptibility tensor, even when no absorption takes place. Conservation of energy in this more general case requires that the susceptibility tensor be Hermitian ($\chi_{ij} = \chi_{ji}^\ast$). See Section ?? for more details.

\(^3\)This inversion assumes the field changes slowly so the forces on the electron are always essentially balanced. This is not true for optical fields, but the proof gives the right flavor for why conservation of energy results in the symmetry. A more formal proof that doesn’t make this assumption can be found in *Principles of Optics, 7th Ed.*, Born and Wolf, pp. 790-791 (Ref. [1]).
where
\[
\begin{pmatrix}
k_{xx} & k_{xy} & k_{xz} \\
k_{yx} & k_{yy} & k_{yz} \\
k_{zx} & k_{zy} & k_{zz}
\end{pmatrix}
\equiv
\frac{N q_e^2}{\epsilon_0}
\begin{pmatrix}
\chi_{xx} & \chi_{xy} & \chi_{xz} \\
\chi_{yx} & \chi_{yy} & \chi_{yz} \\
\chi_{zx} & \chi_{zy} & \chi_{zz}
\end{pmatrix}^{-1}
\] (5.6)

The total work done on an electron in moving it to its displaced position is given by
\[
W = \int_{\text{path}} F(r') \cdot dr'
\] (5.7)

While there are many possible paths for getting the electron to any specific displacement (each path specified by a different history of the electric field), the work done along any of these paths must be the same if the system is conservative (i.e. no absorption). For example, for a final displacement of \( r = x\hat{x} + y\hat{y} \) we could have the following two paths:

We can use (5.5) in (5.7) to calculate the total work done on the electron along path 1:
\[
W = \int_0^x F_x(x', y' = 0, z' = 0) dx' + \int_0^y F_y(x' = x, y', z' = 0) dy'
= \int_0^x k_{xx} x' dx' + \int_0^y (k_{yx} x + k_{yy} y') dy'
= \frac{k_{xx}}{2} x^2 + k_{yx} x y + \frac{k_{yy}}{2} y^2
\]

If we take path 2, the total work is
\[
W = \int_0^y F_y(x' = 0, y', z' = 0) dy' + \int_0^x F_x(x', y' = y, z' = 0) dx'
= \int_0^y k_{yy} y' dy' + \int_0^x (k_{xx} x' + k_{xy} y) dx'
= \frac{k_{yy}}{2} y^2 + k_{xy} x y + \frac{k_{xx}}{2} x^2
\]

Since the work must be the same for these two paths, we clearly have \( k_{xy} = k_{yx} \).

Similar arguments for other pairs of dimensions ensure that the matrix of \( k \) coefficients is symmetric. From linear algebra, we learn that if the inverse of a matrix is symmetric then the matrix itself is also symmetric. When we combine this result with the definition (5.6), we see that the assumption of no absorption requires the susceptibility matrix to be symmetric.

### 5.3 Plane Wave Propagation in Crystals

Now we are ready to search for solutions to the wave equation in a crystal. As a trial solution, we consider a plane wave with frequency \( \omega \), similar to the plane-wave solution we have studied in isotropic materials. In this case, the fields \( E, B, \)
and $P$ are all associated with the same plane wave according to

$$
E = E_0 e^{i(k \cdot r - \omega t)}
$$

$$
B = B_0 e^{i(k \cdot r - \omega t)}
$$

$$
P = P_0 e^{i(k \cdot r - \omega t)}
$$

As usual, the phase of each wave is included in the amplitudes $E_0$, $B_0$, and $P_0$.

We can make several observations about the behavior of these fields by applying Maxwell's equations directly. Gauss's law for electric fields requires

$$
\nabla \cdot (\varepsilon_0 E + P) = 0
$$

and Gauss's law for magnetism gives

$$
\nabla \cdot B = 0
$$

When our trial solutions (5.8) are inserted into these, we find

$$
k \cdot (\varepsilon_0 E + P) = 0
$$

and

$$
k \cdot B = 0
$$

Notice that we have the following peculiarity: From its definition, the Poynting vector $S \equiv E \times B/\mu_0$ is perpendicular to both $E$ and $B$, and by (5.12) the $k$-vector is perpendicular to $B$. However, by (5.11) the $k$-vector is not necessarily perpendicular to $E$, since in general $k \cdot E \neq 0$ if $P$ points in a direction other than $E$. Therefore, $k$ and $S$ are not necessarily parallel in a crystal. In other words, the flow of energy and the direction of the wave propagation can be different.

Now we consider the behavior of our trial fields in the wave equation (1.49). Under the assumption $J_{\text{free}} = 0$, we have

$$
\nabla^2 E - \mu_0 \varepsilon_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P}{\partial t^2} + \nabla (\nabla \cdot E)
$$

As with the isotropic case, we next substitute our trial solutions (5.8) into the wave equation to find a dispersion relation that imposes requirements that $k$ and $\omega$ must satisfy in order for the fields to be consistent with Maxwell's equations. Example 5.2 illustrates how the dispersion relation is found.

**Example 5.2**

Obtain the dispersion relation in a crystal by substituting the proposed solution (5.8) into the wave equation (5.13).

**Solution:** After substituting the fields into the wave equation and carrying out the derivatives, we find

$$
k^2 E - \omega^2 \mu_0 (\varepsilon_0 E + P) = k (k \cdot E)
$$
Inserting the constitutive relation (5.3) for crystals into (5.14) yields

\[ k^2 \mathbf{E} - \omega^2 \mu_0 \varepsilon_0 \left[ (1 + \chi_x) E_x \hat{x} + (1 + \chi_y) E_y \hat{y} + (1 + \chi_z) E_z \hat{z} \right] = \mathbf{k} (\mathbf{k} \cdot \mathbf{E}) \]  

(5.15)

This relationship is unwieldy because of the mix of electric field components that appear in the expression. This was not 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.

Relation (5.15) actually contains three equations, one for each dimension. Explicitly, these equations are

\[ \begin{align*}
    \left[ k^2 - \frac{\omega^2}{c^2} (1 + \chi_x) \right] E_x &= k_x (\mathbf{k} \cdot \mathbf{E}) \\
    \left[ k^2 - \frac{\omega^2}{c^2} (1 + \chi_y) \right] E_y &= k_y (\mathbf{k} \cdot \mathbf{E}) \\
    \left[ k^2 - \frac{\omega^2}{c^2} (1 + \chi_z) \right] E_z &= k_z (\mathbf{k} \cdot \mathbf{E})
\end{align*} \]  

(5.16)\( (5.17)\) and

(5.18)

We have replaced the constants \( \mu_0 \varepsilon_0 \) with \( 1/c^2 \) according to (1.51). We multiply (5.16)–(5.18) 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

\[
\frac{k_x^2 (\mathbf{k} \cdot \mathbf{E})}{k^2 - \frac{\omega^2(1 + \chi_x)}{c^2}} + \frac{k_y^2 (\mathbf{k} \cdot \mathbf{E})}{k^2 - \frac{\omega^2(1 + \chi_y)}{c^2}} + \frac{k_z^2 (\mathbf{k} \cdot \mathbf{E})}{k^2 - \frac{\omega^2(1 + \chi_z)}{c^2}} = k_x E_x + k_y E_y + k_z E_z = (\mathbf{k} \cdot \mathbf{E})
\]  

(5.19)

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 \( \omega^2/c^2 \) we have our dispersion relation unencumbered by field components:

\[ \frac{k_x^2}{k^2 c^2/\omega^2 - (1 + \chi_x)} + \frac{k_y^2}{k^2 c^2/\omega^2 - (1 + \chi_y)} + \frac{k_z^2}{k^2 c^2/\omega^2 - (1 + \chi_z)} = \frac{\omega^2}{c^2} \]  

(5.20)

The dispersion relation (5.20) found in Example 5.2 allows us to find a suitable \( \mathbf{k} \), given values for \( \omega, \chi_x, \chi_y, \) and \( \chi_z \). 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 (i.e. we must choose the ratios between \( k_x, k_y, \) and \( k_z \)). To remind ourselves of this fact, we introduce a unit vector that points in the direction of the \( \mathbf{k} \):

\[
\mathbf{k} = k_x \hat{x} + k_y \hat{y} + k_z \hat{z} = k (u_x \hat{x} + u_y \hat{y} + u_z \hat{z}) = k \hat{u}
\]  

(5.21)

With this unit vector inserted, the dispersion relation (5.20) for plane waves in a crystal becomes

\[ \frac{u_x^2}{k^2 c^2/\omega^2 - (1 + \chi_x)} + \frac{u_y^2}{k^2 c^2/\omega^2 - (1 + \chi_y)} + \frac{u_z^2}{k^2 c^2/\omega^2 - (1 + \chi_z)} = \frac{\omega^2}{k^2 c^2} \]  

(5.22)
5.4 Fresnel’s Equation

We are now ready to introduce the refractive index for anisotropic materials. As we have seen before, the speed of a wave having the form (5.8) 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 (2.19)). Therefore, the refractive index for the wave is

\[
 n = \frac{kc}{\omega}
\]  

(5.23)

Although (5.23) looks innocent enough—and we have seen it before—the relationship between \( k \) and \( \omega \) now depends on the direction of propagation in the crystal according to (5.22).

Motivated by the relation between the index and the susceptibility in the isotropic case (2.19), we replace the susceptibility parameters in (5.22) with three new constants:

\[
 n_x \equiv \sqrt{1 + \chi_x}
\]

\[
 n_y \equiv \sqrt{1 + \chi_y}
\]

\[
 n_z \equiv \sqrt{1 + \chi_z}
\]

(5.24)

Using these definitions along with (5.23) we can write the dispersion relation (5.22) in the following convenient form:

\[
 \frac{u_x^2}{(n^2 - n_x^2)} + \frac{u_y^2}{(n^2 - n_y^2)} + \frac{u_z^2}{(n^2 - n_z^2)} = \frac{1}{n^2}
\]

(5.25)

Equation (5.25) 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 \( \mathbf{k} \)-vector (i.e. the direction of the unit vector \( \mathbf{\hat{u}} \)). For a given \( \mathbf{k} \)-vector there are two possible values for \( n \), one associated with each of the two polarization directions of the electric field.

When Fresnel’s equation (5.25) is solved to find the two values of \( n \) associated with a given \( \mathbf{\hat{u}} \) (see P 5.1), the resulting solution is quadratic in \( n^2 \). The solutions can be written as\(^4\)

\[
 n^2 = \frac{B \pm \sqrt{B^2 - 4AC}}{2A}
\]

(5.26)

where

\[
 A \equiv u_x^2 n_x^2 + u_y^2 n_y^2 + u_z^2 n_z^2
\]

(5.27)

\[
 B \equiv u_x^2 n_x^2 \left( n_y^2 + n_z^2 \right) + u_y^2 n_y^2 \left( n_x^2 + n_z^2 \right) + u_z^2 n_z^2 \left( n_x^2 + n_y^2 \right)
\]

(5.28)

\[
 C \equiv n_x^2 n_y^2 n_z^2
\]

(5.29)

The upper and lower signs in (5.26) give two solutions for \( n \), one large and one small (associated with the “+” and “−”, respectively). These are often referred to as the slow and fast index, respectively, because the waves associated

\(^4\)It is possible to write this solution in many equivalent forms using the identity \( u_x^2 + u_y^2 + u_z^2 = 1 \).
with these indexes propagate at speed \( v = c/n \). In the special cases of propagation along one of the principal axes of the crystal, the index \( n \) takes on the two values \( n_x, n_y, \) or \( n_z \) associated with the axes orthogonal to propagation (see Example 5.3).

**Example 5.3**

Calculate the two possible values for the index of refraction when \( \mathbf{k} \) is in the \( \hat{z} \) direction (in the crystal principal frame).

**Solution:** With \( u_z = 1 \) and \( u_x = u_y = 0 \) we have

\[
A = n_z^2; \quad B = n_z^2 \left( n_x^2 + n_y^2 \right); \quad C = n_x^2 n_y^2 n_z^2
\]

The square-root term is then

\[
\sqrt{B^2 - 4AC} = \sqrt{n_z^4 \left( n_x^4 + 2n_x^2 n_y^2 + n_y^4 \right) - 4n_x^2 n_y^2 n_z^4}
\]

\[
= \sqrt{n_x^4 (n_x^2 - n_y^2)^2 - 4n_x^2 n_y^2 n_z^4}
\]

\[
= n_x^2 \left( n_x^2 - n_y^2 \right)
\]

Inserting this expression into Fresnel’s equation, we find the two values for the index

\[
n = n_x, n_y
\]

The index \( n_x \) is experienced by light polarized in the \( x \)-dimension, and the index \( n_y \) is experienced by light traveling in the \( y \)-dimension.

It is often convenient to use spherical coordinates to represent the components of \( \mathbf{\hat{u}} \):

\[
\begin{align*}
    u_x &= \sin \theta \cos \phi \\
    u_y &= \sin \theta \sin \phi \\
    u_z &= \cos \theta
\end{align*}
\]  

(5.30)

Here \( \theta \) is the polar angle measured from the \( z \)-axis of the crystal and \( \phi \) is the azimuthal angle measured from the \( x \)-axis of the crystal. These equations emphasize the fact that there are only two degrees of freedom when specifying propagation direction (\( \theta \) and \( \phi \)). It is important to remember that these angles must be specified in the frame of the crystal’s principal axes, which is usually not aligned with the faces of a cut crystal in an optical setup.

### 5.5 Polarization in Crystals

The two values for \( n \) associated with a given \( \mathbf{\hat{u}} \) correspond to two distinct polarization components of the field. In Example 5.3 the natural polarization components
were along the principal axes, but for propagation in an arbitrary direction the polarization components are in other directions. Therefore, every propagation direction \( \hat{u} \) has its own natural set of polarization components. The two polarization components travel at different speeds, so even though the frequency is the same for both components, the wavelength for each is different (within the crystal).

To find the direction of the polarization components associated with the two values of \( n \), we return to Maxwell’s equations. When substitute our solution (5.8) into Faraday’s Law (1.44) and Ampere’s Law (1.45) (under the assumption \( J_{\text{free}} = 0 \)) we obtain the following requirements on the solution:

\[
\mathbf{k} \times \mathbf{E} = \omega \mathbf{B} \quad (5.31)
\]

and

\[
\mathbf{k} \times \mathbf{B} = -\mu_0 \omega (\varepsilon_0 \mathbf{E} + \mathbf{P}) \quad (5.32)
\]

We can combine these two equations to eliminate \( \mathbf{B} \)

\[
\mathbf{k} \times (\mathbf{k} \times \mathbf{E}) + \mu_0 \omega^2 (\varepsilon_0 \mathbf{E} + \mathbf{P}) = 0 \quad (5.33)
\]

and then apply the linear constitutive relation (5.2) to find (after carrying out the cross products and a little algebra) the following requirement on the electric field vector

\[
\begin{bmatrix}
\frac{n_x^2}{n^2} - u_x^2 - u_y^2 & u_x u_y & u_x u_z \\
u_x u_y & \frac{n_y^2}{n^2} - u_y^2 - u_z^2 & u_y u_z \\
u_x u_z & u_y u_z & \frac{n_z^2}{n^2} - u_x^2 - u_y^2
\end{bmatrix}
\begin{bmatrix}
E_x \\
E_y \\
E_z
\end{bmatrix} = 0 \quad (5.34)
\]

For (5.34) to have a non-trivial solution, the determinant of the matrix must be zero. Imposing this requirement is an equivalent way to derive Fresnel’s equation (5.25) for \( n \).

Given a direction for \( \hat{u} \) and a value for \( n \) (from Fresnel’s equation), we can use (5.34) to determine which direction the electric field associated with that index must oscillate. When all three equations represented by (5.34) are coupled, the appropriate field direction for a value of \( n \) is given by (see P 5.3)

\[
\begin{bmatrix}
E_x \\
E_y \\
E_z
\end{bmatrix} \propto \begin{bmatrix}
u_x \\
n^2 - n_x^2 \\
u_y \\
n^2 - n_y^2 \\
u_z \\
n^2 - n_z^2
\end{bmatrix} \quad (5.35)
\]

This is a proportionality rather than an equation, since Maxwell’s equation only specifies the direction of \( \mathbf{E} \)—we are free to choose the amplitude. Because Fresnel’s equation gives two values for \( n \), (5.35) specifies two distinct polarization
components associated with each \( \hat{u} \). These polarization components form a natural basis for describing light propagation in a crystal. When light is composed of a mixture of these two polarizations, the two polarization components experience different indexes of refraction.

If any of the components of \( \hat{u} \) (i.e. \( u_x, u_y, \) or \( u_z \)) is precisely zero, (5.35) fails to provide a well-defined direction for the polarization. This is because at least one of the dimensions in the system of equations represented by (5.34) is decoupled from the others. In these cases, it is necessary to go back to (5.34) and re-solve for the polarization directions.

**Example 5.4**

Determine the directions of the two polarization components associated with the values of \( n \) for a propagation direction \( \hat{u} = \hat{z} \)

**Solution:** In this case we have \( u_x = u_y = 0 \), so as noted above, have to go back to (5.34) and re-solve. In our case, the set of equations becomes

\[
\begin{bmatrix}
\frac{n_x^2}{n^2} - 1 & 0 & 0 \\
0 & \frac{n_y^2}{n^2} - 1 & 0 \\
0 & 0 & \frac{n_z^2}{n^2}
\end{bmatrix}
\begin{bmatrix}
E_x \\
E_y \\
E_z
\end{bmatrix} = 0
\]

(5.36)

Notice that all three dimensions are decoupled in this system (i.e. there are no off-diagonal terms). In Example 5.3 we found that the two values of \( n \) associated with \( \hat{u} = \hat{z} \) are \( n_x \) and \( n_y \). If we use \( n = n_x \) in our set of equations, we have

\[
\begin{bmatrix}
0 & 0 & 0 \\
0 & \frac{n_y^2}{n_x^2} - 1 & 0 \\
0 & 0 & \frac{n_z^2}{n_x^2}
\end{bmatrix}
\begin{bmatrix}
E_x \\
E_y \\
E_z
\end{bmatrix} = 0
\]

Assuming \( n_x, n_y, \) and \( n_z \) are all unique so that \( n_y/n_x \neq 1 \) and \( n_z/n_x \neq 1 \), these equations require \( E_y = E_z = 0 \) but allow \( E_x \) to be non-zero. This proves our earlier assertion that the index \( n_x \) is associated with light polarized in the \( x \)-dimension in the special case of \( \hat{u} = \hat{z} \). Similarly, when \( n_y \) is inserted into (5.36), we find that it is associated with light polarized in the \( y \)-dimension.

We can use (5.35) to study the behavior of polarization direction as the direction of propagation varies. Figure 5.3 shows plots of the polarization direction (i.e. normalized \( E_x, E_y, \) and \( E_z \)) in Potassium Niobate as the propagation direction is varied. The plot is created by inserting the spherical representation of \( \hat{u} \) (5.30) into Fresnel’s equation (5.26) for a chosen sign of the \( \pm \), and then inserting the resulting \( n \) into (5.35) to find the associated electric field. As we saw in Example 5.4, at \( \theta = 0 \) the light associated with the slow index is polarized along the \( y \)-axis and the light associated with the fast index is polarized along the \( x \)-axis.
In Fig. 5.3(c) we have plotted the angle between the two polarization components. At \( \theta = 0 \), the two polarization components are orthogonal, as one would expect. However, notice that in other propagation directions the two polarization components are not orthogonal. This feature is one of the things that makes describing light in crystals challenging.

Before moving on, let us briefly summarize what has been accomplished so far. Given values for \( \chi_x, \chi_y, \) and \( \chi_z \) associated with light in a crystal at a given frequency, one defines the indices \( n_x, n_y, \) and \( n_z, \) according to (5.24). Next, a direction for the \( k \)-vector is chosen (i.e. \( u_x, u_y, \) and \( u_z, \) or equivalently \( \theta \) and \( \phi \)). This direction potentially has two values for the index of refraction associated with it, found using Fresnel's equation (5.26). Each index can then be associated with a polarization direction for the electric field, found using (5.35).

### 5.6 Biaxial and Uniaxial Crystals

All anisotropic crystals have certain special propagation directions where the two values for \( n \) from Fresnel's equation are equal. These directions are referred to as the *optic axes* of the crystal. When propagating along an optic axis, all polarization components experience the same index of refraction. If the values of \( n_x, n_y, \) and \( n_z \) are all unique, a crystal will have two optic axes, and hence is referred to as a biaxial crystal. By convention, we order the crystal axes for biaxial crystals so that \( n_x < n_y < n_z. \) Under this convention, the two optic axes occur in the \( x-z \) plane (\( \phi = 0 \)) at the following polar angles \( \theta, \) measured from the \( z \)-axis (see P 5.4):

\[
\cos \theta = \pm \frac{n_x}{n_y} \sqrt{\frac{n_z^2 - n_y^2}{n_z^2 - n_x^2}} \quad \text{(Optic axes directions, biaxial crystal)} \quad (5.37)
\]

Describing light in biaxial crystals can get complicated because, among other things, the polarization components specified by (5.35) are not necessarily orthogonal to one another. Thus, for the remainder of this chapter, we will focus on the simpler case of uniaxial crystals.

![Figure 5.3](image-url) Polarization direction associated with the two values of \( n \) in Potassium Niobate (KNbO\(_3\)) at \( \lambda = 500 \) nm \( (n_x = 2.22, \ n_y = 2.34, \) and \( n_z = 2.41) \) and \( \phi = \pi/4. \) Frame (c) shows the angle between the two polarization components.
In uniaxial crystals two of the coefficients $\chi_x$, $\chi_y$, and $\chi_z$ are the same. In this case, there is only one optic axis for the crystal (hence the name uniaxial). By convention, in uniaxial crystals we label the dimension that has the unique susceptibility as the $z$-axis (i.e. $\chi_x = \chi_y \neq \chi_z$). This makes the $z$-axis the optic axis, since the the two values for $n$ in this direction are the same (i.e. $n_x = n_y$). The unique index of refraction is called the extraordinary index

$$n_z = n_e$$

(5.38)

and the other index is the ordinary index

$$n_x = n_y = n_0$$

(5.39)

These names were coined by Huygens, one of the early scientists to study light in crystals (see appendix 5.B). A uniaxial crystal with $n_e > n_0$ is referred to as a positive crystal, and one with $n_e < n_0$ is referred to as a negative crystal.

To calculate the index of refraction for a wave propagating in a uniaxial crystal, we use definitions (5.38) and (5.39) along with the spherical representation of $\hat{u}$ (5.30) in Fresnel’s equation (5.26) to find the following two values for $n$ (see P 5.5):

$$n = n_0 \quad \text{(uniaxial crystal)} \quad (5.40)$$

and

$$n = n_e(\theta) \equiv \frac{n_0 n_e}{\sqrt{n_0^2 \sin^2 \theta + n_e^2 \cos^2 \theta}} \quad \text{(uniaxial crystal)} \quad (5.41)$$

The index $n_e(\theta)$ in (5.41) is also commonly referred to as the extraordinary index along with the constant $n_e = n_z$. While this has the potential for some confusion, the practice is so common that we will continue it here. In this book we will write $n_e(\theta)$ when the angle dependent quantity specified by (5.41) is required, and write $n_e$ in formulas where the constant (5.38) is called for (as in the right hand side of (5.41)). Notice that $n_e(\theta)$ depends only on $\theta$ (the polar angle measured to the optic axis $\hat{z}$) and not $\phi$ (the azimuthal angle).

The first index (5.40) corresponds to a polarization component which points orthogonal to the plane containing $\hat{u}$ and $\hat{z}$ (e.g. if $\hat{u}$ is in the $x$-$z$ plane, $n_0$ is associated with light polarized in the $y$-dimension).

$$\mathbf{E}_0(\hat{u}) \propto \begin{bmatrix} -\sin \phi \\ \cos \phi \\ 0 \end{bmatrix} \quad (5.42)$$

This is shown by inserting $n = n_0$ into the requirement (5.34), and finding the allowed fields (see P 5.6). This field component is associated with the ordinary wave because just as in an isotropic medium such as glass, the index of refraction for light with this polarization does not vary with $\theta$. The polarization component
associated with $n_e(\theta)$ is found by using (5.35):

$$
E_e(\hat{u}) \propto \begin{bmatrix}
\sin \theta \cos \phi \\
\frac{n_e^2(\theta) - n_o^2}{n_e^2(\theta) - n_o^2} \\
\sin \theta \sin \phi \\
\frac{n_e^2(\theta) - n_o^2}{n_e^2(\theta) - n_o^2} \\
\cos \theta \\
\frac{n_e^2(\theta) - n_o^2}
\end{bmatrix}$$

(5.43)

Notice that this polarization component is partially directed along the optic axis (i.e. it has a $z$-component), and it is not perpendicular to $k$ since $\hat{u} \cdot E_e(\hat{u}) \neq 0$ (see P 5.7). It is, however, perpendicular to the ordinary polarization component, since $E_o \cdot E_o = 0$.

If $\theta = 0$, then the $k$-vector is directed exactly along the optic axis, and no polarization component experiences the unusual dimension (i.e. the $z$-direction). Notice that when $\theta = 0$, (5.41) reduces to $n = n_o$ so that both indices are the same. On the other hand, if $\theta = \pi/2$ then (5.41) reduces to $n = n_e$. This is why a wave plate is cut with the optic axis parallel to the surface. For light entering the wave plate at normal incidence, the propagation direction is $\theta = \pi/2$, and there is a slow and a fast axis with indices $n_o$ and $n_e$.

### 5.7 Refraction at a Crystal Surface

Now we consider refraction as light enters a uniaxial crystal. Snell’s law (3.7) describes the connection between the $k$-vectors incident upon and transmitted through the surface. In a crystal, one must consider the portion of the incoming light that will experience the ordinary and extraordinary indexes separately. Because they experience different indexes, the ordinary and extraordinary 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.

The ordinary polarization component obeys Snell’s law as usual, with index $n_o$. If we assume that the index outside of the crystal is $n_i = 1$, Snell’s law may be written as

$$
\sin \theta_i = n_o \sin \theta_t \quad \text{(ordinary polarized light)}
$$

(5.44)

where $n$ is the index inside the crystal. The extraordinary polarized light still obeys Snell’s law, but now the index of refraction in the crystal depends on direction of propagation inside the crystal relative to the optic axis. Mathematically, we have

$$
\sin \theta_i = n_e(\theta') \sin \theta_t \quad \text{(extraordinary polarized light)}
$$

(5.45)

where $\theta'$ is the angle between the optic axis inside the crystal and the direction of propagation in the crystal (given by $\theta_t$ in the plane of incidence). When the optic axis is at an arbitrary angle with respect to the surface the connection between $\theta'$
and $\theta_t$ must be specified by rotation matrices, and getting a general expression for Snell's law involves solving a quartic equation so that a general solution is awkward. In Example 5.5 we illustrate a specific case where it is possible to find a simple analytic expression for Snell's law, but in the general case one would usually just solve the equations numerically.

**Example 5.5**

Find Snell's for extraordinary polarized light in the special case where a crystal is cut such that the optic axis lies perpendicular to the surface (not the way a wave plate is cut) as shown in Fig. 5.4.

![Figure 5.4](image_url)

*Figure 5.4 Propagation of light in a uniaxial crystal with its optic axis perpendicular to the surface.*

**Solution:** In this case, the connection between propagation direction and the optic axis in this case is simply $\theta_t = \theta'$. If the light hits the crystal surface straight on, the index of refraction is $n_o$, regardless of the orientation of polarization since $\theta' = 0$ for normal incidence. When the light strikes the surface at an angle, $s$-polarized light continues to experience the index $n_o$, but $p$-polarized light experiences an index that varies with angle via (5.41) since the electric field has a component of polarization along the optic axis. (The correspondence between $s$ and $p$ and ordinary and extraordinary polarization components is specific to the orientation of the optic axis in this example. For arbitrary orientations of the optic axis with respect to the surface, the ordinary and extraordinary components will generally be mixtures of $s$ and $p$ polarized light.)

When we insert (5.41) into Snell's law (5.45) with $\theta' = \theta_t$, we can invert it to find the transmitted angle $\theta_t$ in terms of $\theta_i$ (see P 5.8):

$$\tan \theta_t = \frac{n_e \sin \theta_i}{n_o \sqrt{n_e^2 - \sin^2 \theta_i}} \quad \text{(extraordinary polarized, optic axis \perp surface)} \quad (5.46)$$
As strange as this formula looks, it is Snell’s law, but with an angularly dependent index.

## 5.8 Poynting Vector in a Uniaxial Crystal

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*. Since the Poynting vector dictates the direction of energy flow, it is the direction of $S$ that determines the separation of the double image seen when looking through a birefringent crystal.

Snell’s law dictates the connection between the directions of the incident and transmitted $k$-vectors. The Poynting vector $S$ for purely ordinary polarized light points in the same direction as the $k$-vector, so the direction of energy flow for ordinary polarized light also obeys Snell’s law. However, for extraordinary polarized light, the Poynting vector $S$ is not parallel to $k$ (recall the discussion in connection with (5.11) and (5.12)). Thus, the energy flow associated with extraordinary polarized light does not obey Snell’s law. When Huygens saw this, he said “how extraordinary!” (See Appendix 5.B.)

To analyze this situation, it is necessary to derive an expression for extraordinary polarized light similar to Snell’s law, but which applies to $S$ rather than to $k$. This describes the direction that the energy associated with *extraordinary rays* takes upon entering the crystal. To calculate the direction that the extraordinary polarized $S$ takes upon entering a crystal, we first calculate the direction of $k$ inside the crystal using Snell’s law (5.45). Then we use the expression (5.43) for $E$ along with $B = (k \times E)/\omega$, to evaluate $S = E \times B/\mu_0$. In general, this process is best done numerically, since Snell’s law (5.45) for extraordinary polarized light usually does not have simple analytic solutions. In Example 5.6 we study a special case where an analytic solution can be found relatively easily.

### Example 5.6

Derive the connection between the incident and transmitted Poynting vectors when extraordinary polarized light is incident on a uniaxial crystal cut with its optic axis perpendicular to the surface as in Fig. 5.4.

**Solution:** This is the same orientation as Example 5.5, so we can use the expression derived there for Snell’s law. We assume that the light propagates in the $y$-$z$ plane.
as shown in Fig. 5.4 (i.e. $\phi = \pi/2$). From (5.43) we have

$$
E_e(\hat{u}) = E_0 \begin{bmatrix}
0 \\
\sin \theta \\
\cos \theta \\
\frac{n_e^2(\theta_t) - n_o^2}{n_e^2(\theta_t) - n_o^2}
\end{bmatrix}
$$

where $\theta_t$ describes the direction of the $k$-vector:

$$
k = k \begin{bmatrix}
0 \\
\sin \theta_t \\
\cos \theta_t
\end{bmatrix}
$$

To find the direction of energy flow, we must calculate $S = E \times B/\mu_0$.

To do this we need to know $E$. From the constitutive relation (5.3) and the definitions (5.24) we have

$$
\epsilon_0 E + P = \epsilon_0 \left( (1 + \chi_x) E_x \hat{x} + (1 + \chi_y) E_y \hat{y} + (1 + \chi_z) E_z \hat{z} \right) = \epsilon_0 \left( n_o^2 E_x \hat{x} + n_o^2 E_y \hat{y} + n_o^2 E_z \hat{z} \right)
$$

Upon substitution of (5.47) and (5.48) into (5.11) we have

$$
k \cdot (\epsilon_0 E + P) = k \left( \hat{y} \sin \theta_t + \hat{z} \cos \theta_t \right) \cdot \epsilon_0 \left( n_o^2 E_x \hat{x} + n_o^2 E_y \hat{y} + n_o^2 E_z \hat{z} \right) = \epsilon_0 k \left( n_o^2 E_x \sin \theta_t + n_o^2 E_z \cos \theta_t \right) = 0
$$

Therefore, the $y$ and $z$ components of the field are related through

$$
E_z = -\frac{n_o^2 E_y}{n_e^2} \tan \theta_t
$$

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

$$
E = E_y \left( \hat{y} - \hat{z} \frac{n_o^2}{n_e^2} \tan \theta_t \right) \quad \text{(ordinary polarized)}
$$

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 (5.47) to find

$$
B = \frac{k \times E}{\omega} = \frac{k \left( \hat{y} \sin \theta_t + \hat{z} \cos \theta_t \right) \times E_y \left( \hat{y} - \hat{z} \frac{n_o^2}{n_e^2} \tan \theta_t \right)}{\omega} = -\hat{x} \frac{k E_y}{\omega} \left( \frac{n_o^2}{n_e^2} \sin \theta_t \tan \theta_t + \cos \theta_t \right)
$$
We proceed with the computation of the Poynting vector. Using (5.51) and (5.52) we get
\[
S = E \times B = \frac{\mu_0}{\omega} \left( \frac{n_o^2}{n_e^2} \sin \theta_i \tan \theta_t + \cos \theta_t \right) \hat{x} \quad (5.53)
\]
Keep in mind that \( \theta_t \) refers to the direction of the \( \mathbf{k} \)-vector. The above equation demonstrates that the Poynting vector \( \mathbf{S} \) lies along another direction. Let us label the direction of the Poynting vector with the angle \( \theta_S \). This angle can be obtained from the ratio of the two vector components of \( \mathbf{S} \) as follows:
\[
\tan \theta_S = \frac{S_y}{S_z} = \frac{n_o^2}{n_e^2} \tan \theta_t \quad \text{(extraordinary polarized)} \quad (5.54)
\]
While the \( \mathbf{k} \)-vector is characterized by the angle \( \theta_t \), the Poynting vector is characterized by the angle \( \theta_S \). We can find the connection between the incident angle \( \theta_i \) and \( \theta_S \) by taking advantage of the already known connection between \( \theta_i \) and \( \theta_t \). Combining (5.46) and (5.54), we obtain
\[
\tan \theta_S = \frac{n_o \sin \theta_i}{n_e \sqrt{n_e^2 - \sin^2 \theta_i}} \quad \text{(extraordinary polarized)} \quad (5.55)
\]
As we noted in the last example, we have the case where ordinary polarized light is \( s \)-polarized light, and extraordinary polarized light is \( p \)-polarized light due to our specific choice of orientation for the optic axis in this section. 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 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
\[
P = \varepsilon_0 \chi E \quad (5.56)
\]
where
\[
\mathbf{E} = \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \quad \mathbf{P} = \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} \quad \chi = \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{xy} & \chi_{yy} & \chi_{yz} \\ \chi_{xz} & \chi_{yz} & \chi_{zz} \end{bmatrix} \quad (5.57)
\]
Our task is to find a new coordinate system \( x', y', z' \) for which the susceptibility tensor is diagonal. That is, we want to choose \( x', y', z' \) such that
\[
P' = \varepsilon_0 \chi' E' \quad (5.58)
\]
where

\[
\begin{align*}
E' &\equiv \begin{bmatrix} E'_x' \\ E'_y' \\ E'_z' \end{bmatrix} \\
P' &\equiv \begin{bmatrix} P'_x' \\ P'_y' \\ P'_z' \end{bmatrix} \\
\chi' &\equiv \begin{bmatrix} \chi'_{x'x'} & 0 & 0 \\ 0 & \chi'_{y'y'} & 0 \\ 0 & 0 & \chi'_{z'z'} \end{bmatrix}
\end{align*}
\]

(5.59)

To arrive at the new coordinate system, we are free to make pure rotation transformations. From (4.32), a rotation through an angle \( \gamma \) about the \( z \)-axis, followed by a rotation through an angle \( \beta \) about the resulting \( y \)-axis, and finally a rotation through an angle \( \alpha \) about the new \( x \)-axis, can be written as

\[
R = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & \sin \alpha \\ 0 & -\sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \cos \beta & 0 & \sin \beta \\ 0 & 1 & 0 \\ -\sin \beta & 0 & \cos \beta \end{bmatrix} \begin{bmatrix} \cos \gamma & \sin \gamma & 0 \\ -\sin \gamma & \cos \gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}
\]

(5.60)

The matrix \( R \) produces an arbitrary rotation of coordinates in three dimensions. Specifically, we can write:

\[
\begin{align*}
E' &= RE \\
P' &= RP
\end{align*}
\]

(5.61)

These transformations can be inverted to give

\[
\begin{align*}
E &= R^{-1}E' \\
P &= R^{-1}P'
\end{align*}
\]

(5.62)

where

\[
R^{-1} = \begin{bmatrix} \cos \beta \cos \gamma & -\cos \alpha \sin \gamma - \sin \alpha \sin \beta \cos \gamma & \sin \alpha \sin \gamma - \cos \alpha \sin \beta \cos \gamma \\ \cos \beta \sin \gamma & \cos \alpha \cos \gamma - \sin \alpha \sin \beta \sin \gamma & \sin \alpha \cos \gamma - \cos \alpha \sin \beta \sin \gamma \\ \sin \beta & \sin \alpha \cos \beta & \cos \alpha \cos \beta \end{bmatrix}
\]

\[
= \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{12} & R_{22} & R_{23} \\ R_{13} & R_{23} & R_{33} \end{bmatrix} = R^T
\]

(5.63)

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 (5.62) into (5.56) we have

\[
R^{-1}P' = \epsilon \chi R^{-1}E'
\]

(5.64)

or

\[
P' = \epsilon \chi R R^{-1}E'
\]

(5.65)
From this equation we see that the new susceptibility tensor we seek for (5.58) is

\[
\chi' \equiv R \chi R^{-1} = \begin{bmatrix}
R_{11} & R_{12} & R_{13} \\
R_{21} & R_{22} & R_{23} \\
R_{31} & R_{32} & R_{33}
\end{bmatrix}
\begin{bmatrix}
\chi_{xx} & \chi_{xy} & \chi_{xz} \\
\chi_{xy} & \chi_{yy} & \chi_{yz} \\
\chi_{xz} & \chi_{yz} & \chi_{zz}
\end{bmatrix}
\begin{bmatrix}
R_{11} & R_{21} & R_{31} \\
R_{22} & R_{22} & R_{32} \\
R_{33} & R_{33} & R_{33}
\end{bmatrix}
\]

(5.66)

We have expressly indicated that the off-diagonal terms of \(\chi'\) are symmetric (i.e. \(\chi'_{ij} = \chi'_{ji}\)). This can be verified by performing the multiplication in (5.66). It is a consequence of \(\chi\) being symmetric and \(R^{-1}\) being equal to \(R^T\).

The three off-diagonal elements of \(\chi'\) (appearing both above and below the diagonal) are found by performing the matrix multiplication in the second line of (5.66). The specific expressions for these three elements are not particularly enlightening. The important point is that we can make all three of them equal to zero since we have three degrees of freedom in the angles \(\alpha, \beta, \text{ and } \gamma\). Although, we do not expressly solve for the angles, we have demonstrated that it is always possible to set

\[
\begin{align*}
\chi'_{x'y'} &= 0 \\
\chi'_{x'z'} &= 0 \\
\chi'_{y'z'} &= 0
\end{align*}
\]

(5.67)

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 wavelet 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 (5.55) (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 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
Christiaan Huygens (1629–1695, Dutch) Huygens championed the wave theory of light. He was able to explain birefringence in terms of different indexes of refraction that vary 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.

After rearranging, the equation of the ellipse inside the crystal can also be written as

\[ z = \frac{c t}{n_o} \sqrt{1 - \frac{y^2}{(c t n_o)^2}} \]  

(5.69)

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 \( c t / \sin \theta_i \) above the original point. This distance is represented by the hypotenuse of the right triangle seen in Fig. 5.5. 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

\[ \frac{d z}{d y} = -\frac{z}{c t \sin \theta_i - z \tan \phi'} \]  

(5.70)

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

\[ \frac{d z}{d y} = -\frac{y n_o^2}{n_o c t \sqrt{1 - \frac{y^2}{(c t n_o)^2}}} = -\frac{n_o^2}{n_o^2 c t \sqrt{1 - \frac{y^2}{(c t n_o)^2}}} \]  

(5.71)

\[ = -\frac{n_o^2}{n_o^2} \tan \phi' \]
We would like these two slopes to be the same. We therefore set them equal to each other:

\[- \frac{n_e^2 \tan \phi'}{n_o^2} = - \frac{z}{ct \sin \theta_i - z \tan \phi'} \Rightarrow \frac{ct}{z} \frac{n_e^2 \tan \phi'}{n_o^2 \sin \theta_i} = \frac{n_e^2}{n_o^2} \tan^2 \phi' + 1 \quad (5.72)\]

If we evaluate (5.68) for the point \((y, z) = (z \tan \phi', z)\), we obtain

\[\frac{ct}{z} = n_0 \sqrt{\frac{n_e^2}{n_o^2} \tan^2 \phi' + 1} \quad (5.73)\]

Upon substitution of this into (5.72) we arrive at

\[\frac{n_e^2 \tan \phi'}{n_o^2 \sin \theta_i} = \sqrt{\frac{n_e^2}{n_o^2} \tan^2 \phi' + 1} \Rightarrow \frac{n_e^2}{n_o^2} \tan^2 \phi' = \frac{n_e^2}{n_o^2} \tan^2 \phi' + 1 \quad (5.74)\]

\[\Rightarrow \left[ \frac{n_e^2}{n_o^2} - 1 \right] \tan^2 \phi' = \frac{n_o^2}{n_e^2} \sin \theta_i \Rightarrow \tan \phi' = \frac{n_o \sin \theta_i}{n_e \sqrt{n_e^2 - \sin^2 \theta_i}} \quad (5.75)\]

This agrees with (5.55) as anticipated. Again, Huygens’ approach obtained the correct direction of the Poynting vector associated with the extraordinary wave.
Exercises

Exercises for 5.4 Fresnel’s Equation

P5.1 Solve Fresnel’s equation (5.25) to find the two values of \( n \) associated with a given \( \hat{u} \). Show that both solutions yield a positive index of refraction.

HINT: Show that (5.25) can be manipulated into the form

\[
0 = \left[ (u_x^2 + u_y^2 + u_z^2) - 1 \right] n^6 \\
+ \left[ \left( n_x^2 + n_y^2 + n_z^2 \right) - u_x^2 \left( n_y^2 + n_z^2 \right) \right] n^4 \\
- \left[ \left( n_x^2 n_y^2 + n_x^2 n_z^2 + n_y^2 n_z^2 \right) - u_x^2 n_y^2 n_z^2 - u_x^2 n_x^2 n_z^2 - u_x^2 n_x^2 n_y^2 \right] n^2 + n_x^2 n_y^2 n_z^2
\]

The coefficient of \( n^6 \) is identically zero since by definition we have \( u_x^2 + u_y^2 + u_z^2 = 1 \).

P5.2 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 \( \hat{u} = (\hat{x} + 2\hat{y} + 3\hat{z})/\sqrt{14} \) direction.

Exercises for 5.6 Biaxial and Uniaxial Crystals

P5.3 Show that (5.35) is a solution to (5.34).

P5.4 Given that the optic axes are in the \( x \)-\( z \) plane, show that the direction of the optic axes are given by (5.37).

HINT: The two indexes are the same when \( B^2 - 4AC = 0 \). You will want to use polar coordinates for the direction unit vector, as in (5.30). Set \( \phi = 0 \) so you are in the \( x \)-\( z \) plane. Use \( \sin^2 \theta + \cos^2 \theta = 1 \) to get an equation that only has cosine terms and solve for \( \cos^2 \theta \).

P5.5 Use definitions (5.38) and (5.39) along with the spherical representation of \( \hat{u} \) (5.30) in Fresnel’s equation (5.26) to calculate the two values for the index in a uniaxial crystal (i.e. (5.40) and (5.41)).

HINT: First show that

\[
A = n_o^2 \sin^2 \theta + n_e^2 \cos^2 \theta \\
B = n_o^2 n_e^2 + n_o^4 \sin^2 \theta + n_e^2 n_e^2 \cos^2 \theta \\
C = n_o^4 n_e^2
\]

and then use these expressions to evaluate Fresnel’s equation.
P5.6 Show that the field polarization component associated with \( n = n_o \) in a uniaxial crystal is directed perpendicular to the plane containing \( \hat{u} \) and \( \hat{z} \) by substituting this value for \( n \) into (5.34) and determining what combination of field components are allowable.

HINT: Use (5.30) to represent \( \hat{u} \) with \( \phi = 0 \) (the index is the same for all \( \phi \), so you may as well use one that makes calculation easy). When you substitute into (5.33) you will find that \( E_y \) can be any value because of the location of zeros in the matrix.

To get a requirement on \( E_x \) and \( E_z \), collapse the matrix equation down to a \( 2 \times 2 \) system. For non-trivial solutions to exist (i.e. \( E_x \neq 0 \) or \( E_y \neq 0 \)), the determinant of the matrix must be zero. Show that this is only the case if \( n_o = n_e \) (i.e. the crystal is isotropic).

P5.7 Show that the electric field for extraordinary polarized light \( E_e(\hat{u}) \) in a uniaxial crystal is not perpendicular to \( \mathbf{k} \) (i.e. \( \hat{u} \)), but that it is perpendicular to the ordinary polarization component \( E_o(\hat{u}) \).

P5.8 Derive (5.46).

P5.9 A quartz plate (uniaxial crystal with the optic axis perpendicular to the surfaces) has thickness \( d = 0.96 \) mm. The indices of refraction are \( n_o = 1.54424 \) and \( n_e = 1.55335 \). A plane wave with wavelength \( \lambda_{vac} = 633 \) nm passes through the plate. After emerging from the crystal, there is a phase difference \( \Delta \) between the two polarization components of the plane wave, and this phase difference depends on incident angle \( \theta_i \). Use a computer to plot \( \Delta \) as a function of incident angle from zero to 90°.

![Figure 5.6 Diagram for P 5.9.](image-url)
HINT: For \( s \)-polarized light, show that the number of wavelengths that fit in the plate is \( \frac{d}{(\lambda_{\text{vac}}/n_s) \cos \phi_s} \). For \( p \)-polarized light, show that the number of wavelengths that fit in the plate and the extra leg \( \delta \) outside of the plate (see Fig. 5.6) is \( \frac{d}{(\lambda_{\text{vac}}/n_p) \cos \phi_p} + \frac{\delta}{\lambda_{\text{vac}}} \), where \( \delta = d |\tan \phi_s - \tan \phi_p| \sin \theta_i \) and \( n_p \) is given by (5.41). Find the difference between these expressions and multiply by \( 2\pi \) to find \( \Delta \).

**L5.10**

In the laboratory, send a HeNe laser (\( \lambda_{\text{vac}} = 633 \text{ nm} \)) through two crossed polarizers, oriented at 45° and 135°. Place the quartz plate described in P 5.9 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 5.7](image)

**Figure 5.7** Schematic for L 5.10.

If the phase shift between the two paths is an odd integer times \( \pi \), 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 \( \pi \), 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 5.8](image)

**Figure 5.8** Plot for P 5.9 and L 5.10.
# 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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R1</strong></td>
<td>T or F: The optical index of <em>any</em> material (not vacuum) varies with frequency.</td>
</tr>
<tr>
<td><strong>R2</strong></td>
<td>T or F: The frequency of light can change as it enters a <em>crystal</em> (consider low intensity—no nonlinear effects).</td>
</tr>
<tr>
<td><strong>R3</strong></td>
<td>T or F: The entire expression $E_0 e^{i(kr−ωt)}$ associated with a light field (both the real part and the imaginary parts) is physically relevant.</td>
</tr>
<tr>
<td><strong>R4</strong></td>
<td>T or F: The <em>real</em> part of the refractive index cannot be less than one.</td>
</tr>
<tr>
<td><strong>R5</strong></td>
<td>T or F: <em>s</em>-polarized light and <em>p</em>-polarized light experience the <em>same</em> phase shift upon reflection from a material with complex index.</td>
</tr>
<tr>
<td><strong>R6</strong></td>
<td>T or F: When light is incident upon a material interface at Brewster’s angle, only one polarization can <em>transmit</em>.</td>
</tr>
<tr>
<td><strong>R7</strong></td>
<td>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 <em>direction of the reflected</em> $\mathbf{k}$-vector.</td>
</tr>
<tr>
<td><strong>R8</strong></td>
<td>T or F: The critical angle for total internal reflection exists on both sides of a material interface.</td>
</tr>
<tr>
<td><strong>R9</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>R10</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>R11</strong></td>
<td>T or F: An evanescent wave travels <em>parallel</em> to the surface interface on the <em>transmitted</em> side.</td>
</tr>
</tbody>
</table>
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 $J_{\text{free}} = 0$ and $P = \varepsilon_0 \chi E$. Note: $\nabla \times (\nabla \times \mathbf{f}) = \nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f}$.

(c) Show by direct substitution that $E(r, t) = E_0 e^{i(k \cdot r - \omega t)}$ is a solution to the wave equation. Find the resulting connection between $k$ and $\omega$. Give appropriate definitions for $c$ and $n$, assuming that $\chi$ is real.

(d) If $k = k \hat{z}$ and $E_0 = E_0 \hat{x}$, find the associated $B$-field.

(e) The Poynting vector is $S = E \times B / \mu_0$, 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.
(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 \( \pi \), 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 \( \theta \). What is the ratio of the intensity exiting the polarizer to the incoming intensity as a function of \( \theta \)?

Figure 5.10 Polarizing Elements

R19  
Consider an interface between two isotropic media where the incident field is defined by

\[
E_i = \left[ E_i^{(p)}(\hat{y} \cos \theta_i - \hat{z} \sin \theta_i) + \hat{x} E_i^{(s)} \right] e^{i[k(y \sin \theta_i + z \cos \theta_i - \omega_i t)]}
\]
The plane of incidence is shown in Fig. 5.11

(a) By inspection of the figure, write down similar expressions for the reflected and transmitted fields (i.e. \( E_r \) and \( E_t \)).

(b) Find an expression relating \( E_i \), \( E_r \), and \( E_t \) 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 \( \mathbf{B} \) must be continuous leads to

\[
\begin{align*}
    n_i(E_i^{(p)} - E_i^{(p)}) &= n_tE_i^{(p)} \\
    n_i(E_i^{(s)} - E_i^{(s)}) \cos \theta_i &= n_tE_i^{(s)} \cos \theta_t
\end{align*}
\]

Use this and the results from part (b) to derive

\[
    r_p \equiv \frac{E_r^{(p)}}{E_i^{(p)}} = -\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}
\]

You may use the identity

\[
\frac{\sin(\theta_i \cos \theta_i) - \sin(\theta_l \cos \theta_l)}{\sin(\theta_i \cos \theta_i) + \sin(\theta_l \cos \theta_l)} = \frac{\tan(\theta_i - \theta_l)}{\tan(\theta_i + \theta_l)}
\]

---

**Figure 5.11**

The Fresnel equations are

\[
\begin{align*}
    r_s &\equiv \frac{E_r^{(s)}}{E_i^{(s)}} = \frac{\sin \theta_i \cos \theta_i - \sin \theta_l \cos \theta_t}{\sin \theta_i \cos \theta_i + \sin \theta_l \cos \theta_t} \\
    t_s &\equiv \frac{E_t^{(s)}}{E_i^{(s)}} = \frac{2 \sin \theta_i \cos \theta_i}{\sin \theta_i \cos \theta_i + \sin \theta_l \cos \theta_t}
\end{align*}
\]
(a) Find what each of these equations reduces to when $\theta_i = 0$. Give your answer in terms of $n_i$ and $n_t$.

(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 5.12

(a) Derive and calculate Brewster's angle $\theta_B$. You may use the results of R19 (c).

(b) Calculate $\phi$.

(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 $\theta$ 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.
(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.

Second harmonic generation (the conversion of light with frequency $\omega$ into light with frequency $2\omega$) 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 downstream joins in phase with the harmonic light produced upstream, 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 $\phi$ 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 $\phi$ 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 $\theta_0$ 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 ~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.

### 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 \( E_{\text{s}}^{(t)} \) and \( E_{\text{p}}^{(t)} \) in terms of the incident fields \( E_{\text{s}}^{(i)} \) and \( E_{\text{p}}^{(i)} \). Similarly, we can also find the reflected fields \( E_{\text{s}}^{(r)} \) and \( E_{\text{p}}^{(r)} \) in terms the incident fields \( E_{\text{s}}^{(i)} \) and \( E_{\text{p}}^{(i)} \).

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 into the page.

**Figure 6.1** Waves propagating through a dual interface between materials.
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 $E_{1-0}^s$ and $E_{1-0}^p$ or by $E_{0-1}^s$ and $E_{0-1}^p$, separated into their $s$ or $p$-components, as usual. Similarly, $E_{1-0}^e$ and $E_{0-1}^e$ as well as $E_{2-0}^e$ and $E_{2-0}^e$ 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

\[ r_s^{0\rightarrow 1} = \frac{\sin \theta_1 \cos \theta_0 - \sin \theta_0 \cos \theta_1}{\sin \theta_1 \cos \theta_0 + \sin \theta_0 \cos \theta_1} \]
\[ t_s^{0\rightarrow 1} = \frac{2 \sin \theta_0 \cos \theta_1}{\sin \theta_1 \cos \theta_0 + \sin \theta_0 \cos \theta_1} \]
\[ r_p^{0\rightarrow 1} = \frac{\cos \theta_1 \sin \theta_1 - \cos \theta_0 \sin \theta_0}{\cos \theta_1 \sin \theta_1 + \cos \theta_0 \sin \theta_0} \]
\[ t_p^{0\rightarrow 1} = \frac{2 \cos \theta_0 \sin \theta_1}{\cos \theta_1 \sin \theta_1 + \cos \theta_0 \sin \theta_0} \]

The notation $0 \to 1$ 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

\[ r_s^{1\rightarrow 0} = -r_s^{0\rightarrow 1} \]
\[ t_s^{1\rightarrow 0} = \frac{2 \sin \theta_0 \cos \theta_1}{\sin \theta_0 \cos \theta_1 + \sin \theta_1 \cos \theta_0} \]
\[ r_p^{1\rightarrow 0} = -r_p^{0\rightarrow 1} \]
\[ t_p^{1\rightarrow 0} = \frac{2 \cos \theta_1 \sin \theta_0}{\cos \theta_0 \sin \theta_0 + \cos \theta_1 \sin \theta_1} \]

The notation $1 \to 0$ 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

\[
\begin{align*}
    r_s^{1\to2} & = \frac{\sin \theta_2 \cos \theta_1 - \sin \theta_1 \cos \theta_2}{\sin \theta_2 \cos \theta_1 + \sin \theta_1 \cos \theta_2} \\
    t_s^{1\to2} & = \frac{2 \sin \theta_2 \cos \theta_1}{\sin \theta_2 \cos \theta_1 + \sin \theta_1 \cos \theta_2} \\
    r_p^{1\to2} & = \frac{\cos \theta_2 \sin \theta_1 - \cos \theta_1 \sin \theta_2}{\cos \theta_2 \sin \theta_1 + \cos \theta_1 \sin \theta_2} \\
    t_p^{1\to2} & = \frac{2 \cos \theta_1 \sin \theta_2}{\cos \theta_2 \sin \theta_1 + \cos \theta_1 \sin \theta_2}
\end{align*}
\]

(6.3)

The notation $1 \to 2$ 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 $E_{1s}^{(i)}$ 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

\[
E_{2s}^{(i)} = t_s^{1\to2} E_{1s}^{(i)} e^{i k_1 d \cos \theta_1}
\]

where we have adjusted the phase of the field in (6.4) by $k_1 \cdot r = k_1 d \cos \theta_1$.

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 $E_{1s}^{(o)}$ stand for the transmitted field at the point $(y, z) = (0, d)$; its phase is built into its definition. The factor $t_s^{1\to2}$ 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

\[
E_{1s}^{(o)} e^{-i k_1 d \cos \theta_1} = t_s^{1\to2} E_{1s}^{(i)} e^{i k_1 d \cos \theta_1}
\]

(6.5)

Here again we have chosen to let $E_{2s}^{(o)}$ represent a plane wave field referenced to the origin $(y, z) = (0, 0)$. Therefore, the factor $e^{-i k_1 d \cos \theta_1}$ is needed at $(y, z) = (0, d)$.
(i.e. \( r = 2d \)) since the \( k \)-vector for the reverse-traveling field in the middle region is \( k_{1^-} = k_1 (\hat{y} \sin \theta_1 - \hat{z} \cos \theta_1) \).

We next connect the two plane waves in the middle region with the incident plane wave. In this case we must simultaneously connect \( E_{1^-}^{(s)} \) with both \( E_{0^-}^{(s)} \) and \( E_{1^=}^{(s)} \) since they each give a contribution:

\[
E_{1^-}^{(s)} = t_{s}^{0\to 1} E_{0^-}^{(s)} + r_{s}^{1\to 0} E_{1^=}^{(s)}
\]  

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 \( E_{0^-}^{(s)} \), 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 \( E_{1^-}^{(s)} \) and \( E_{1^=}^{(s)} \) from the expressions. Equation (6.4) can be inverted as follows:

\[
E_{1^-}^{(s)} = \frac{E_{2^-}^{(s)}}{t_{s}^{1\to 2} e^{i k_1 d \cos \theta_1}}
\]

When this is substituted into (6.5), we obtain

\[
E_{1^-}^{(s)} = \frac{r_{s}^{1\to 2} e^{i k_1 d \cos \theta_1}}{t_{s}^{1\to 2}} E_{2^-}^{(s)}
\]

Substitution of (6.7) and (6.8) into (6.6) yields

\[
\frac{E_{2^-}^{(s)}}{t_{s}^{1\to 2} e^{i k_1 d \cos \theta_1}} = t_{s}^{0\to 1} E_{0^-}^{(s)} + r_{s}^{1\to 0} \frac{r_{s}^{1\to 2} e^{i k_1 d \cos \theta_1}}{t_{s}^{1\to 2}} E_{2^-}^{(s)}
\]

This can be simplified to

\[
\frac{E_{2^-}^{(s)}}{E_{0^-}^{(s)}} = \frac{t_{s}^{0\to 1} t_{s}^{1\to 2}}{e^{-i k_1 d \cos \theta_1} - r_{s}^{1\to 0} r_{s}^{1\to 2} e^{i k_1 d \cos \theta_1}}
\]

where the factor

\[
k_1 d \cos \theta_1 = \frac{2 \pi n_1 d \cos \theta_1}{\lambda_{\text{vac}}}
\]

represents the phase acquired by either plane wave in traversing the middle region (see (2.25) and (2.27)).

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

\[
\tau_{s}^{\text{tot}} = \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \left| \frac{E_{2^-}^{(s)}}{E_{0^-}^{(s)}} \right|^2 (\theta_2 \text{ real}).
\]
Of course the relationship
\[ T_s^{\text{tot}} + R_s^{\text{tot}} = 1 \]  \hspace{1cm} (6.13)
still applies, but it is convenient for us to compute \( T_s^{\text{tot}} \) directly through (6.12) instead of indirectly from \( R_s^{\text{tot}} \).

When the transmitted angle \( \theta_2 \) is real, we may write the fraction of the transmitted power as
\[
T_s^{\text{tot}} = \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \left| \frac{e^{-ik_1 d \cos \theta_1} - r_s^{1-0} e^{ik_1 d \cos \theta_1}}{e^{ik_1 d \cos \theta_1} - r_s^{1-0} e^{-ik_1 d \cos \theta_1}} \right|^2 \hspace{1cm} (\theta_2 \text{ real}) \hspace{1cm} (6.14)
\]
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 \( \theta_1 \) is complex. Thus, it can be applied to the case of evanescent waves “tunneling” through a gap where \( \theta_0 \) 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 \( \theta_1 \) is complex, the angle \( \theta_2 \) 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 \( \theta_1 \) 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:

\[
T_s^{\text{tot}} = \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \left| \frac{t_s^{0-1} r_s^{1-2}}{t_s^{0-1} r_s^{1-2}} \right|^2 = \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \left( 1 + \left| r_s^{1-0} \right|^2 \left| r_s^{1-2} \right|^2 - 2 \text{Re} \left\{ r_s^{1-0} r_s^{1-2} e^{2ik_1 d \cos \theta_1} \right\} \right) \hspace{1cm} (6.15)
\]

On the last line we have introduced the definitions
\[ \delta \equiv 2k_1 d \cos \theta_1 \]  \hspace{1cm} (6.16)
and
\[ \delta_{r_s} \equiv \delta_{r_s}^{1-0} + \delta_{r_s}^{1-2} \]  \hspace{1cm} (6.17)

The phase terms \( \delta_{r_s}^{1-0} \) and \( \delta_{r_s}^{1-2} \) are defined indirectly and may be extracted from the relationships
\[
r_s^{1-0} = \left| r_s^{1-0} \right| e^{i \delta_{r_s}^{1-0}} \]  \hspace{1cm} (6.18)
and
\[
r_s^{1-2} = \left| r_s^{1-2} \right| e^{i \delta_{r_s}^{1-2}} \]  \hspace{1cm} (6.19)
We can continue our simplification of (6.15) by using the following identity:

\[
\cos \Phi = 1 - 2 \sin^2 \frac{\Phi}{2}
\]  

(6.20)

where \( \Phi \equiv \delta + \delta r_s \). With this, (6.15) can be written as

\[
T_{\text{tot}}^s = \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \frac{\left|r_s^{0-1}\right|^2 \left|t_s^{1-2}\right|^2}{1 + \left|r_s^{0-1}\right|^2 - 2 \left|r_s^{0-1}\right| \left|r_s^{1-2}\right| \left[1 - 2 \sin^2 \left(\frac{\Phi}{2}\right)\right]}
\]

\[
= \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \frac{\left|r_s^{0-1}\right|^2 \left|t_s^{1-2}\right|^2}{1 - \left|r_s^{1-2}\right|^2}
\]

\[
= \frac{T_{s}^{\text{max}}}{1 + F_s \sin^2 \left(\frac{\Phi}{2}\right)} \quad (\theta_2 \text{ and } \theta_1 \text{ real})
\]

(6.21)

where

\[
T_{s}^{\text{max}} \equiv \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \left|r_s^{0-1}\right|^2 \left|t_s^{1-2}\right|^2
\]

(6.22)

\[
F_s \equiv \frac{4 \left|r_s^{1-0}\right| \left|r_s^{1-2}\right|}{\left(1 - \left|r_s^{1-0}\right|^2 \left|r_s^{1-2}\right|^2\right)^2}
\]

(6.23)

The quantity \( T_{s}^{\text{max}} \) 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 \( \lambda_{\text{vac}} \) (causing \( \Phi \) to vary).

The maximum transmittance \( T_{s}^{\text{max}} \) can be manipulated as follows:

\[
T_{s}^{\text{max}} = \frac{n_2 \cos \theta_2}{n_0 \cos \theta_0} \frac{\left|r_s^{0-1}\right|^2 \left|t_s^{1-2}\right|^2}{1 - \left|r_s^{0-1}\right|^2 \left|r_s^{1-2}\right|^2} = \frac{T_{s}^{0-1} T_{s}^{1-2}}{1 - \sqrt{R_{s}^{0-1} R_{s}^{1-2}}}
\]

(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:

\[
F_s = \frac{4 \sqrt{R_{s}^{1-0} R_{s}^{1-2}}}{1 - \sqrt{R_{s}^{1-0} R_{s}^{1-2}}}
\]

(6.25)

Please note that \( R_{s}^{1-0} = R_{s}^{0-1} \), 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 \( p \).
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.

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 \(\lambda_{\text{vac}} = 633\) 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

\[ n_0 = 1.5 \]
\[ n_1 = 1.38 \]
\[ n_2 = 1 \]
\[ \theta_2 = 45^\circ \]

\[ n_1 \sin \theta_1 = n_2 \sin \theta_2 \Rightarrow \theta_1 = \sin^{-1} \left( \frac{\sin 45^\circ}{1.38} \right) = 30.82^\circ \]

\[ r_s^{1-o} = -\frac{\sin (\theta_1 - \theta_2)}{\sin (\theta_1 + \theta_2)} = -\frac{\sin (30.82^\circ - 45^\circ)}{\sin (30.82^\circ + 45^\circ)} = 0.253 \]

\[ n_0 \sin \theta_0 = n_2 \sin \theta_2 \Rightarrow \theta_0 = \sin^{-1} \left( \frac{\sin 45^\circ}{1.5} \right) = 28.13^\circ \]

\[ r_s^{1-o} = -\frac{\sin (\theta_1 - \theta_0)}{\sin (\theta_1 + \theta_0)} = -\frac{\sin (30.82^\circ - 28.13^\circ)}{\sin (30.82^\circ + 28.13^\circ)} = -0.0549 \]

\[ R_s^{1-o} \equiv |r_s^{1-o}|^2 = 0.0030 \]
\[ R_s^{1-o} \equiv |0.253|^2 = 0.0640 \]
\[ T_s^{0-1} = T_s^{1-0} = 1 - R_s^{1-o} = 1 - 0.0030 = 0.997 \]
\[ T_s^{1-2} = 1 - R_s^{1-2} = 1 - 0.0640 = 0.936 \]
\[ \delta_{s_1} = \delta_{s_1}^{1-o} + \delta_{s_1}^{1-2} = \pi + 0 = \pi \]

\[ F = \frac{4 \sqrt{R_{s_1} R_{s_2}}}{(1 - \sqrt{R_{s_1} R_{s_2}})^2} = \frac{4 \sqrt{(0.0030)(0.0640)}}{(1 - \sqrt{(0.0030)(0.0640)})^2} = 0.0570 \]

\[ T_s^{\text{max}} = \frac{T_s^{0-1} T_s^{1-2}}{(1 - \sqrt{T_s^{0-1} T_s^{1-2}})^2} = \frac{(0.997)(0.936)}{(1 - \sqrt{(0.0030)(0.0640)})^2} = 0.960 \]

\[ T_s^{\text{tot}} = \frac{0.960}{1 + 0.0570 \sin^2 \left( \frac{\delta + \pi}{2} \right)} \]

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 \( \theta_1 \) 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 \( \theta_0 \) and \( \theta_2 \) 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 \( \theta_1 \) when calculating the single-boundary Fresnel coefficients, so we rewrite \( \sin \theta_1 \) using Snell’s law:

\[
\sin \theta_1 = \frac{n_0}{n_1} \sin \theta_0 = \frac{n_2}{n_1} \sin \theta_2
\]

(6.26)

and \( \cos \theta_1 \) as

\[
\cos \theta_1 = i \sqrt{\sin^2 \theta_1 - 1}
\]

(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 \( \lambda_{\text{vac}} \) 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 \( \theta_0 = \theta_2 = 45^\circ \). Neglect possible reflections from the exterior surfaces of the prisms.

![Figure 6.3 Frustrated total internal reflection in two prisms.](image)

**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
We also need (6.26) and (6.27) to handle the complex angles:

\[
|I_p^{0-1}|^2 = \frac{2 \cos \theta_0 \sin \theta_1}{\cos \theta_1 \sin \theta_1 + \cos \theta_0 \sin \theta_0} \left( \frac{2 \cos \theta_0 (n \sin \theta_0)}{i \sqrt{n^2 \sin^2 \theta_0 - 1}} (n \sin \theta_0) + \cos \theta_0 \sin \theta_0 \right) = 5.76
\]

(6.28)

\[
|I_p^{1-2}|^2 = \frac{2 \cos \theta_1 \sin \theta_2}{\cos \theta_2 \sin \theta_2 + \cos \theta_1 \sin \theta_1} \left( \frac{2 \left( i \sqrt{n^2 \sin^2 \theta_2 - 1} \right) \sin \theta_2}{\cos \theta_2 \sin \theta_2 + \left( i \sqrt{n^2 \sin^2 \theta_0 - 1} \right) (n \sin \theta_0)} \right)^2 = 0.64
\]

(6.29)

\[
I_p^{1-2} = I_p^{0-0} = -r_p^{0-1} = -\frac{\cos \theta_1 \sin \theta_1 - \cos \theta_0 \sin \theta_0}{\cos \theta_1 \sin \theta_1 + \cos \theta_0 \sin \theta_0}
\]

(6.30)

\[
= -\left( i \sqrt{n^2 \sin^2 \theta_0 - 1} \right) (n \sin \theta_0) - \cos \theta_0 \sin \theta_0
\]

\[
= e^{-i \cdot 1.287}
\]

We also need

\[
k_1 d \cos \theta_1 = \frac{2 \pi}{\lambda_{\text{vac}}} \frac{d \cos \theta_1}{ \lambda_{\text{vac}}}
\]

\[
= 2 \pi \left( i \sqrt{n^2 \sin^2 \theta_0 - 1} \right) \frac{d}{ \lambda_{\text{vac}}}
\]

(6.31)

\[
= i 2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right)
\]

Now we are ready to compute the net transmittance (6.14). Since \( \theta_0 = \theta_2 \) and \( n_0 = n_2 \), we have

\[
T_p^{\text{tot}} = \frac{|I_p^{0-1}|^2 |I_p^{1-2}|^2}{|e^{-i k_1 d \cos \theta_1} - r_p^{0-0} e^{i k_1 d \cos \theta_1}|^2}
\]

(5.76)(0.64)

\[
= e^{-i \left( 2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right) \right)} - e^{-i1.287} e^{-i1.287} e^{i \left( 2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right) \right)}
\]

\[
|2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right) - 2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right) + 2.574| e^{2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} - e^{-2.22 \left( \frac{d}{ \lambda_{\text{vac}}} \right) + 2.574}
\]

\[
= 3.69
\]

(6.32)

\[
e^{4.44 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} + e^{-4.44 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} - 2 \left( e^{2.574} + e^{-2.574} \right)
\]

\[
= 3.69
\]

\[
e^{4.44 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} + e^{-4.44 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} - 2 \cos(2.574)
\]

\[
= 3.69
\]

\[
e^{4.44 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} + e^{-4.44 \left( \frac{d}{ \lambda_{\text{vac}}} \right)} + 1.69
\]

© 2004-2009 Peatross and Ware
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: \( T_{\text{tot}}^{p}(d/\lambda_{\text{vac}} = 0) = 1 \). 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 \( (T_{\text{tot}}^{p}(d/\lambda_{\text{vac}} \gg 1) \approx 0) \).

### 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

\[
T_{\text{tot}} = \frac{T_{\text{max}}^{1}}{1 + F \sin^2 \left( \frac{\Phi}{2} \right)} \tag{6.33}
\]

In the case of identical interface on the incident and transmitted sides, the transmittance and reflectance coefficients are the same at each surface (i.e. \( T = T^{s} = T^{p} \) and \( R = R^{s} = R^{p} \)). In this case, the maximum transmittance and the finesse coefficient are

\[
T_{\text{max}}^{s} = \frac{T^2_{\text{tot}}}{(1 - R)^2} \tag{6.34}
\]

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 $T_{\text{tot}}$ as the parameter $\Phi$ is varied (see (6.16) and (6.20)). The parameter $\Phi$ can be varied by altering $d$, $\theta_1$, or $\lambda$ as prescribed by

$$\Phi = \frac{4\pi n_1 d}{\lambda_{\text{vac}}} \cos \theta_1 + \delta_r$$  \hspace{1cm} (6.36)

To increase the sensitivity of the instrument, it is desirable to have the transmittance $T_{\text{tot}}$ vary strongly when $\Phi$ is varied. By inspection of (6.33), we see that $T_{\text{tot}}$ varies strongest if the finesse coefficient $F$ is large. We achieve a large finesse coefficient by increasing the reflectance $R$.

The total transmittance $T_{\text{tot}}$ (6.33) through a Fabry-Perot instrument is depicted in Fig. 6.5 as a function of $\Phi$. The various curves correspond to different values of $F$.

Typical values of $\Phi$ can be extremely large. For example, suppose that the instrument is used at near-normal incidence (i.e. $\cos \theta_1 \approx 1$) with a wavelength of $\lambda_{\text{vac}} = 500$ nm and an interface separation of $d_0 = 1$ cm. From (6.36) the value of $\Phi$ (ignoring the constant phase term $\delta_r$) is approximately

$$\Phi_0 = \frac{4\pi (1 \text{ cm})}{500 \text{ nm}} = 80,000\pi$$  \hspace{1cm} (6.37)

As we vary $d$, $\lambda$, or $\theta_1$ by small amounts, we can easily cause $\Phi$ to change by $2\pi$ as depicted in Fig. 6.5. The figure shows small changes in $\Phi$ above a value $\Phi_0$, which represents a large multiple of $2\pi$.

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

$$R + T + A = 1 \quad (6.38)$$

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 $T_{\text{max}}$ through the instrument (see (6.34) and (6.35)). However, in Fig. 6.5, each curve is plotted in terms of its own $T_{\text{max}}$.

The reflection phase $\delta_r$ 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 $\delta_r$ (depending on both the complex index of the coating material and its thickness). Whatever the value of $\delta_r$, we only care that it is constant. Experimentally, we can always compensate for the $\delta_r$ 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 $\theta_1$ 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 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, $\Phi$ 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 $T_{\text{tot}}$ (6.33) as a function of angle
6.6 Setup of a Fabry-Perot Instrument

Figure 6.8 Transmittance as the separation $d$ is varied ($F = 100$). $d_o$ represents a large distance for which $\Phi$ is a multiple of $2\pi$.

Figure 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 instru-
ment varies with surface separation \( d \) and angle \( \theta_1 \). 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 \( \Phi \) (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 \( \Phi \) is a multiple of \( 2\pi \) for the wavelength \( \lambda_{\text{vac}} \). 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 \( \Phi \) 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 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 \( \Phi \) 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
\( \lambda_0 \). That is, we have

\[
\Phi_0 = \frac{4\pi n_1 d_0 \cos \theta_1}{\lambda_0} + \delta_r \tag{6.39}
\]

where \( \Phi_0 \) is an integer multiple of \( 2\pi \). Now consider what happens to \( \Phi \) as the
wavelength increases. At a new wavelength (all else remaining the same) we have

\[
\Phi = \frac{4\pi n_1 d_0 \cos \theta_1}{\lambda_0 + \Delta \lambda} + \delta_r \tag{6.40}
\]

The change in wavelength \( \Delta \lambda \) is usually very small compared to \( \lambda_0 \), so we can
represent the denominator with the first two terms of a Taylor-series expansion:

\[
\frac{1}{\lambda_0 + \Delta \lambda} = \frac{1}{\lambda_0} \left(1 + \frac{\Delta \lambda}{\lambda_0}\right) \approx \frac{1}{\lambda_0} - \frac{\Delta \lambda}{\lambda_0^2} \tag{6.41}
\]

Then (6.40) can be rewritten as

\[
\Phi_0 - \Phi = \frac{4\pi n_1 d_0 \cos \theta_1}{\lambda_0^2} \Delta \lambda \tag{6.42}
\]
6.7 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument

Figure 6.9 A diverging monochromatic beam traversing a Fabry-Perot etalon.

Figure 6.10 (a) Transmittance as the angle $\theta_1$ is varied. It is assumed that the distance $d$ is chosen such that $\Phi$ is a multiple of $2\pi$ 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 $\lambda_0$, and the dashed line plots the transmittance of a wavelength shorter than $\lambda_0$. Note that the fringes shift positions for different wavelengths.

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 $\Phi$ to decrease by $2\pi$, 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 $2\pi$. After rearranging, we get

$$\Delta\lambda_{\text{FSR}} = \frac{\lambda^2_{\text{vac}}}{2n_1d_0\cos\theta_1}$$

(6.43)

If the wavelength is $\lambda_{\text{vac}} = 500$ nm and the spacing is $d_0 = 1$ cm, the free spectral range is $\Delta\lambda_{\text{FSR}} = (500 \text{ nm})^2 / 2(1 \text{ cm}) = 0.033$ 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
6.7 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument

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.

is the largest change in wavelength permissible while avoiding confusion. To convert this wavelength difference $\Delta \lambda_{\text{PSR}}$ into a corresponding frequency difference, one differentiates $\omega = \frac{2\pi c}{\lambda_{\text{vac}}}$ to get

$$|\Delta \omega| = \frac{2\pi c \Delta \lambda}{\lambda_{\text{vac}}^2}$$

(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 $\Phi_0$ be a multiple of $2\pi$ so that a peak in transmittance occurs when $\Phi = \Phi_0$. In this case, we have from (6.33) that

$$T_{\text{tot}} = \frac{T_{\text{max}}}{1 + F \sin^2 \left( \frac{\Phi_0}{2} \right)} = T_{\text{max}}$$

(6.45)

If $\Phi$ varies from $\Phi_0$ to $\Phi_0 \pm \Phi_{\text{FWHM}}/2$, then, by definition, the transmittance drops to one half. Therefore, we may write

$$T_{\text{tot}} = \frac{T_{\text{max}}}{1 + F \sin^2 \left( \frac{\Phi_0 \pm \Phi_{\text{FWHM}}}{2} \right)} = \frac{T_{\text{max}}}{2}$$

(6.46)

We solve (6.46) for $\Phi_{\text{FWHM}}$, and we see that this equation requires

$$F \sin^2 \left( \frac{\Phi_{\text{FWHM}}}{4} \right) = 1$$

(6.47)
where we have taken advantage of the fact that $\Phi_0$ is a multiple of $2\pi$. Next, we suppose that $\Phi_{\text{FWHM}}$ 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

$$\Phi_{\text{FWHM}} \approx \frac{4}{\sqrt{F}}.$$  \hfill (6.48)

The ratio of the period between peaks $2\pi$ to the width $\Phi_{\text{FWHM}}$ of individual peaks is called the reflecting finesse (or just finesse).

$$f \equiv \frac{2\pi}{\Phi_{\text{FWHM}}} = \frac{\pi \sqrt{F}}{2}$$ \hfill (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 $2\pi$ compared to $\Phi_{\text{FWHM}}$, or the reflecting finesse $f$. Therefore, we have

$$\Delta \lambda_{\text{FWHM}} = \frac{\Delta \lambda_{\text{FSR}}}{f} = \frac{\lambda_{\text{vac}}^2}{\pi n_1 d_0 \cos \theta_1 \sqrt{F}}$$ \hfill (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

$$\Delta \lambda_{\text{FWHM}} = \frac{(500 \text{ nm})^2}{\pi (1 \text{ cm}) \sqrt{100}} = 0.00080 \text{ nm}$$ \hfill (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 $\lambda_0$ to $\Delta \lambda_{\min}$, where $\Delta \lambda_{\min}$ is the minimum change of wavelength that the instrument can distinguish in the neighborhood of $\lambda_0$. (We are not as impressed when $\Delta \lambda_{\min}$ is small, if $\lambda_0$ also is small.) This ratio is called the resolving power of the instrument:

$$\text{RP} \equiv \frac{\lambda_0}{\Delta \lambda_{\min}}$$ \hfill (6.52)

In the case of a Fabry-Perot instrument, we have $\Delta \lambda_{\min} = \Delta \lambda_{\text{FWHM}}$. Fabry-Perot instruments tend to have very high resolving powers since they respond to very small differences in wavelength. When $\Delta \lambda_{\min} = 0.00080$ nm and $\lambda_0 = 500$ nm, the resolving power of a Fabry-Perot instrument is an impressive $\text{RP} = 600,000$. 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-deriv 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:

$$n_0 \sin \theta_0 = n_1 \sin \theta_1 = \cdots = n_N \sin \theta_N = n_{N+1} \sin \theta_{N+1}$$  \hspace{1cm} (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.

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 arrows pointing right indicate plane wave fields in individual layers that travel roughly in the forward (incident) direction, and the arrows pointing left indicate plane wave fields that travel roughly in the backward (reflected) direction. In the final region, there is only one plane wave traveling with a forward direction ($E_{N+1}^{p}$) which gives 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

$$\cos \theta_0 \left( E_{0}^{p} + E_{0}^{p} \right) = \cos \theta_1 \left( E_{1}^{p} + E_{1}^{p} \right)$$  \hspace{1cm} (6.54)
Figure 6.14 Light propagation through multiple layers.

and

$$ n_0 (E_{00}^{(p)} - E_{00}^{(p)}) = n_1 (E_{11}^{(p)} - E_{11}^{(p)}) $$ (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

$$ \cos \theta_j \left( E^{(p)}_{j+1,1} e^{i k_j \ell_j \cos \theta_j} + E^{(p)}_{j+1,2} e^{-i k_j \ell_j \cos \theta_j} \right) = \cos \theta_{j+1} \left( E^{(p)}_{j+1,1} + E^{(p)}_{j+1,2} \right) $$ (6.56)

and

$$ n_j \left( E^{(p)}_{j+1,1} e^{i k_j \ell_j \cos \theta_j} - E^{(p)}_{j+1,2} e^{-i k_j \ell_j \cos \theta_j} \right) = n_{j+1} \left( E^{(p)}_{j+1,1} - E^{(p)}_{j+1,2} \right) $$ (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 $k_j \ell_j \cos \theta_j$. This corresponds to the phase acquired by the plane wave field in traversing the layer with thickness $\ell_j$. The right-hand sides of (6.56) and (6.57) need no phase adjustment since the $(j+1)$th field is evaluated on the left side of its layer.

At the final interface, the boundary conditions reduce to

$$ \cos \theta_N \left( E^{(p)}_{N,1} e^{i k_N \ell_N \cos \theta_N} + E^{(p)}_{N,2} e^{-i k_N \ell_N \cos \theta_N} \right) = \cos \theta_{N+1} E^{(p)}_{N+1,1} $$ (6.58)

and

$$ n_N \left( E^{(p)}_{N,1} e^{i k_N \ell_N \cos \theta_N} - E^{(p)}_{N,2} e^{-i k_N \ell_N \cos \theta_N} \right) = n_{N+1} E^{(p)}_{N+1,1} $$ (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 $E^{(p)}_{N+1,1} \equiv 0$.

At this point we are ready to solve (6.54)–(6.59). We would like to eliminate all fields besides $E^{(p)}_{00,1}$, $E^{(p)}_{00,2}$, and $E^{(p)}_{N+1,1}$. 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:

\[
\begin{pmatrix}
\cos \theta_j e^{i \beta_j} & \cos \theta_j e^{-i \beta_j} \\
n_j e^{i \beta_j} & -n_j e^{-i \beta_j}
\end{pmatrix}
\begin{pmatrix}
E_{\text{j}+1}^p \\
E_{\text{j}+1}^s
\end{pmatrix} =
\begin{pmatrix}
\cos \theta_{j+1} & \cos \theta_{j+1} \\
n_{j+1} & -n_{j+1}
\end{pmatrix}
\begin{pmatrix}
E_{\text{j}+1}^p \\
E_{\text{j}+1}^s
\end{pmatrix}
\tag{6.60}
\]

where

\[
\beta_j = \begin{cases} 
0 & j = 0 \\
k_j \ell_j \cos \theta_j & 1 \leq j \leq N 
\end{cases}
\tag{6.61}
\]

and

\[
E_{N+1\rightarrow}^p \equiv E_{N+1\rightarrow}^s 
\tag{6.62}
\]

Then we solve (6.60) for the incident fields as follows:

\[
\begin{pmatrix}
E_{\text{j}+1}^p \\
E_{\text{j}+1}^s
\end{pmatrix} =
\begin{pmatrix}
\cos \theta_0 & \cos \theta_0 \\
n_0 & -n_0
\end{pmatrix}^{-1}
\begin{pmatrix}
\cos \theta_{j+1} & \cos \theta_{j+1} \\
n_{j+1} & -n_{j+1}
\end{pmatrix}
\begin{pmatrix}
E_{\text{j}+1}^p \\
E_{\text{j}+1}^s
\end{pmatrix}
\tag{6.63}
\]

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

\[
\begin{pmatrix}
E_{\text{0\rightarrow}}^p \\
E_{\text{0\rightarrow}}^s
\end{pmatrix} =
\begin{pmatrix}
\cos \theta_0 & \cos \theta_0 \\
n_0 & -n_0
\end{pmatrix}^{-1}
\begin{pmatrix}
\cos \theta_1 & \cos \theta_1 \\
n_1 & -n_1
\end{pmatrix}
\begin{pmatrix}
E_{\text{1\rightarrow}}^p \\
E_{\text{1\rightarrow}}^s
\end{pmatrix}
\tag{6.64}
\]

where we have grouped the matrices related to the \( j = 1 \) layer together via

\[
M_{1\rightarrow}^p \equiv \begin{pmatrix}
\cos \theta_1 & \cos \theta_1 \\
n_1 & -n_1
\end{pmatrix}
\begin{pmatrix}
\cos \theta_1 e^{i \beta_1} & \cos \theta_1 e^{-i \beta_1} \\
n_1 e^{i \beta_1} & -n_1 e^{-i \beta_1}
\end{pmatrix}^{-1}
\tag{6.65}
\]

By repeating this procedure for all \( N \) layers, we connect the fields in the initial
medium with the final medium as follows:

\[
\begin{pmatrix}
E_{\text{N\rightarrow}}^p \\
E_{\text{N\rightarrow}}^s
\end{pmatrix} =
\begin{pmatrix}
\cos \theta_0 & \cos \theta_0 \\
n_0 & -n_0
\end{pmatrix}^{-1}
\left( \prod_{j=1}^{N} M_j^p \right)
\begin{pmatrix}
\cos \theta_{N+1} & \cos \theta_{N+1} \\
n_{N+1} & -n_{N+1}
\end{pmatrix}
\begin{pmatrix}
E_{\text{N\rightarrow}}^p \\
E_{\text{N\rightarrow}}^s
\end{pmatrix}
\tag{6.66}
\]

where the matrices related to the \( j^{th} \) layer are grouped together according to

\[
M_j^p \equiv \begin{pmatrix}
\cos \theta_j & \cos \theta_j \\
n_j & -n_j
\end{pmatrix}
\begin{pmatrix}
\cos \theta_j e^{i \beta_j} & \cos \theta_j e^{-i \beta_j} \\
n_j e^{i \beta_j} & -n_j e^{-i \beta_j}
\end{pmatrix}^{-1}
\tag{6.67}
\]
The matrix inversion in the first line was performed using (0.46). The symbol $\Pi$ signifies the product of the matrices with the lowest subscripts on the left:

$$\prod_{j=1}^{N} M_j^{(p)} \equiv M_1^{(p)} M_2^{(p)} \cdots M_N^{(p)}$$  \hspace{1cm} (6.68)

As a finishing touch, we divide (6.64) by the incident field $E_{0^+}^{(p)}$ and perform the matrix inversion using (0.46) to obtain

$$\begin{bmatrix} 1 \\ E_{0^+}^{(p)} / E_{0^-}^{(p)} \end{bmatrix} = A^{(p)} \begin{bmatrix} E_{N+1^-}^{(p)} / E_{0^-}^{(p)} \\ 0 \end{bmatrix}$$  \hspace{1cm} (6.69)

where

$$A^{(p)} \equiv \begin{bmatrix} a_{11}^{(p)} & a_{12}^{(p)} \\ a_{21}^{(p)} & a_{22}^{(p)} \end{bmatrix} = \frac{1}{2n_0 \cos \theta_0} \begin{bmatrix} n_0 \cos \theta_0 & -n_0 \cos \theta_0 \\ n_0 \cos \theta_0 & 1 \end{bmatrix} \left( \prod_{j=1}^{N} M_j^{(p)} \right) \begin{bmatrix} \cos \theta_{N+1} & 0 \\ 0 & n_{N+1} \end{bmatrix}$$  \hspace{1cm} (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 permissible 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 $E_{0^+}^{(p)} / E_{0^-}^{(p)}$ and $E_{N+1^-}^{(p)} / E_{0^-}^{(p)}$. Once the matrix $A^{(p)}$ is computed, this is a relatively simple task:

$$t_p \equiv \frac{E_{N+1^-}^{(p)}}{E_{0^-}^{(p)}} = \frac{1}{a_{11}^{(p)}} \hspace{1cm} \text{(Multilayer)}$$  \hspace{1cm} (6.71)

$$r_p \equiv \frac{E_{0^+}^{(p)}}{E_{0^-}^{(p)}} = \frac{a_{21}^{(p)}}{a_{11}^{(p)}} \hspace{1cm} \text{(Multilayer)}$$  \hspace{1cm} (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 $2 \times 2$ 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

$$\begin{bmatrix} 1 \\ E_{0^+}^{(s)} / E_{0^-}^{(s)} \end{bmatrix} = A^{(s)} \begin{bmatrix} E_{N+1^-}^{(s)} / E_{0^-}^{(s)} \\ 0 \end{bmatrix}$$  \hspace{1cm} (6.73)

where

$$A^{(s)} \equiv \begin{bmatrix} a_{11}^{(s)} & a_{12}^{(s)} \\ a_{21}^{(s)} & a_{22}^{(s)} \end{bmatrix} = \frac{1}{2n_0 \cos \theta_0} \begin{bmatrix} n_0 \cos \theta_0 & -n_0 \cos \theta_0 \\ n_0 \cos \theta_0 & 1 \end{bmatrix} \left( \prod_{j=1}^{N} M_j^{(s)} \right) \begin{bmatrix} 1 & 0 \\ n_{N+1} \cos \theta_{N+1} & 0 \end{bmatrix}$$  \hspace{1cm} (6.74)
and

\[
M_j^{(p)} = \begin{bmatrix}
\cos \beta_j & -i \sin \beta_j / (n_j \cos \theta_j) \\
-i n_j \cos \theta_j \sin \beta_j & \cos \beta_j
\end{bmatrix} \tag{6.75}
\]

We can then compute the transmission and reflection coefficients in the same manner that we found the \(p\)-components

\[
t_s \equiv \frac{E_s^{(s)}}{E_{0s}} = \frac{1}{d_{1j}} \tag{Multilayer} \tag{6.76}
\]

\[
r_s \equiv \frac{E_{0s}}{E_{0s}} = \frac{a_{2j}^{(s)}}{a_{1j}^{(s)}} \tag{Multilayer} \tag{6.77}
\]

### 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

\[
\beta_j = \frac{\pi}{2} \quad \text{(high reflector)} \tag{6.78}
\]

This amounts to the condition on the thickness of

\[
\ell_j = \frac{\lambda_{\text{vac}}}{4 n_j \cos \theta_j} \quad \text{(high reflector)} \tag{6.79}
\]

Since the layers alternate high and low indices, at every other boundary there is a phase shift of \(\pi\) 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} \quad \text{(high reflector, \(p\)-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

\[
\begin{bmatrix}
0 & -i \cos \theta_H / n_H \\
-i n_L / \cos \theta_L & 0
\end{bmatrix} \begin{bmatrix}
0 & -i \cos \theta_L / n_L \\
-i n_H / \cos \theta_H & 0
\end{bmatrix} = \begin{bmatrix}
-n_L \cos \theta_H / n_H \cos \theta_L & 0 \\
0 & -n_L \cos \theta_H / n_H \cos \theta_L
\end{bmatrix} \tag{6.81}
\]

To extend to \(q = N/2\) layer pairs, we have

\[
\prod_{j=1}^{N} M_j^{(p)} = \begin{bmatrix}
-n_L \cos \theta_H / n_H \cos \theta_L & 0 \\
0 & -n_L \cos \theta_H / n_H \cos \theta_L
\end{bmatrix}^q = \begin{bmatrix}
\left( \frac{n_L \cos \theta_H}{n_H \cos \theta_L} \right)^q & 0 \\
0 & \left( \frac{n_L \cos \theta_H}{n_H \cos \theta_L} \right)^q
\end{bmatrix} \tag{6.82}
\]
and using (6.82) we can compute $A^{[p]}$

$$A^{[p]} = \frac{1}{2} \begin{bmatrix} \left( \frac{n_H \cos \theta_H}{n_L \cos \theta_L} \right)^q \cos \theta_{N+1} & \cos \theta_0 & 0 \\ \left( \frac{n_H \cos \theta_H}{n_L \cos \theta_L} \right)^q \cos \theta_{N+1} & \cos \theta_0 & 0 \\ \left( \frac{n_H \cos \theta_H}{n_L \cos \theta_L} \right)^q \cos \theta_{N+1} & \cos \theta_0 & 0 \end{bmatrix}$$

(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 0.5) comes in handy:

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

(6.84)

where

$$\cos \theta \equiv \frac{1}{2} (A + D) .$$

(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**

**Exercises for 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).

**Exercises for 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

\[
T_{\text{tot}} = \frac{1}{1 + \frac{(n^2 - 1)^2}{4n^2} \sin^2 \left(\frac{2\pi nd}{\lambda_{\text{vac}}}\right)}
\]

HINT: Find

\[
r_1^{1-2} = r_1^{1-0} = -r_0^{0-1} = \frac{n - 1}{n + 1}
\]

and then use

\[
T_{1-2} = 1 - R_0^{0-1}
\]

\[
T_{1-2} = 1 - R_{1-2}
\]

(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 6.1. 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 \(\lambda_{\text{vac}} = 633\ nm\).
Exercises for 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 (\( \lambda_{\text{vac}} = 3 \text{ cm} \)) 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](image)

(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). Figure 6.16 shows a plot of some typical data taken with this setup. Presumably experimental error causes some discrepancy, but the trend is clear.

![Figure 6.16](image)

Exercises for 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 \( d = 0.5 \text{ cm} \) 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 $\theta_1$ is varied. Let $T_{\text{max}} = 1$, $F = 10$, $\lambda = 500$ nm, $d = 1$ cm, and $n_1 = 1$.

(a) Plot $T$ vs. $\theta_1$ over the angular range used in Fig. 6.10(a).

(c) Suppose $d$ was slightly different, say 1.00001 cm. Make a plot of $T$ vs $\theta_1$ 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 0th fringe corresponding to the center). HINT: $\cos \theta \approx 1 - \theta^2/2$. 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](image1)

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](image2)
Exercises for 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.

Exercises for 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 \( n_H = 2.32 \) and \( n_L = 1.38 \), deposited on a glass substrate with index \( n = 1.5 \). Let the wavelength be \( \lambda_{\text{vac}} = 633 \) 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 \( \lambda = 550 \text{ nm} \) (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 \( \lambda \), but more importantly, the thicknesses used in the coatings will not be \( \lambda/4 \) for other wavelengths. Let’s ignore the index change with \( \lambda \) and focus on the wavelength dependence. Use the matrix techniques and a computer to plot \( R(\lambda) \) 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:

\[ E(r, t) = \sum_j E_j e^{i(k_j \cdot r - \omega_j t)} \] (7.1)

The corresponding magnetic field (see (2.54)) is

\[ B(r, t) = \sum_j B_j e^{i(k_j \cdot r - \omega_j t)} = \sum_j \frac{k_j \times E_j}{\omega_j} e^{i(k_j \cdot r - \omega_j t)} \] (7.2)

In section 7.2, we show that the intensity of this overall field under certain assumptions can be expressed as

\[ I(r, t) = \frac{n \varepsilon_0 c^2}{2} E(r, t) \cdot E^*(\mathbf{r}, t) \] (7.3)

where \( E(\mathbf{r}, t) \) represents the entire complex expression for the electric field rather than just the real part. Although this expression is reminiscent of (2.60), 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 \( k_j \) 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 \( k_j \)'s. We can relax the restriction of parallel \( k_j \)'s slightly and apply (7.3) also to plane waves with nearly parallel \( k_j \)'s such as occurs in a Young's two-slit diffraction experiment (studied in the next chapter).
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 \( \omega \). We discuss superpositions of plane waves in terms of Fourier theory. (For an introductory overview of Fourier transforms, see section 0.4.) 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 (2.50) is

\[
S(r, t) = \frac{\text{Re}[E(r, t)] \times \text{Re}[B(r, t)]}{\mu_0}
\]  

(7.4)

Upon substitution of (7.1) and (7.2) into the above expression, we obtain

\[
S(r, t) = \sum_{j,m} \frac{1}{\omega_m \mu_0} \left[ k_m \left\{ \text{Re}\left\{ E_j e^{i(k_j \cdot r - \omega_j t)} \right\} \cdot \text{Re}\left\{ E_m e^{i(k_m \cdot r - \omega_m t)} \right\} \right\} 
- \text{Re}\left\{ E_m e^{i(k_m \cdot r - \omega_m t)} \right\} \left\{ \text{Re}\left\{ E_j e^{i(k_j \cdot r - \omega_j t)} \right\} \cdot k_m \right\}
\]

(7.5)

The last term in (7.6) can be dismissed if all of the \( k_m \) are perpendicular to each other. This can only be ensured if all \( k \)-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(r, t) = \varepsilon_0 c \sum_{j,m} n_m \text{Re}\left\{ E_j e^{i(k_j \cdot r - \omega_j t)} \right\} \cdot \text{Re}\left\{ E_m e^{i(k_m \cdot r - \omega_m t)} \right\} \quad \text{parallel \( k \)-vectors}
\]

(7.7)

where in accordance with (1.51) and (2.22) we have introduced

\[
\frac{k_m}{\omega_m \mu_0} = n_m \varepsilon_0 c.
\]

(7.8)

Here \( n_m \) refers to the refractive index associated with the frequency \( \omega_m \). 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 (0.30))

\[
S(r, t) = n \varepsilon_0 c \sum_{j,m} \frac{\left[ E_j e^{i(k_j \cdot r - \omega_j t)} + E_j^* e^{-i(k_j \cdot r - \omega_j t)} \right]}{2} \cdot \frac{\left[ E_m e^{i(k_m \cdot r - \omega_m t)} + E_m^* e^{-i(k_m \cdot r - \omega_m t)} \right]}{2}
= \frac{n \varepsilon_0 c}{4} \sum_{j,m} \left[ E_j \cdot E_m e^{i[(k_j + k_m) \cdot r - (\omega_j + \omega_m) t]} + E_j^* \cdot E_m^* e^{-i[(k_j + k_m) \cdot r - (\omega_j + \omega_m) t]} \right]
+ \left[ E_j \cdot E_m^* e^{i[(k_j - k_m) \cdot r - (\omega_j - \omega_m) t]} + E_j^* \cdot E_m e^{-i[(k_j - k_m) \cdot r - (\omega_j - \omega_m) t]} \right]
\]

(parallel \( k \)-vectors, constant \( n \))

(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

$$\langle S(r, t) \rangle_{osc.} = \frac{n\epsilon_0 c}{2} \sum_{j,m} E_j \cdot E_m^* e^{i[(k_j - k_m) \cdot r - (\omega_j - \omega_m)t]} + E_j^* \cdot E_m e^{-i[(k_j - k_m) \cdot r - (\omega_j - \omega_m)t]}$$

$$= \frac{n\epsilon_0 c}{2} \text{Re} \left\{ \sum_j E_j e^{i(k_j \cdot r - \omega_j t)} \cdot \sum_m E_m^* e^{-i(k_m \cdot r - \omega_m t)} \right\}$$

$$= \frac{n\epsilon_0 c}{2} \text{Re} \left[ E(r, t) \cdot E^*(r, t) \right].$$

(parallel $k$-vectors, constant $n$, time-averaged over rapid oscillations)

(7.10)

In writing the final line we have again invoked (7.1). Notice that the expression $E(r, t) \cdot E^*(r, t)$ 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 $k_m$ to get (7.7) we assumed that all $k_m$ are nearly parallel to each other. If some of the $k_m$ 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

$$I(r, t) \cong \frac{n\epsilon_0 c}{2} E(r, t) \cdot E^*(r, t)$$

(parallel or antiparallel $k$-vectors, constant $n$)

(7.11)

where $E(r, t)$ 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:

$$E(r, t) = E_0 e^{i(k_1 \cdot r - \omega_1 t)} + E_0 e^{i(k_2 \cdot r - \omega_2 t)}$$

(7.12)
As we previously studied (see P 1.10), the velocities of the individual wave crests are

\[ v_{p1} = \omega_1 / k_1 \]
\[ v_{p2} = \omega_2 / k_2 \]  \hspace{1cm} (7.13)

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:

\[
I(r, t) = \frac{n \varepsilon_0 c^2}{2} \mathbf{E}_0 \cdot \mathbf{E}_0^* \left[ e^{i(k_1 \cdot r - \omega_1 t)} + e^{i(k_2 \cdot r - \omega_2 t)} \right] \left[ e^{-i(k_1 \cdot r - \omega_1 t)} + e^{-i(k_2 \cdot r - \omega_2 t)} \right] \\
= \frac{n \varepsilon_0 c^2}{2} \mathbf{E}_0 \cdot \mathbf{E}_0^* \left[ 2 + e^{i(k_2 - k_1) \cdot (r - (\omega_2 - \omega_1) t)} \right] \\
= n \varepsilon_0 c \mathbf{E}_0 \cdot \mathbf{E}_0^* \left[ 1 + \cos[(k_2 - k_1) \cdot r - (\omega_2 - \omega_1) t]\right] \\
= n \varepsilon_0 c \mathbf{E}_0 \cdot \mathbf{E}_0^* \left[ 1 + \cos(\Delta k \cdot r - \Delta \omega t)\right] \\ 
\]  \hspace{1cm} (7.14)

where

\[ \Delta k = k_2 - k_1 \]
\[ \Delta \omega = \omega_2 - \omega_1 \]  \hspace{1cm} (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

\[ v_g = \frac{\Delta \omega}{\Delta k} \]  \hspace{1cm} (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 2.7) for which the index is real over a range of frequencies. The index of refraction is given by

\[ n_{\text{plasma}}(\omega) = \sqrt{1 - \omega_p^2 / \omega^2} < 1 \] (assuming \( \omega > \omega_p \))  \hspace{1cm} (7.17)
The phase velocity for each frequency is computed by

\[
v_{p1} = \frac{c}{n_{\text{plasma}}(\omega_1)}
\]
\[
v_{p2} = \frac{c}{n_{\text{plasma}}(\omega_2)}
\]

(7.18)

Since \( n_{\text{plasma}} < 1 \), both of these velocities exceed \( c \). However, the group velocity is

\[
v_g = \frac{\Delta \omega}{\Delta k} \equiv \frac{d \omega}{dk} \equiv \left( \frac{dk}{d\omega} \right)^{-1} = \left[ \frac{d \omega n_{\text{plasma}}(\omega)}{d\omega} \frac{1}{c} \right]^{-1} = n_{\text{plasma}}(\omega) \frac{c}{\omega}
\]

(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 \( \omega_1 \) and \( \omega_2 \) 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, 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

\[
E(\mathbf{r}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(\mathbf{r}, \omega) e^{-i\omega t} d\omega 
\] (7.20)

The function \( E(\mathbf{r}, \omega) \) 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 \( \exp \{ i\mathbf{k}(\omega) \cdot \mathbf{r} \} \). The function \( E(\mathbf{r}, \omega) \) is distinguished from the function \( E(\mathbf{r}, t) \) by its argument (i.e. \( \omega \) instead of \( t \)). The factor \( 1/\sqrt{2\pi} \) is introduced to match our Fourier transform convention.

Given knowledge of \( E(\mathbf{r}, \omega) \), the waveform \( E(\mathbf{r}, t) \) can be constructed. Similarly, if the waveform \( E(\mathbf{r}, t) \) is known, the field per frequency can be obtained via

\[
E(\mathbf{r}, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(\mathbf{r}, t) e^{i\omega t} dt 
\] (7.21)

This operation, which produces \( E(\mathbf{r}, \omega) \) from \( E(\mathbf{r}, t) \), is called a Fourier transform. The operation (7.20) is called the inverse Fourier transform. For a review of Fourier theory, see section 0.4.

Even though \( E(\mathbf{r}, t) \) can be written as a real function (since, after all, only the real part is relevant), \( E(\mathbf{r}, \omega) \) is in general complex. The real and imaginary parts of \( E(\mathbf{r}, \omega) \) keep track of how much cosine and how much sine, respectively, make up \( E(\mathbf{r}, t) \). Keep in mind that both positive and negative frequency components go into the cosine and sine according to (0.19). Therefore, it should not seem strange that we integrate (7.20) over all frequencies, both positive and negative. If \( E(\mathbf{r}, t) \) is taken to be a real function, then we have the symmetry relation

\[
E(\mathbf{r}, -\omega) = E^* (\mathbf{r}, \omega) \quad \text{(if } E(\mathbf{r}, t) \text{ is real)}
\] (7.22)

However, often \( E(\mathbf{r}, t) \) is written in complex notation, where taking the real part is implied. For example, the real waveform

\[
E_r (\mathbf{r}, t) = E_{0r} (\mathbf{r}) e^{-t^2/2\tau^2} \cos (\omega_0 t - \phi)
\] (7.23)
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 $\tau = 5T$.

is usually written as

$$E_c(r, t) = E_{0c}(r) e^{-t^2 / 2\tau^2} e^{-i\omega_0 t}$$  \hspace{1cm} (7.24)

where $E_c(r, t) = \text{Re}(E_c(r, t))$. The phase $\phi$ is hidden within the complex amplitude $E_{0c}(r)$, 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 $\tau$.

Consider the Fourier transforms of the waveform (7.23). Upon applying (7.21)
we get (see P 0.27)

\[ E_r (r, \omega) = \tau E_0 r (r) e^{i\phi} e^{-\frac{r^2 (\omega + \omega_0)^2}{2}} + e^{i\phi} e^{-\frac{r^2 (\omega - \omega_0)^2}{2}} \]  

(7.25)

Similarly, the Fourier transform of (7.24), i.e. the complex version of the same waveform, is

\[ E_c (r, \omega) = \tau E_{0c} r (r) e^{-\frac{r^2 (\omega - \omega_0)^2}{2}} \]  

(7.26)

The latter transform is less cumbersome to perform, and for this reason more often used.

Figure 7.3 shows graphs of \(|E_r (r, \omega)|^2\) associated with the waveforms in Fig. 7.2. Figure 7.4 shows graphs of \(E_c (r, \omega) \cdot E^*_c (r, \omega)/2\) 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 \(\tau\)) 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 \(\tau\) 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 \(\omega_0\), 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 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).
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 $\tau = 5T$. 
7.4 Frequency Spectrum of Light

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 $\tau = 5T$. 
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:

\[
\int_{-\infty}^{\infty} I(r, t) \, dt = \int_{-\infty}^{\infty} I(r, \omega) \, d\omega
\]  

(7.27)

where

\[
I(r, t) \equiv \frac{n \epsilon_0 c}{2} E(r, t) \cdot E^*(r, t)
\]

\[
I(r, \omega) \equiv \frac{n \epsilon_0 c}{2} E(r, \omega) \cdot E^*(r, \omega)
\]  

(7.28)

The power spectrum \( I(r, \omega) \) 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): \( I(r, \omega) \) is not the Fourier transform of \( I(r, t) \)!

## 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 (2.23)). A phase shift associated with a displacement \( \Delta r \) modifies the field according to

\[
E(r_0 + \Delta r, \omega) = E(r_0, \omega) e^{i k(\omega) \cdot \Delta r}
\]  

(7.29)

The \( k \)-vector contains the pertinent information about the material via \( k = n(\omega) \omega / c \). (A complex wave vector \( k \) 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 \( E(r_0, t) \) to find the plane-wave coefficients \( E(r_0, \omega) \) at the beginning of propagation. Apply the phase adjustment in (7.29) to find the plane wave coefficients \( E(r_0 + \Delta r, \omega) \) at the end of propagation. Then take the inverse Fourier transform to determine the
waveform \( E(\mathbf{r}_0 + \Delta \mathbf{r}, t) \) at the new position:

\[
E(\mathbf{r}_0 + \Delta \mathbf{r}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(\mathbf{r}_0 + \Delta \mathbf{r}, \omega) e^{-i\omega t} \, d\omega \\
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(\mathbf{r}_0, \omega) e^{i\mathbf{k}(\omega) \cdot \Delta \mathbf{r}} \cdot e^{-i\omega t} \, d\omega
\]  

(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 \( \bar{\omega} \):

\[
\mathbf{k} \cdot \Delta \mathbf{r} \equiv \left[ \mathbf{k}_{\bar{\omega}} + \frac{\partial \mathbf{k}}{\partial \omega} \bigg|_{\bar{\omega}} (\omega - \bar{\omega}) + \frac{1}{2} \frac{\partial^2 \mathbf{k}}{\partial \omega^2} \bigg|_{\bar{\omega}} (\omega - \bar{\omega})^2 + \cdots \right] \cdot \Delta \mathbf{r}
\]  

(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

\[
E(\mathbf{r}_0 + \Delta \mathbf{r}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(\mathbf{r}_0, \omega) e^{i \left[ \left( \mathbf{k}_{\bar{\omega}} + \frac{\partial \mathbf{k}}{\partial \omega} \bigg|_{\bar{\omega}} (\omega - \bar{\omega}) \right) \cdot \Delta \mathbf{r} \right]} \cdot e^{-i\omega t} \, d\omega \\
= e^{i \left[ \mathbf{k}_{\bar{\omega}} \cdot \Delta \mathbf{r} - \bar{\omega} t' \right]} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(\mathbf{r}_0, \omega) e^{i\omega t - i\omega t'} \, d\omega
\]  

(7.32)

where in the last line we have used the definition

\[
t' \equiv \frac{\partial \mathbf{k}}{\partial \omega} \bigg|_{\bar{\omega}} \cdot \Delta \mathbf{r}.
\]  

(7.33)

If we assume that the imaginary part of \( \mathbf{k} \) is constant near \( \bar{\omega} \) so that \( t' \) is real, i.e.

\[
t' = \frac{\partial \text{Re} \mathbf{k}}{\partial \omega} \bigg|_{\bar{\omega}} \cdot \Delta \mathbf{r}
\]  

(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

\[
E(\mathbf{r}_0 + \Delta \mathbf{r}, t) = e^{i \left[ \mathbf{k}_{\bar{\omega}} \cdot \Delta \mathbf{r} - \bar{\omega} t' \right]} E(\mathbf{r}_0, t - t')
\]  

(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)):

\[
v_p^{-1}(\bar{\omega}) = \frac{k(\bar{\omega})}{\bar{\omega}}
\]  

(7.36)
To compare the intensity profile of the pulse at \( r_0 + \Delta r \) with the profile at \( r_0 \) we compute the square magnitude of (7.35)

\[
I(r_0 + \Delta r, t) \propto |E(r_0, t - t')|^2 e^{-2 \text{Im}(\omega) \cdot \Delta r}.
\]  

(7.37)

In (7.37) we see that (to first order) \( t' \) is the time required for the pulse to traverse the displacement \( \Delta r \). 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 \( \partial \text{Re} k / \partial \omega \cdot \Delta r \) is known as the group delay function, and in (7.34) it is evaluated only at the carrier frequency \( \omega \). Traditional group velocity is obtained by dividing the displacement \( \Delta r \) by the group delay time \( t' \) to obtain

\[
v^{-1}_g(\omega) = \left. \frac{\partial \text{Re}(k(\omega))}{\partial \omega} \right|_{\omega} \]

(7.38)

Group delay (or group velocity) essentially tracks the center of the packet.

In our derivation we have assumed that the phase delay \( k(\omega) \cdot \Delta r \) 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 \( \omega \) 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 \( E(r, \omega) \) does not change, but merely its phase according to (7.29). In other words, the plane-wave components that 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 \( r_0 \) at the start of the glass at \( z = 0 \) and assume that all plane-wave components travel in the \( \hat{z} \)-direction, so that \( k \cdot \Delta r = k z \). 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

$$E(0, t) = E_0 e^{-\frac{t^2}{2\tau^2}} e^{-i\omega_0 t}$$

$$E(0, \omega) = \tau E_0 e^{-\frac{\omega^2}{2\tau^2}} e^{-i\omega t}$$

(7.39)

To find the field downstream we invoke (7.29), which gives the appropriate phase shift for each plane wave component:

$$E(z, \omega) = E(0, \omega) e^{ik(\omega)z} = \tau E_0 e^{-\frac{\omega^2}{2\tau^2}} e^{ik(\omega)z}$$

(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 $k_{vac}(\omega) = \omega/c$. In this case, the final waveform is

$$E(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_0 \tau e^{-\frac{\omega^2}{2\tau^2}} e^{i\frac{\omega z}{c}} e^{-i\omega t} d\omega = E_0 \tau e^{-\frac{1}{2}(\frac{\omega^2}{\omega_0^2} + \frac{z^2}{c^2})} e^{i(k_0z - \omega_0 t)}$$

(vacuum)

(7.41)

where $k_0 \equiv \omega_0/c$. Not surprisingly, after traveling a distance $z$ through vacuum, the pulse looks identical to the original pulse, only its peak occurs at a later time $z/c$. The term $k_0z$ 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 $\hat{\omega} = \omega_0$, so the expansion is

$$k(\omega) z \approx k(\omega_0) z + \frac{\partial k}{\partial \omega}\bigg|_{\omega_0} (\omega - \omega_0) z + \frac{1}{2} \frac{\partial^2 k}{\partial \omega^2}\bigg|_{\omega_0} (\omega - \omega_0)^2 z + \cdots$$

$$\approx k_0 z + v_g^{-1} (\omega - \omega_0) z + \alpha (\omega - \omega_0)^2 z$$

(7.42)
where

\[ k_0 = k(\omega_0) = \frac{\omega_0 n(\omega_0)}{c} \]  \hspace{1cm} (7.43)

\[ \nu_k^{-1} = \left. \frac{\partial k}{\partial \omega} \right|_{\omega_0} = \frac{n(\omega_0)}{c} + \frac{\omega_0 n'(\omega_0)}{c} \]  \hspace{1cm} (7.44)

\[ \alpha \equiv \frac{1}{2} \left. \frac{\partial^2 k}{\partial \omega^2} \right|_{\omega_0} = \frac{n'(\omega_0)}{c} + \frac{\omega_0 n''(\omega_0)}{2c} \]  \hspace{1cm} (7.45)

With this approximation for \( k(\omega) \), we are now able to perform the inverse Fourier transform on (7.40):

\[
E(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_0 e^{i(k_0 z - \omega_0 t)} e^\left[ -\frac{(\omega - \omega_0)^2}{2} \right] e^{i\nu_k^{-1}(\omega - \omega_0)z + i\alpha(\omega - \omega_0)z^2} e^{-i\omega t} d\omega
\]

\[
= \frac{\tau E_0 e^{i(k_0 z - \omega_0 t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\left[ -\frac{(\omega - \omega')^2}{2} \right]} e^{i\nu_k^{-1}(\omega - \omega_0)z - i(t - \omega_0 t)} d\omega' \]  \hspace{1cm} (7.46)

We can avoid considerable clutter if we change variables to \( \omega' \equiv \omega - \omega_0 \). Then the inverse Fourier transform becomes

\[
E(z, t) = \frac{\tau E_0 e^{i(k_0 z - \omega_0 t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\left[ -\frac{(\omega - \omega')^2}{2} \right]} e^{i\nu_k^{-1}(\omega - \omega_0)z - i(t - \omega_0 t)} d\omega' \]  \hspace{1cm} (7.47)

The above integral can be performed with the aid of (0.52). The result is

\[
E(z, t) = \frac{\tau E_0 e^{i(k_0 z - \omega_0 t)}}{\sqrt{2\pi}} \sqrt{\frac{\pi}{\frac{\tau^2}{4} \left( 1 - i2\alpha z / \tau^2 \right)^2}} e^{-\frac{(z - \nu_k\tau)^2}{4\left( 1 - i2\alpha z / \tau^2 \right)^2}}
\]

\[
= E_0 e^{i(k_0 z - \omega_0 t)} \left[ e^{\frac{i}{\tau} \tan^{-1} \frac{2\alpha z}{\tau^2}} \right] e^{-\frac{(z - \nu_k\tau)^2}{4\left( 1 + (2\alpha z / \tau^2)^2 \right)}} \]  \hspace{1cm} (7.48)

Next, we spruce up the appearance of this rather cumbersome formula as follows:

\[
E(z, t) = \frac{E_0}{\sqrt{T(z)/\tau}} e^{-\frac{1}{\tau} \left[ \frac{2\alpha z}{\tau^2} \right]} e^{-\frac{i}{\tau} \left[ \frac{2\alpha z}{\tau^2} \right]} \Phi(z) + i(k_0 z - \omega_0 t) + i \frac{1}{2} \tan^{-1} \Phi(z) \]  \hspace{1cm} (7.49)

where

\[
\Phi(z) \equiv \frac{2\alpha}{\tau^2} z \]  \hspace{1cm} (7.50)

and

\[
T(z) \equiv \tau \sqrt{1 + \Phi^2(z)} \]  \hspace{1cm} (7.51)

We can immediately make a few observations 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

$$e^{-\frac{1}{2} \left( \frac{t-z/v_g}{T(z)} \right)^2}$$

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 $T(z)$ 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 $k(\omega)$ 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 $k(\omega) \cdot \Delta r$. 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 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:

$$\langle t \rangle_r \equiv \int_{-\infty}^{\infty} t \rho(r, t) \, dt$$  \hspace{1cm} (7.52)

Here $\rho(r, t)$ is a normalized distribution function associated with the intensity:

$$\rho(r, t) \equiv \frac{I(r, t)}{\int_{-\infty}^{\infty} I(r, t) \, dt}$$  \hspace{1cm} (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 $r_0$ to point $r = r_0 + \Delta r$ in a homoge-
neous medium (see Fig. 7.9). The difference in arrival times at the two points is

$$\Delta t \equiv \langle t \rangle_r - \langle t \rangle_{r_0}$$  \hspace{1cm} (7.54)

The pulse shape can evolve in complicated ways between the two points, spread-
ing with different portions being absorbed (or amplified) during transit. Never-
theless, (7.54) renders an unambiguous time interval between the passage of the
pulse center at each point.

Figure 7.8 Pulse undergoing distortion during transit.

Figure 7.9 Transit time defined as the difference between arrival time at two points.
This difference in arrival time can be shown to consist of two terms (see P 7.12):

\[ \Delta t = \Delta t_G (\mathbf{r}) + \Delta t_R (\mathbf{r}_0) \]  

(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 \( \mathbf{r} = \mathbf{r}_0 + \Delta \mathbf{r} \):

\[ \Delta t_G (\mathbf{r}) = \int_{-\infty}^{\infty} \rho (\mathbf{r}, \omega) \left( \frac{\partial \text{Re} \mathbf{k}}{\partial \omega} \cdot \Delta \mathbf{r} \right) \, d\omega \]  

(7.56)

where the spectral weighting function is

\[ \rho (\mathbf{r}, \omega) \equiv \frac{I (\mathbf{r}, \omega)}{\int_{-\infty}^{\infty} I (\mathbf{r}, \omega') \, d\omega'} \]  

(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.

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 \( \partial \text{Re} \mathbf{k} / \partial \omega \), 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 \( \partial \text{Re} \mathbf{k} / \partial \omega |_{\bar{\omega}} \cdot \Delta \mathbf{r} \), 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*:

\[ \Delta t_R (\mathbf{r}_0) = \langle t \rangle_{\mathbf{r}_0} \big|_{\text{altered}} - \langle t \rangle_{\mathbf{r}_0} \]  

(7.58)

Here \( \langle t \rangle_{\mathbf{r}_0} \big|_{\text{altered}} \) 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, \( \langle t \rangle_{\mathbf{r}_0} \) is the arrival time of a pulse associated with the modified field \( \mathbf{E} (\mathbf{r}_0, \omega) e^{-i \mathbf{k} \cdot \Delta \mathbf{r}} \). Notice that \( \mathbf{E} (\mathbf{r}_0, \omega) e^{-i \mathbf{k} \cdot \Delta \mathbf{r}} \) 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
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 $\gamma$ is the damping term in the Lorentz model described in section 2.3. Let the pulse travel a distance $\Delta r = 2c/(10\gamma)$ through the absorbing medium (as depicted in Fig. 7.10), which has a resonance at frequency $\omega_0$. The index of refraction is shown in Fig. 7.11. Its resonance has a width of $\gamma$. Fig. 7.12 shows the delay between the pulse arrival times at $r_0$ and $r = r_0 + \Delta r$ as the pulse’s central frequency $r = r_0 + \Delta r$ is varied in the neighborhood of the resonance. The solid line gives the total delay $\Delta t \approx \Delta t_G(r)$ experienced by the narrowband pulse in traversing the displacement. The reshaping delay in this case is negligible (i.e. $\Delta t_R(r) \approx 0$) 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 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 $\tau_2$. 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 superluminal 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 (2.49), the total energy density stored in an electromagnetic field and in a medium is given by

\[ u(r, t) = u_{\text{field}}(r, t) + u_{\text{exchange}}(r, t) + u(r, -\infty) \quad (7.59) \]

This expression for the energy density includes all (relevant) forms of energy, including a non-zero integration constant \( u(r, -\infty) \) corresponding to energy stored in the medium before the arrival of any pulse (important in the case of an amplifying medium). \( u_{\text{field}}(r, t) \) and \( u_{\text{exchange}}(r, t) \) are both zero before the arrival of the pulse (i.e. at \( t = -\infty \)). In addition, \( u_{\text{field}}(r, t) \), given by (2.51), 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

\[ u_{\text{exchange}}(r, t) = \int_{-\infty}^{t} E(r, t') \cdot \frac{\partial P(r, t')}{\partial t'} \, dt' \quad (7.60) \]

where we ignore the possibility of any free current \( J_{\text{free}} \) in (2.52). As \( u_{\text{exchange}} \) increases, the energy in the medium increases. Conversely, as \( u_{\text{exchange}} \) decreases, the medium surrenders energy to the electromagnetic field. While it is possible for \( u_{\text{exchange}} \) 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 (2.49) has the form of a continuity equation which when integrated spatially over a small volume \( V \) yields

\[ \oint_A S \cdot da = -\frac{\partial}{\partial t} \int_V u \, dV \quad (7.61) \]

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 \( S \) to be uniform throughout \( V \). The energy transport velocity (directed along \( S \)) 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 \( S \)). The energy transport velocity as traditionally written is then

\[ v_E \equiv \frac{S}{u} \quad (7.62) \]

© 2004-2009 Peatross and Ware
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 $u_{\text{field}}$ 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 $u_{\text{field}}$. 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

$$\langle v_E \rangle = \frac{\int v_E u \, d^3r}{\int u \, d^3r} = \frac{\int S \, d^3r}{\int u \, d^3r}$$

(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

$$\langle v_E \rangle = -\frac{\int \mathbf{r} \cdot \nabla \cdot S \, d^3r}{\int u \, d^3r} = \frac{\int \frac{\partial u}{\partial t} \, d^3r}{\int u \, d^3r}$$

(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 (2.49) can be written with no source terms (i.e. $\nabla \cdot S + \partial u / \partial t = 0$). 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

$$\langle v_E \rangle = \frac{\partial \langle r \rangle}{\partial t}$$

(7.65)

where

$$\langle r \rangle = \frac{\int \mathbf{r} u \, d^3r}{\int u \, d^3r}$$

(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 $u_{\text{field}}$ 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

\[
\left\langle \frac{S}{u_{\text{field}}} \right\rangle \neq \frac{\partial}{\partial t} \int u_{\text{field}} d^3 r
\]  

(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 \(u_{\text{exchange}}\) within a frequency-domain context. We utilize the field represented in terms of an inverse Fourier transform (7.20). Similarly, the polarization \(P\) can be written as an inverse Fourier transform:

\[
P(r, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} P(r, \omega) e^{-i\omega t} d\omega \Rightarrow \frac{\partial P(r, t)}{\partial t} = \frac{-i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega P(r, \omega) e^{-i\omega t} d\omega
\]  

(7.68)

In an isotropic medium, the polarization for an individual plane wave can be written in terms of the linear susceptibility defined in (1.46):

\[
P(r, \omega) = \epsilon_0 \chi(r, \omega) E(r, \omega)
\]  

(7.69)

With (7.21), (7.68), and (7.69), the exchange energy density (7.60), can be written as

\[
u_{\text{exchange}}(r, t) = \int_{-\infty}^{t} \left[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(r, \omega') e^{-i\omega' t'} d\omega' \right] \cdot \left[ -\frac{i\epsilon_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega \chi(r, \omega) E(r, \omega) e^{-i\omega t} d\omega \right] d t'
\]  

(7.70)

After interchanging the order of integration, the expression becomes

\[
u_{\text{exchange}}(r, t) = -i\epsilon_0 \int_{-\infty}^{\infty} d\omega \omega \chi(r, \omega) E(r, \omega) \cdot \int_{-\infty}^{\infty} d\omega' E(r, \omega') \frac{1}{2\pi} \int_{-\infty}^{t} e^{-i(\omega+\omega') t'} d t'
\]  

(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 \(r\) experiences the entire pulse, the final amount of energy density exchanged between the field and the medium at that point is

\[
u_{\text{exchange}}(r, +\infty) = -i\epsilon_0 \int_{-\infty}^{\infty} \omega \chi(r, \omega) E(r, \omega) \cdot E(r, -\omega) d\omega
\]  

(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
\[ P^* (r, \omega) = P (r, -\omega) \] (7.73)
and hence
\[ \chi^* (r, \omega) = \chi (r, -\omega). \] (7.74)
Then we obtain
\[ u_{\text{exchange}} (r, +\infty) = \varepsilon_0 \int_{-\infty}^{\infty} \omega \text{Im} \chi (r, \omega) E (r, \omega) \cdot E^* (r, \omega) \, d\omega \] (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):
\[ u_{\text{exchange}} (r, t) = \varepsilon_0 \int_{-\infty}^{\infty} \omega \text{Im} \chi (r, \omega) E_t (r, \omega) \cdot E_t^* (r, \omega) \, d\omega \] (7.76)
where
\[ E_t (r, \omega) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} E (r, t') e^{i\omega t'} \, dt' \] (7.77)
This time dependence enters only through \( E_t (r, \omega) \cdot E_t^* (r, \omega) \), 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 \( E_t(\omega) \) 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 \( \chi(r, \omega) \).

If the entire pulse passing through point \( r \) 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 \( r_0 \) 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 \( u_{\text{exchange}} \) as a function of time, where 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 \( u_{\text{exchange}} \)) 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.
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 $E_0^2 / \gamma^2$. Spectra are assessed at the times indicated in (a) and (c).
Exercises for 7.2 Intensity

P7.1 (a) Let \( \hat{x}E_1 e^{i(kz - \omega t)} \) and \( \hat{x}E_2 e^{i(-kz - \omega t)} \) be two counter-propagating plane waves where \( E_1 \) and \( E_2 \) are both real. Show that their sum can be written as

\[ \hat{x}E_{tot}(z) e^{i(\Phi(z) - \omega t)} \]

where

\[ E_{tot}(z) = E_1 \sqrt{\left(1 - \frac{E_2}{E_1}\right)^2 + 4 \frac{E_2}{E_1} \cos^2 k z} \]

and

\[ \Phi(z) = \tan^{-1} \left[ \frac{(1 - E_2/E_1)}{(1 + E_2/E_1)} \tan k z \right] \]

Outside the range \(-\frac{\pi}{2} \leq k z \leq \frac{\pi}{2}\) 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

\[ E(r, t) = \hat{x}E_0 e^{i[(k_2 \hat{x} - \omega_2 - \omega_1) \cdot r - \omega_2 t]} + \hat{y}E_0 e^{i[(k_1 \hat{y} - \omega_1) \cdot r - \omega_1 t]} \]

according to (7.11) is uniform throughout space. Of course (7.11) does not apply since the \( k \)-vectors are not parallel. Show that the time-average of \( S(r, t) \) according to (7.6) exhibits interference in the distribution of net energy flow.

Exercises for 7.3 Group vs. Phase Velocity: Sum of Two Plane Waves

P7.3 Show that (7.12) can be written as

\[ E(r, t) = 2E_0 e^{i\left[\frac{k_2 + k_1}{2} \cdot r - \frac{\omega_2 + \omega_1}{2} t\right]} \cos \left(\frac{\Delta k}{2} \cdot r - \frac{\Delta \omega}{2} t\right) \]

From this show that the speed at which the rapid-oscillation peaks move in Fig. 7.1 is

\[ \left( v_{p1} + v_{p2} \right) / 2 \]

P7.4 Confirm the right-hand side of (7.19).


**Exercises for 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: 
\[ E(\mathbf{r}, t) = E_0 e^{i(k_0 z - \omega_0 t)}. \]
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 \( \omega_0 \) (see (0.45)).

(b) Compute the power spectrum after the shutter if it is opened during the interval \(-\tau/2 \leq t \leq \tau/2\). 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 
\[ \text{sinc} \alpha \equiv \frac{\sin \alpha}{\alpha}. \]

**P7.6**

(a) Determine the Full-Width-at-Half-Maximum of the intensity (i.e. the width of \( I(\mathbf{r}, t) \) represented by \( \Delta t_{\text{FWHM}} \)) 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 \( \tau \).

(b) Give an uncertainty principle for the product of \( \Delta t_{\text{FWHM}} \) and \( \Delta \omega_{\text{FWHM}} \).

**P7.7** Verify (7.27) for the Gaussian pulse defined by (7.24) and (7.26).

---

**Exercises for 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 \( \omega_0 \) 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

\[ n(\omega) \equiv 1.4948 + 0.016 \frac{\omega}{\omega_0}. \]

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 \( T(z) \to \tau \Phi(z) \) and \( \tan^{-1} \Phi(z) \to \pi/2 \), show that the instantaneous frequency of the pulse is

\[ \omega_0 + \frac{t - 2z/v_g}{4az}. \]
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.

Exercises for 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

\[
\lim_{\tau \to \infty} \Delta t_G(r) = \frac{\partial \text{Re} k}{\partial \omega} \Delta \bar{\omega} \right| \omega
\]

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

\[
\lim_{\tau \to 0} \Delta t_G(r) = \frac{\Delta r}{c}
\]

assuming \( k \) and \( \Delta r \) are parallel. This implies 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 \( \kappa \) 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

\[
I_{\text{det}}(\tau) = c\epsilon_0 E_0^2 \left[ \left| E_0 \right|^2 + 2E_0 \cdot E_0^* \cos(\omega \tau) \right]
\]

(8.1)

where \( I_0 \equiv \frac{c\epsilon_0}{2} E_0 \cdot E_0^* \) 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
upons 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 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 \( E(t) \) that has traveled through the first arm of a Michelson interferometer to arrive at the detector in Fig. 8.1. Specifically, \( E(t) \) is the value of the field at the detector when the second arm of the interferometer is blocked. The waveform \( E(t) \) in general may be composed of many frequency components according to the inverse Fourier transform (7.20). For convenience we will think of \( E(t) \) 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 \( \tau \) according to the path difference between the two arms. Thus, \( E(t - \tau) \) indicates the field at the detector from the second arm when the first arm of the interferometer is blocked. Again, \( \tau \) 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:

\[
E_{\text{det}}(t, \tau) = E(t) + E(t - \tau)
\] (8.2)
With (7.28) we compute the intensity at the detector:

\[
I_{\text{det}}(t, \tau) = \frac{c\epsilon_0}{2} E_{\text{det}}(t, \tau) \cdot E^*_{\text{det}}(t, \tau)
\]

\[
= \frac{c\epsilon_0}{2} \left[ E(t) \cdot E^*(t) + E(t) \cdot E^*(t - \tau) + E(t - \tau) \cdot E^*(t) + E(t - \tau) \cdot E^*(t - \tau) \right]
\]

\[
= I(t) + I(t - \tau) + \frac{c\epsilon_0}{2} \left[ E(t) \cdot E^*(t - \tau) + E(t - \tau) \cdot E^*(t) \right]
\]

\[
= I(t) + I(t - \tau) + c\epsilon_0 \text{Re} \{ E(t) \cdot E^*(t - \tau) \}
\]

\( (8.3) \)

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 \( I_{\text{det}}(t, \tau) \) in addition to the dependence on the path delay \( \tau \). 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

\[
\int_{-\infty}^{\infty} I_{\text{det}}(t, \tau) \, dt = \int_{-\infty}^{\infty} I(t) \, dt + \int_{-\infty}^{\infty} I(t - \tau) \, dt + c\epsilon_0 \text{Re} \int_{-\infty}^{\infty} E(t) \cdot E^*(t - \tau) \, dt
\]

\( (8.4) \)

The final integral remains unchanged if we take a Fourier transform followed by an inverse Fourier transform:

\[
\int_{-\infty}^{\infty} E(t) \cdot E^*(t - \tau) \, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\omega e^{-i\omega t} \left[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\tau e^{i\omega \tau} \int_{-\infty}^{\infty} E(t) \cdot E^*(t - \tau) \, dt \right]
\]

\( (8.5) \)

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):

\[
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\tau e^{i\omega \tau} \int_{-\infty}^{\infty} E(t) \cdot E^*(t - \tau) \, dt = \sqrt{2\pi} E(\omega) \cdot E^*(\omega) = \sqrt{2\pi} \frac{2I(\omega)}{c\epsilon_0}
\]

\( (8.6) \)

We can apply Parseval’s theorem (see (7.27)) to the first two integrals on the right-hand side of (8.4):

\[
\int_{-\infty}^{\infty} I(t) \, dt = \int_{-\infty}^{\infty} I(t - \tau) \, dt = \int_{-\infty}^{\infty} I(\omega) \, d\omega
\]

\( (8.7) \)

Notice that the middle integral is insensitive to the delay \( \tau \) 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

\[ \int_{-\infty}^{\infty} I_{\text{det}}(t, \tau) \, dt = 2 \int_{-\infty}^{\infty} I(\omega) \, d\omega + 2 \text{Re} \int_{-\infty}^{\infty} I(\omega) e^{-i\omega \tau} \, d\omega \]

\[ = \left[ 2 \int_{-\infty}^{\infty} I(\omega) \, d\omega \right] \left[ 1 + \frac{\text{Re} \int I(\omega) e^{-i\omega \tau} \, d\omega}{\int_{-\infty}^{\infty} I(\omega) \, d\omega} \right] \tag{8.8} \]

It is convenient to rewrite this in terms of the Degree of Coherence function \( \gamma(\tau) \):

\[ \int_{-\infty}^{\infty} I_{\text{det}}(t, \tau) \, dt = \left[ 2 \int_{-\infty}^{\infty} I(t) \, dt \right] \left[ 1 + \text{Re} \gamma(\tau) \right] \tag{8.9} \]

where

\[ \gamma(\tau) \equiv \frac{\int_{-\infty}^{\infty} I(\omega) e^{-i\omega \tau} \, d\omega}{\int_{-\infty}^{\infty} I(\omega) \, d\omega} \tag{8.10} \]

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 \( \tau \) is entirely contained in the function \( \gamma(\tau) \).

### 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 \( \omega \) 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:

\[ \int_{-\infty}^{\infty} I_{\text{det}}(\omega, \tau) \, d\omega = \int_{-\infty}^{\infty} 2I(\omega) [1 + \cos(\omega \tau)] \, d\omega \tag{8.11} \]

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 \textit{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:

$$
\int_{-\infty}^{\infty} I_{\text{det}}(\omega, \tau) d\omega = \left[ 2 \int_{-\infty}^{\infty} I(\omega) d\omega \right] \left[ 1 + \frac{\int_{-\infty}^{\infty} I(\omega) \cos(\omega\tau) d\omega}{\int_{-\infty}^{\infty} I(\omega) d\omega} \right] \quad (8.12)
$$

This is the same as (8.8) since we can replace $\cos(\omega\tau)$ with $\Re\{e^{-i\omega\tau}\}$, 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:

$$
I_{\text{ave}} \equiv \langle I(t) \rangle_t = \frac{1}{T} \int_{-T/2}^{T/2} I(t) dt \quad \text{(continuous source)} \quad (8.13)
$$

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

$$
\langle I_{\text{det}}(t, \tau) \rangle_t = 2 \langle I(t) \rangle_t \left[ 1 + \Re\gamma(\tau) \right] \quad \text{(continuous source)} \quad (8.14)
$$

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 $E(\omega)$. It depends only on the amount of light associated with each frequency component through
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 $\text{Re} \gamma(\tau)$ is analogous to $\cos(\omega\tau)$ in (8.1). For large delays $\tau$, 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

$$\tau_c \equiv \int_{-\infty}^{\infty} |\gamma(\tau)|^2 d\tau = 2 \int_{0}^{\infty} |\gamma(\tau)|^2 d\tau$$ (8.15)

© 2004-2009 Peatross and Ware
8.4 Fringe Visibility and Coherence Length

The coherence length is the distance that light travels in this time:

$$\ell_c \equiv c \tau_c$$  \hspace{1cm} (8.16)

Another useful concept is fringe visibility. The fringe visibility is defined in the following way:

$$V(\tau) \equiv \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} \quad \text{(continuous)}$$  \hspace{1cm} (8.17)

or

$$V(\tau) \equiv \frac{E_{\text{max}} - E_{\text{min}}}{E_{\text{max}} + E_{\text{min}}} \quad \text{(pulsed)}$$  \hspace{1cm} (8.18)

where $$E_{\text{max}} \equiv \max \int_{-\infty}^{\infty} I_{\text{det}}(t, \tau) \, dt$$ 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)). $$E_{\text{min}}$$ 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 $$E_{\text{min}}$$ and $$E_{\text{max}}$$ 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

$$V(\tau) = |\gamma(\tau)|$$  \hspace{1cm} (8.19)

In the case of a Gaussian spectral distribution (7.26)

$$I(\omega) = I(\omega_0) e^{-\left(\frac{\omega - \omega_0}{\Delta \omega}\right)^2}$$  \hspace{1cm} (8.20)
the result of (8.10) is
\[ \gamma(\tau) = e^{-i\tau\omega_0 - \frac{(\Delta\omega)^2 \tau^2}{4}} \] (8.21)
Figure 8.2 plots the magnitude and real part of (8.21). From (8.15) the coherence time is
\[ \tau_c = \frac{\sqrt{2\pi}}{\Delta\omega} \] (8.22)
Figure 8.3 shows \(1 + \text{Re}\gamma(\tau)\), which is proportional to the energy (per area) arriving at the detector. As expected, the fringes die off for a delay interval of \(\tau_c\).

### 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
\[
\text{Sig}(\tau) \propto \int_{-\infty}^{\infty} I_{\text{det}}(t, \tau) \, dt = \left[ 2 \int_{-\infty}^{\infty} I(t) \, dt \right] \left[ 1 + \text{Re}\gamma(\tau) \right] \] (8.23)
where
\[
\gamma(\tau) \equiv \frac{\int_{-\infty}^{\infty} I(\omega)e^{-i\omega\tau} \, d\omega}{\int_{-\infty}^{\infty} I(\omega) \, d\omega} \] (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 \(\tau\) is large (i.e. fringe visibility goes to zero: \(\gamma(\tau) = 0\)). In this case, the normalized signal must approach
\[
\lim_{\tau \to \infty} \eta \text{Sig}(\tau) = 2\varepsilon_0 \] (8.25)
where \(\eta\) is the appropriate normalization constant that changes the proportionality (8.23) into an equation, and
\[
\varepsilon_0 = \int_{-\infty}^{\infty} I(t) \, dt = \int_{-\infty}^{\infty} I(\omega) \, d\omega \] (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 \(\text{Sig}(\tau)\), 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 \(\varepsilon_0\)) 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

\[
\eta \text{Sig}(\tau) = 2\mathcal{E}_0 + 2\text{Re} \int_{-\infty}^{\infty} I(\omega) e^{-i\omega\tau} d\omega \quad (8.27)
\]

Next, we take the Fourier transform of this equation:

\[
\mathcal{F}\{\eta \text{Sig}(\tau)\} = \mathcal{F}\{2\mathcal{E}_0\} + \mathcal{F}\left\{2\text{Re} \int_{-\infty}^{\infty} I(\omega) e^{-i\omega\tau} d\omega\right\} \quad (8.28)
\]

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:

\[
\mathcal{F}\{2\mathcal{E}_0\} = 2\mathcal{E}_0 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega\tau} d\tau = 2\mathcal{E}_0 \sqrt{2\pi} \delta(\omega) \quad (8.29)
\]

Notice that (8.29) is zero everywhere except where \( \omega = 0 \), where a spike occurs. This represents the DC component of \( \mathcal{F}\{\eta \text{Sig}(\tau)\} \).
The second term of (8.28) can be written as

\[
\begin{align*}
\mathcal{F}\{2\text{Re} \int_{-\infty}^{\infty} I(\omega) e^{-i\omega \tau} d\omega\} &= \mathcal{F}\left\{ \int_{-\infty}^{\infty} I(\omega) e^{-i\omega \tau} d\omega + \int_{-\infty}^{\infty} I(\omega) e^{i\omega \tau} d\omega \right\} \\
&= \int_{-\infty}^{\infty} \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I(\omega') e^{-i\omega' \tau} d\omega' \right) e^{i\omega \tau} d\tau + \int_{-\infty}^{\infty} \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I(\omega') e^{i\omega' \tau} d\omega' \right) e^{-i\omega \tau} d\tau \\
&= \sqrt{2\pi} \left[ \int_{-\infty}^{\infty} I(\omega') \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(\omega'-\omega) \tau} d\tau \right) d\omega' + \int_{-\infty}^{\infty} I(\omega') \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(\omega'+\omega) \tau} d\tau \right) d\omega' \right] \\
&= \sqrt{2\pi} \left[ I(\omega) \delta(\omega - \omega) + \int_{-\infty}^{\infty} I(\omega') \delta(\omega' + \omega) d\omega' \right] \\
&= \sqrt{2\pi} I(\omega) + I(-\omega)
\end{align*}
\] (8.30)

With (8.29) and (8.30) we can write (8.28) as

\[
\mathcal{F}\left\{ \eta \text{Sig}(\tau) \right\} = \sqrt{2\pi} I(\omega) + I(-\omega)
\] (8.31)

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 \(\omega\) (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 \(\tau\) looks something like that in Fig. 8.3. The oscillation frequency of the fringes lies in the neighborhood of \(\omega_0\). To obtain \(I(\omega)\) the procedure is clear: Record \(\text{Sig}(\tau)\); if desired, normalize by its value at large \(\tau\); 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 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

\[
I = 2I_0 \left[ 1 + \cos \left[ k(d_2 - d_1) + \phi_2 - \phi_1 \right] \right] = 2I_0 \left[ 1 + \cos \left( khy/D + \Delta \phi \right) \right] \quad (8.32)
\]

where \(\phi_1\) and \(\phi_2\) 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 \(\tau\) (the delay introduced by moving a mirror in the Michelson interferometer). To obtain the final expression in (8.32) we have made the approximations

\[
d_1(y) = \sqrt{(y - h/2)^2 + D^2} = D \sqrt{1 + \left( \frac{y-h/2}{D} \right)^2} \approx D \left[ 1 + \frac{(y-h/2)^2}{2D^2} + \cdots \right] \quad (8.33)
\]

and

\[
d_2(y) = \sqrt{(y + h/2)^2 + D^2} = D \sqrt{1 + \left( \frac{y+h/2}{D} \right)^2} \approx D \left[ 1 + \frac{(y+h/2)^2}{2D^2} + \cdots \right] \quad (8.34)
\]

![Figure 8.5](image)

**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.
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 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 \( \omega \) 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

\[
E_1(y'_j) = |E(y'_j)| e^{i\left[k\left[r_1(y'_j) + d_1(y)\right] - \omega t + \phi(y'_j)\right]} \tag{8.35}
\]

and

\[
E_2(y'_j) = |E(y'_j)| e^{i\left[k\left[r_2(y'_j) + d_2(y)\right] - \omega t + \phi(y'_j)\right]} \tag{8.36}
\]
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

\[
E_{\text{net}}(h) = \sum_j \left[ E_1(y'_j) + E_2(y'_j) \right]
\]  

(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

\[
I_{\text{net}}(h) = \frac{\varepsilon_0 c}{2} |E_{\text{net}}(h)|^2 = \frac{\varepsilon_0 c}{2} \left[ \sum_j E_1(y'_j) + E_2(y'_j) \right] \left[ \sum_m E_1(y'_m) + E_2(y'_m) \right]^* 
\]  

(8.38)

When inserting the field expressions (8.35) and (8.36) into this expression for the intensity at the screen, we get

\[
I_{\text{net}}(h) = \frac{\varepsilon_0 c}{2} \sum_{j,m} \left| E(y'_j) \right| \left| E(y'_m) \right| e^{ik\left[r_1(y'_j) - r_1(y'_m)\right]} e^{i\left[\phi(y'_j) - \phi(y'_m)\right]} \\
+ \left| E(y'_j) \right| \left| E(y'_m) \right| e^{ik\left[r_2(y'_j) - r_2(y'_m)\right]} e^{i\left[\phi(y'_j) - \phi(y'_m)\right]} \\
+ 2\text{Re} \left| E(y'_j) \right| \left| E(y'_m) \right| e^{ik\left[r_1(y'_j) - r_2(y'_m)\right]} e^{ik[d_1(y) - d_2(y)]} e^{i\left[\phi(y'_j) - \phi(y'_m)\right]} 
\]  

(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

\[
\left< e^{i [\phi(y') - \phi(y'\prime)_{m}]} \right>_t = \delta_{j, m} \equiv \begin{cases} 1 & \text{if } j = m, \\ 0 & \text{if } j \neq m. \end{cases} \quad \text{(random phase assumption)} \quad (8.40)
\]

The function \( \delta_{j, m} \) is known as the Kronecker delta function.

The time-averaged intensity under the random-phase assumption (8.40) becomes

\[
\left< I_{\text{net}}(h) \right>_t = \sum_j I(y_j') + \sum_j I(y_j') + 2 \text{Re} \sum_j I(y_j') e^{ik \left[ r_1(y_j') - r_2(y_j') \right]} e^{i k \left[ d_1(y_j) - d_2(y_j) \right]} \quad (8.41)
\]

We may use (8.33) to simplify \( d_1(y_j) - d_2(y_j) \equiv h y_j / D \), and similarly, we may simplify \( r_1(y_j') - r_2(y_j') \equiv y_j' h / R \) with the approximations

\[
r_1(y_j') = \sqrt{\left( y_j' - h/2 \right)^2 + R^2} \approx R \left[ 1 + \frac{\left( y_j' - h/2 \right)^2}{2 R^2} + \cdots \right] \quad (8.42)
\]

and

\[
r_2(y_j') = \sqrt{\left( y_j' + h/2 \right)^2 + R^2} \approx R \left[ 1 + \frac{\left( y_j' + h/2 \right)^2}{2 R^2} + \cdots \right] \quad (8.43)
\]

With these simplifications, (8.41) becomes

\[
\left< I_{\text{net}}(h) \right>_t = 2 \sum_j I\left( y_j' \right) + 2 \text{Re} e^{-i \frac{kh}{2D}} \sum_j I\left( y_j' \right) e^{-i \frac{kh}{\pi}} \quad \text{(random phase assumption)} \quad (8.44)
\]

The only thing left to do is to put this formula into a slightly more familiar form:

\[
\left< I_{\text{net}}(h) \right>_t = \left[ 2 \sum_j I\left( y_j' \right) \right] \left[ 1 + \text{Re} \gamma(h) \right] \quad \text{(random phase assumption)} \quad (8.45)
\]

where

\[
\gamma(h) \equiv \frac{e^{-i \frac{kh}{\pi}} \sum_j I\left( y_j' \right) e^{-i \frac{kh}{\pi}}}{\sum_j I\left( y_j' \right)} \quad (8.46)
\]

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 \( \exp(-i k h y / D) \) 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 \( \gamma(h) \) tends to diminish until the intensity at the screen becomes uniform. When
the two slits have very small separation (such that $e^{-i \frac{kh}{\lambda}} \approx 1$ wherever $I(y')$ is significant) then we have $|\gamma(h)| = 1$ and very good fringe visibility results. As the slit separation $h$ increases, the fringe visibility

$$V(h) = |\gamma(h)|$$ (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:

$$h_c \equiv 2 \int_{0}^{\infty} |\gamma(h)|^2 dh$$ (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

$$\langle I_{\text{net}}(h) \rangle_t = 2 \langle I_{\text{oneslit}} \rangle_t [1 + \text{Re}\gamma(h)] \quad \text{(random phase assumption)}$$ (8.49)

where

$$\gamma(h) \equiv \frac{e^{-i \frac{kh}{\lambda}} \int_{-\infty}^{\infty} I(y') e^{-i \frac{kh}{\lambda}} dy'}{\int_{-\infty}^{\infty} I(y') dy'}$$ (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:
\[
\sum_j E_1(y_j') \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_1(y') dy' \\
\sum_m E_1(y_m') \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_1(y'') dy'' \\
\sum_j E_2(y_j') \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_2(y') dy' \\
\sum_m E_2(y_m') \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_2(y'') dy''
\] 
\text{(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

\[
I_{\text{net}}(h) = \varepsilon_0 c^2 \left[ \int_{-\infty}^{\infty} |E(y')| e^{ik_1(y')} e^{i\phi(y')} dy' \int_{-\infty}^{\infty} |E(y'')| e^{-ik_1(y'')} e^{-i\phi(y'')} dy'' \\
+ \frac{1}{2\pi} \int_{-\infty}^{\infty} |E(y')| e^{ik_2(y')} e^{i\phi(y')} dy' \int_{-\infty}^{\infty} |E(y'')| e^{-ik_2(y'')} e^{-i\phi(y'')} dy'' \\
+ 2\text{Re} \left[ \frac{e^{ik(d_1(y') - d_2(y))}}{2\pi} \int_{-\infty}^{\infty} |E(y')| e^{ik_1(y')} e^{i\phi(y')} dy' \int_{-\infty}^{\infty} |E(y'')| e^{-ik_2(y'')} e^{-i\phi(y'')} dy'' \right] \right]
\] 
\text{(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 \(\phi(y_j') \to Ky',\) 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

\[
\left\langle e^{i|\phi(y_j') - \phi(y_m')}\right\rangle_t = \delta_{j,m} \to \int_{-\infty}^{\infty} e^{iK(y'' - y')} dK = 2\pi \delta(y'' - y')
\] 
\text{(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

$$
\int_{-\infty}^{\infty} I_{\text{net}}(h) dK = \frac{\epsilon_0 c}{2} \left[ \int_{-\infty}^{\infty} |E(y')| e^{ikr_1(y')} dy' \int_{-\infty}^{\infty} |E(y'')| e^{-ikr_1(y'')} \delta(y'' - y') dy'' + \int_{-\infty}^{\infty} |E(y')| e^{ikr_2(y')} dy' \int_{-\infty}^{\infty} |E(y'')| e^{-ikr_2(y'')} \delta(y'' - y') dy'' + 2\text{Re} e^{ik|d_1(y) - d_2(y)|} \int_{-\infty}^{\infty} |E(y')| e^{ikr_1(y')} dy' \int_{-\infty}^{\infty} |E(y'')| e^{-ikr_2(y'')} \delta(y'' - y') dy'' \right]
$$

(random phase assumption)

(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

$$
\int_{-\infty}^{\infty} I_{\text{net}}(h) dK = 2 \int_{-\infty}^{\infty} I(y') dy' + 2\text{Re} e^{-iky'\Delta} \int_{-\infty}^{\infty} I(y') e^{-iky'} dy' \quad \text{(random phase assumption)}
$$

(8.55)

where

$$
I(y') \equiv \frac{1}{2} \epsilon_0 c |E(y')|^2
$$

(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 $I_{\text{net}}(h)$ has usual units of intensity.) We can renormalize these strange units on each side of the equation. We can redefine the left-hand side $\int_{-\infty}^{\infty} I_{\text{net}}(h) dK$ 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

$$I_{\text{net}}(h) = \varepsilon_0 \frac{c^2 p^2 \pi}{2} \left[ \left| \int_{-\infty}^{\infty} |E(y')| e^{i\phi(y') + \frac{ky'y'}{2R}} e^{-i\frac{kyy'}{2R}} dy' \right|^2 + \int_{-\infty}^{\infty} \left| E(y') | e^{i\phi(y') + \frac{ky'y'}{2R}} e^{i\frac{kyy'}{2R}} dy' \right|^2 \right]$$

$$+ 2\text{Re} \frac{e^{i\frac{kyy'}{2R}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left| E(y') | e^{i\phi(y') + \frac{ky'y'}{2R}} e^{-i\frac{kyy'}{2R}} dy' \right|^2$$

$$\left\{ \int_{-\infty}^{\infty} \left| E(y'') | e^{i\phi(y'') + \frac{ky'y''}{2R}} e^{i\frac{kyy''}{2R}} dy'' \right|^2 \right\}^*$$

(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:

$$\phi(y') = -\frac{k y'^2}{2R} \text{ (converging spherical wave)}$$

(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

$$I_{\text{net}}(h) = \frac{\varepsilon_0 c^2}{2} \left[ \left| \int_{-\infty}^{\infty} |E(y')| e^{-i\frac{kyy'}{2R}} dy' \right|^2 + \int_{-\infty}^{\infty} \left| E(y') | e^{i\frac{kyy'}{2R}} dy' \right|^2 \right]$$

$$+ 2\text{Re} \frac{e^{i\frac{kyy'}{2R}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left| E(y') | e^{-i\frac{kyy'}{2R}} dy' \right|^2$$

$$\left\{ \int_{-\infty}^{\infty} \left| E(y'') | e^{i\frac{kyy''}{2R}} dy'' \right|^2 \right\}^*$$

(8.59)

(converging spherical wave)

There is a close resemblance between the expression

$$|E_{\text{slitone}}(h/2)| \equiv \left| \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |E(y')| e^{-i\frac{kyy'}{2R}} dy' \right|$$

(8.60)
and the magnitude of the degree of coherence $V = |\gamma(h)|$ from (8.50). Here $E_{\text{slitone}}$ 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

$$\frac{|E_{\text{slitone}}(h)|}{|E_{\text{slitone}}(0)|} = \left| \frac{\int_{-\infty}^{\infty} |E(y')| e^{-i\frac{kh}{\lambda} y'} dy'}{\int_{-\infty}^{\infty} |E(y')| dy'} \right| (\text{converging spherical wave assumption})$$

(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

Exercises for 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

\[
I = 2I_0 [1 + \cos(\omega_0 \tau)]
\]
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 \( \tau \).

HINT: See (6.53).

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 \( \tau \) 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

\[
E(\omega) = \frac{iE(\omega_0) e^{i\omega_0 z}}{i + \frac{\omega - \omega_0}{\Delta \omega_{\text{FWHM}}/2}}
\]

Perform the inverse Fourier transform on the field and find how the intensity of the light looks a function of time.

HINT:

\[
\int_{-\infty}^{\infty} \frac{e^{-iax}}{x + \beta} dx = \begin{cases} 
-2i\pi e^{i\alpha\beta} & \text{if } a > 0 \\
0 & \text{if } a < 0 
\end{cases} \quad (\text{Im}\beta > 0)
\]

The constants \( I(\omega_0) \), and \( \Delta \omega_{\text{FWHM}} \) will appear in the answer.
Exercises for 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 $\gamma$ that give rise to fringes are due entirely to changes in $\phi$ and that $|\gamma|$ is a slowly varying function in comparison to the oscillations.

(b) What is the coherence time $\tau_c$ 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 $\gamma$ 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 $\ell_c$ and the center wavelength $\lambda_0$ associated with (8.20).

HINT: Derive $\Delta\omega_{\text{FWHM}} = 2\sqrt{\ln2}\Delta\omega$. To convert to a wavelength difference, use $\lambda = \frac{2\pi c}{\omega} \Rightarrow \Delta\lambda \approx -\frac{2\pi c}{\omega^2} \Delta\omega$. You can ignore the minus sign; it simply means that wavelength decreases as frequency increases.

Exercises for 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 $\lambda$ 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?

**Exercises for 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.

(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)

\[
\gamma(h) = \frac{a^2}{\pi} \int_{-a/2}^{a/2} \frac{\sin^2 \frac{x}{a}}{\left( \alpha x \right)^2} dx = \frac{\pi}{4} \sinc \frac{ka}{2R}
\]

Note that

\[
\int_{0}^{\infty} \frac{\sin^2 \alpha x}{\left( \alpha x \right)^2} dx = \frac{\pi}{2a}
\]
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 $\omega$.

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 \( \lambda_{\text{vac}} = 500 \text{ nm} \) 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 \( T_{s}^{\text{tot}} \) for this thickness assuming \( s \)-polarized light?

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

\[
\Delta \lambda_{\text{FSR}} = \frac{\lambda^2}{2nd \cos \theta}
\]

(b) Show that the fringe width \( \Delta \lambda_{\text{FWHM}} \) is

\[
\frac{\lambda^2}{\pi \sqrt{F} nd \cos \theta}
\]

where \( F \equiv \frac{4R}{(1-R)^2} \).

(c) Derive the reflecting finesse \( f = \Delta \lambda_{\text{FSR}} / \Delta \lambda_{\text{FWHM}} \).

R42  For a Fabry-Perot etalon, let \( R = 0.90, \lambda_{\text{vac}} = 500 \text{ nm} \), \( n = 1 \), and \( d = 5.0 \text{ mm} \).
(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 $\cos \theta \approx 1 - \theta^2/2$.

(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.

The matrix equation relating the incident field to the reflected and transmitted fields (at normal incidence) is

$$\begin{pmatrix} 1 & 1 \\ n_0 & -n_0 \end{pmatrix} \begin{pmatrix} E_{\text{incident}}^\text{reflected} \\ E_{\text{incident}}^\text{transmitted} \end{pmatrix} = \begin{pmatrix} \cos k_1 \ell & -i n_1 \sin k_1 \ell \\ -i n_1 \sin k_1 \ell & \cos k_1 \ell \end{pmatrix} \begin{pmatrix} 1 \\ n_t \end{pmatrix} \begin{pmatrix} E_{\text{incident}}^\text{reflected} \\ E_{\text{incident}}^\text{transmitted} \end{pmatrix}$$

(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 $n_1 < n_t$.

(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

$$E(t) = \begin{cases} E_0 e^{-i\omega_0 t}, & |t| \leq \tau/2 \\ 0, & |t| > \tau/2 \end{cases}$$

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

$$E(\omega) = \begin{cases} E_0, & |\omega - \omega_0| \leq \Delta \omega/2 \\ 0, & |\omega - \omega_0| > \Delta \omega/2 \end{cases}$$

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 $\ell$ 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.
(a) Prove Parseval's theorem:

\[ \int_{-\infty}^{\infty} |E(\omega)|^2 d\omega = \int_{-\infty}^{\infty} |E(t)|^2 dt. \]

HINT:

\[ \delta(t' - t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega(t' - t)} d\omega \]

(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 mm². 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

\[ I(\omega) = I(\omega_0) e^{-\frac{(\omega - \omega_0)^2}{\delta\omega^2}} \]

Use an approximate value \( \delta\omega \approx \frac{2\pi c}{\lambda^2} \Delta\lambda \). Find a value and correct units for \( I(\omega_0) \).

HINT:

\[ \int_{-\infty}^{\infty} e^{-Ax^2+Bx+C} dx = \sqrt{\frac{\pi}{A}} e^{B^2/4A+C} \text{ Re} \{A\} > 0 \]

Continuous light entering a Michelson interferometer has a spectrum described by

\[ I(\omega) = \begin{cases} I_0, & |\omega - \omega_0| \leq \Delta\omega/2 \\ 0, & |\omega - \omega_0| > \Delta\omega/2 \end{cases} \]

The Michelson interferometer uses a 50:50 beam splitter. The emerging light has intensity \( \langle I_{\text{det}} (I, \tau) \rangle_t = 2 \langle I(t) \rangle_t [1 + \text{Re} \{\gamma(\tau)\}] \), 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 \( V \equiv (I_{\text{max}} - I_{\text{min}})/(I_{\text{max}} + I_{\text{min}}) \) as a function of \( \tau \) (i.e. the round-trip delay due to moving one of the mirrors).

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

\[ I_{\text{screen}} (y', h) = 2I(y') \left[ 1 + \cos \left( kh \left( \frac{y}{D} + \frac{y'}{R} \right) \right) \right] \]

where an approximation has restricted us to small angles.
(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:

\[
I_{\text{screen}}(h) \equiv \int_{-\infty}^{\infty} I_{\text{screen}}(y', h) \, dy'
\]

(b) Assume that the source has an emission distribution with the form 

\[
I(y') = \left( \frac{I_0}{\Delta y'} \right) e^{-y'^2/\Delta y'^2}.
\]

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:

\[
\int_{-\infty}^{\infty} e^{-Ax^2+Bx+C} \, dx = \sqrt{\frac{\pi}{A}} e^{B^2/4A+C} \quad \text{Re}\{A\} > 0.
\]

(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 

\[
V \equiv \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} = \frac{1}{3}
\]

Selected Answers

R40: (a) 100 nm. (b) 0.55.
R42: (a) 0.074°.
R43: (b) 1.24.
R45: (b) \( 3.8 \times 10^{-16} \text{J/(cm}^2 \cdot \text{s}) \).
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 “εἰκώνς” 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
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 $2 \times 2$ 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

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

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 $f_{\text{eff}}$.

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 (2.21) for a medium with a real index of refraction:

$$\nabla^2 E(r, t) - \frac{n(r)^2}{c^2} \frac{\partial^2 E(r, t)}{\partial t^2} = 0 \quad (9.2)$$

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 \( n(r) \) is an arbitrary function of \( r \). In this case, the usual plane-wave solutions no longer satisfy the wave equation.

We consider the light to have a single frequency \( \omega \). As a trial solution for (9.2), we take

$$E(r, t) = E_0(r) e^{i[k_{\text{vac}} R(r) - \omega t]} \quad (9.3)$$

where

$$k_{\text{vac}} = \frac{\omega}{c} = \frac{2\pi}{\lambda_{\text{vac}}} \quad (9.4)$$

Here \( R(r) \) is a real scalar function (which depends on position) having the dimension of length. By assuming that \( R(r) \) 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 \( R(r) \) 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 \( R(r) = \text{constant} \) can be interpreted as wave fronts of the solution. The wave fronts travel in the direction for which \( R(r) \) varies the

![Figure 9.1 Wave fronts distributed throughout space in the presence of a spatially inhomogeneous refractive index.](image-url)
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 $R(\mathbf{r}) = \mathbf{k} \cdot \mathbf{r} / \kappa_{\text{vac}}$ 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

$$\nabla^2 \left[ \mathbf{E}_0(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} \right] + \frac{n^2(\mathbf{r})}{c^2} E_0(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]} = 0 \quad (9.5)$$

We divide each term by $e^{-i\omega t}$ and utilize (9.4) to rewrite the wave equation as

$$\frac{1}{k_{\text{vac}}^2} \nabla^2 \left[ \mathbf{E}_0(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})]} \right] + n^2(\mathbf{r}) E_0(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})]} = 0 \quad (9.6)$$

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

$$\nabla \left[ E_{0x}(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})]} \right] = [\nabla E_{0x}(\mathbf{r})] e^{i[k_{\text{vac}}R(\mathbf{r})]} + i k_{\text{vac}} E_{0x}(\mathbf{r}) [\nabla R(\mathbf{r})] e^{i[k_{\text{vac}}R(\mathbf{r})]}$$

The Laplacian of the $x$ component is

$$\nabla \cdot \nabla \left[ E_{0x}(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})]} \right] = \{\nabla^2 E_{0x}(\mathbf{r}) - k_{\text{vac}}^2 E_{0x}(\mathbf{r}) [\nabla R(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] + i k_{\text{vac}} E_{0x}(\mathbf{r}) [\nabla^2 R(\mathbf{r})] + 2i k_{\text{vac}} [\nabla E_{0x}(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] \} e^{i[k_{\text{vac}}R(\mathbf{r})]}$$

Upon combining the result for each vector component of $\mathbf{E}_0(\mathbf{r})$, the required spatial derivative can be written as

$$\nabla^2 \left[ E_{0x}(\mathbf{r}) e^{i[k_{\text{vac}}R(\mathbf{r})]} \right] = \{\nabla^2 E_{0x}(\mathbf{r}) - k_{\text{vac}}^2 E_{0x}(\mathbf{r}) [\nabla R(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] + i k_{\text{vac}} E_{0x}(\mathbf{r}) [\nabla^2 R(\mathbf{r})] + 2i k_{\text{vac}} [\nabla E_{0x}(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] + \hat{z} [\nabla E_{0x}(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] \} e^{i[k_{\text{vac}}R(\mathbf{r})]}$$

Using the result from Example 9.1 with some additional rearranging, (9.6) becomes

$$[\nabla R(\mathbf{r}) \cdot \nabla R(\mathbf{r}) - n^2(\mathbf{r})] E_0(\mathbf{r}) = \frac{\nabla^2 E_0(\mathbf{r})}{k_{\text{vac}}^2} + i \frac{\nabla^2 R(\mathbf{r})}{k_{\text{vac}}} + 2i \frac{\hat{x} \nabla E_{0x}(\mathbf{r}) \cdot \nabla R(\mathbf{r})}{k_{\text{vac}}}$$

$$+ 2i \frac{\hat{x} [\nabla E_{0y}(\mathbf{r})] \cdot \nabla R(\mathbf{r}) + \hat{y} [\nabla E_{0z}(\mathbf{r})] \cdot \nabla R(\mathbf{r})}{k_{\text{vac}}}$$

$$\quad (9.7)$$
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

$$[\nabla R(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] = n^2(\mathbf{r}), \quad (9.8)$$

Written another way, this equation is

$$\nabla R(\mathbf{r}) = n(\mathbf{r}) \hat{s}(\mathbf{r}) \quad (9.9)$$

This latter form is called the *eikonal equation* where $\hat{s}$ is a unit vector pointing in the direction $\nabla R(\mathbf{r})$, the direction normal to wave front surfaces.

Under the assumption of an infinitely short wavelength, the Poynting vector is directed along $\hat{s}$ as demonstrated in P 9.2. In other words, the direction of $\hat{s}$ specifies the direction of energy flow. The unit vector $\hat{s}$ 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 $\hat{s}$ 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 $R(\mathbf{r}) = \text{constant}$ 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

$$\nabla \times [n(\mathbf{r}) \hat{s}(\mathbf{r})] = \nabla \times [\nabla R(\mathbf{r})] = 0 \quad (9.10)$$

(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

$$\int_A \nabla \times [n(\mathbf{r}) \hat{s}(\mathbf{r})] \, da = 0 \quad (9.11)$$
We next apply Stokes’ theorem (0.12) to the integral and convert it to a path integral around the perimeter of the area. Then we get

\[ \oint_C \mathbf{n}(\mathbf{r}) \hat{s}(\mathbf{r}) \cdot d\mathbf{\ell} = 0 \]  
(9.12)

The integration of \( n\hat{s} \cdot d\ell \) around a closed loop is always zero. Keep in mind that the proper value for \( \hat{s}(\mathbf{r}) \) 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:

\[ \int_A^B n\hat{s} \cdot d\ell \]  
(9.13)

Now consider points A and B that lie along a path that is always parallel to \( \hat{s} \) (i.e. perpendicular to the wave fronts as depicted in Fig. 9.2). When integrating along the path parallel to \( \hat{s} \), 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 \( \hat{s} \). 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 \( \hat{s} \).

With the dot product removed from (9.13), the result of the integration agrees with the true result only for the path taken along \( \hat{s} \) (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

\[ \int_A^B n\hat{s} \cdot d\ell = \min \left\{ \int_A^B n d\ell \right\} \]  
(9.14)

The integral on the right gives the optical path length (OPL) between A and B:

\[ \text{OPL}_{AB}^B \equiv \int_A^B n d\ell \]  
(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 \( \hat{s} \)) 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 \( \Delta t \) depends not only on the path taken by the
light but also on the velocity of the light $v(r)$, which varies spatially with the refractive index:

$$\Delta t^B_A = \int_A^B \frac{d\ell}{v(r)} = \int_A^B \frac{d\ell}{c/n(r)} = \frac{OPL_A^B}{c}$$

(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 anisotropic 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 (2.27)). 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](image-url) 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

\[ OPL = n_i \sqrt{\Delta x_i^2 + \Delta y_i^2} + n_t \sqrt{\Delta x_t^2 + \Delta y_t^2} \]  

(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 \( \Delta x_i \) and \( \Delta x_t \) as constants. The distances \( \Delta y_i \) and \( \Delta y_t \) are not constants although the combination

\[ \Delta y_{\text{tot}} = \Delta y_i + \Delta y_t \]  

(9.18)

is constant. Thus, we may rewrite (9.17) as

\[ OPL(\Delta y_i) = n_i \sqrt{\Delta x_i^2 + \Delta y_i^2} + n_t \sqrt{\Delta x_t^2 + (\Delta y_{\text{tot}} - \Delta y_i)^2} \]  

(9.19)

where everything in the right-hand side of the expression is constant except for \( \Delta y_i \).

We now minimize the optical path length by taking the derivative and setting it equal to zero:

\[ \frac{dOPL}{d\Delta y_i} = n_i \frac{\Delta y_i}{\sqrt{\Delta x_i^2 + \Delta y_i^2}} + n_t \frac{-(\Delta y_{\text{tot}} - \Delta y_i)}{\sqrt{\Delta x_t^2 + (\Delta y_{\text{tot}} - \Delta y_i)^2}} = 0 \]  

(9.20)

Notice that

\[ \sin \theta_i = \frac{\Delta y_i}{\sqrt{\Delta x_i^2 + \Delta y_i^2}} \]  

and

\[ \sin \theta_t = \frac{\Delta y_t}{\sqrt{\Delta x_t^2 + \Delta y_t^2}} \]  

(9.21)

When these are substituted into (9.20) we obtain

\[ n_i \sin \theta_i = n_t \sin \theta_t \]  

(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.

**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 halfway between the points, which are separated by a distance $2a$, as shown in Fig. 9.5.

![Figure 9.5](image)

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

$$b = \sqrt{(x + a)^2 + y^2} + \sqrt{(x - a)^2 + y^2}$$  \hspace{1cm} (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 the remaining binomial terms

$$(x^2 + a^2 + 2ax) + y^2 = b^2 + (x^2 + a^2 - 2ax) + y^2 - 2b\sqrt{(x - a)^2 + y^2}.$$  

Some nice cancellation 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

$$16a^2x^2 - 4ab^2x + b^4 = 4b^2(x^2 - 2ax + a^2 + y^2),$$

and then cancel and regroup terms to arrive at

$$(16a^2 - 4b^2)x^2 - 4b^2y = 4a^2b^2 - b^4.$$
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

\[ \sin \theta \equiv \theta \quad \text{(9.24)} \]

and similarly

\[ \tan \theta \equiv \theta \quad \text{(9.25)} \]

Here, the angle \( \theta \) (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. 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 $2 \times 2$ 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 $\theta_1$ (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 $\theta_1$:

$$y_2 = y_1 + d \tan \theta_1 \tag{9.27}$$

where $d \equiv z_2 - z_1$. Equation (9.27) is nonlinear in $\theta_1$. 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

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \quad \text{(propagation through a distance $d$)} \tag{9.29}$$

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

$$\begin{bmatrix} y_{\text{mid}} \\ \theta_{\text{mid}} \end{bmatrix} = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$$

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{\text{mid}} \\ \theta_{\text{mid}} \end{bmatrix} \tag{9.30}$$
where the subscript "mid" refers to the ray in the middle position after traversing the distance $a$. If we combine the equations, we get

$$
\begin{bmatrix}
y_2 \\
\theta_2
\end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\
\theta_1
\end{bmatrix}
$$

(9.31)

which is in complete agreement with (9.29) since the ABCD matrix for the entire displacement is

$$
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & a+b \\ 0 & 1 \end{bmatrix}
$$

(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

$$
y_2 = y_1
$$

(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 $\theta_2$ is a negative angle (as opposed to $\theta_1$ which is drawn as a positive angle) and is equal to

$$
\theta_2 = - (\theta_1 + 2\theta_i)
$$

(9.34)
where $\theta_i$ 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 $\theta_i$ 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 $\theta_i$ from (9.34) in favor of $\theta_1$ and $y_1$. By inspection of Fig. 9.7 we can write

$$\frac{y_1}{R} = \sin \phi \equiv \phi$$

(9.35)

where we have applied the paraxial approximation (9.24). (Note that the angles in the figure are exaggerated.) We also have

$$\phi = \theta_1 + \theta_i$$

(9.36)

and when this is combined with (9.35), we get

$$\theta_i = \frac{y_1}{R} - \theta_1$$

(9.37)

With this we are able to put (9.34) into a useful linear form:

$$\theta_2 = -\frac{2}{R} y_1 + \theta_1$$

(9.38)

Equations (9.33) and (9.38) describe a linear transformation that can be concisely formulated as

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2/R & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$$

(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 $\theta_1$ and $\theta_2$ we must use Snell’s law which in the paraxial approximations is

$$n_i \theta_i = n_t \theta_t$$

(9.40)

As seen in the Fig. 9.8, we have

$$\theta_i = \theta_1 + \phi$$

(9.41)
As before, (9.35) applies (i.e. \( \phi \equiv y_1/R \)). When this is used in (9.41) and (9.42), Snell’s law (9.40) becomes

\[
\theta_2 = \left( \frac{n_i}{n_t} - 1 \right) \frac{y_1}{R} + \frac{n_i}{n_t} \theta_1
\]

The compact matrix form of (9.33) and (9.43) turns out to be

\[
\begin{bmatrix}
y_2 \\
\theta_2
\end{bmatrix} =
\begin{bmatrix}
1 & b \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
\begin{bmatrix}
1 \\
-2/R
\end{bmatrix}
\begin{bmatrix}
1 & a \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
y_1 \\
\theta_1
\end{bmatrix}
\]

(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:

\[
\begin{bmatrix}
y_2 \\
\theta_2
\end{bmatrix} =
\begin{bmatrix}
1 & b \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
\begin{bmatrix}
1 \\
-2/R
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
\begin{bmatrix}
1 & a \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
y_1 \\
\theta_1
\end{bmatrix}
\]

(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).

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

$$\delta = R - R \cos \phi$$

(9.46)

In the paraxial approximation, we have $\cos \phi \approx 1 - \phi^2/2$. And since in this approximation we may also write $\phi \approx \sqrt{x^2 + y^2}/R$, (9.46) becomes

$$\delta \approx \frac{x^2 + y^2}{2R}$$

(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

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 - 2b/R & a + b - 2ab/R \\ -2/R & 1 - 2a/R \end{bmatrix}$$

(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 $d_o = a$ before the mirror. In order for an image to occur at $d_i = b$, 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 $\theta_1$) all to converge to a single point $y_2$ at the image. In the following equation we need $y_2$ to be independent of $\theta_1$:

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} Ay_1 + B\theta_1 \\ Cy_1 + D\theta_1 \end{bmatrix}$$

(9.49)

The condition for image formation is therefore

$$B = 0 \quad \text{(condition for image formation)} \quad (9.50)$$

When this condition is applied to (9.48), we obtain

$$d_o + d_i - 2d_0d_i \frac{1}{R} = 0 \Rightarrow \frac{2}{R} = \frac{1}{d_o} + \frac{1}{d_i}$$

(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. $d_o \to \infty$), the image appears at $d_i \to R/2$. This distance is called the focal length and is denoted by

$$f = \frac{R}{2} \quad \text{(focal length of a mirror)} \quad (9.52)$$
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

\[
M \equiv \frac{y_2}{y_1} = A = 1 - \frac{2d_i}{R} = -\frac{d_i}{d_o}
\]  

(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

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix} = \begin{bmatrix}
\frac{1}{R_2} (n-1) & n \\
\frac{1}{R_1} & \frac{1}{n}
\end{bmatrix} \begin{bmatrix}
\frac{1}{R_1} & 0 \\
-(n-1) & 1
\end{bmatrix} 
\]  

(Thin Lens)  

(9.54)

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

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix} = \begin{bmatrix}
1 & 0 \\
-1/f & 1
\end{bmatrix}
\]  

(9.55)

where in the case of the thin lens the focal length is given by the lens maker’s formula

\[
\frac{1}{f} = (n-1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right) \quad \text{(focal length of thin lens)}
\]  

(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):

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix} = \begin{bmatrix}
1 - d_i/f & d_o + d_i - d_o d_i/f \\
-1/f & 1 - d_o/f
\end{bmatrix}
\]  

(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 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:

\[
\begin{bmatrix}
1 & d_i & A & B \\
0 & 1 & C & D \\
\end{bmatrix}
\begin{bmatrix}
1 & d_o \\
0 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
A + d_i C & d_o A + B + d_o d_i C + d_i D \\
C & d_o C + D \\
\end{bmatrix}
\]

(9.58)

Imaging occurs according to (9.50) when \( B' = 0 \), or

\[
d_o A + B + d_o d_i C + d_i D = 0, \quad \text{(general condition for image formation)} \quad (9.59)
\]

with magnification

\[
M = A + d_i C
\]

(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.)

\[\text{Figure 9.10} \quad \text{A multi-element system represented as an ABCD matrix for which principal planes always exist.}\]
Next we demonstrate that \( p_1 \) and \( p_2 \) can always be selected such that we can write

\[
\begin{bmatrix}
  1 & p_2 \\
  0 & 1
\end{bmatrix}
\begin{bmatrix}
  A & B \\
  C & D
\end{bmatrix}
\begin{bmatrix}
  1 & p_1 \\
  0 & 1
\end{bmatrix}
= 
\begin{bmatrix}
  A + p_2 C & p_1 A + B + p_1 p_2 C + p_2 D \\
  C & p_1 C + D
\end{bmatrix}
\begin{bmatrix}
  1 & 0 \\
  -1/f_{\text{eff}} & 1
\end{bmatrix}
\]

(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

\[ f_{\text{eff}} \equiv -1/C \]

(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

\[ p_1 C + D = 1 \Rightarrow p_1 = \frac{1-D}{C} \]

(9.63)

and

\[ A + p_2 C = 1 \Rightarrow p_2 = \frac{1-A}{C} \]

(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

\[
p_1 A + B + p_1 p_2 C + p_2 D = \frac{1-D}{C} A + B + \frac{1-D}{C} 1 - \frac{1-A}{C} C + \frac{1-A}{C} D
\]

\[
= \frac{1}{C} \left[ 1 - AD + BC \right]
\]

(9.65)

\[
= \frac{1}{C} \left[ 1 - \begin{bmatrix}
  A & B \\
  C & D
\end{bmatrix} \right]
\]

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

\[
\begin{bmatrix}
  A & B \\
  C & D
\end{bmatrix} = 1
\]

(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
\[
\begin{bmatrix}
1 & d \\
0 & 1
\end{bmatrix}
\]  
(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 \(f = R/2\))

\[
\begin{bmatrix}
1 + \frac{d}{R_1} \left(\frac{1}{n} - 1\right) \\
(1 - n) \left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{d}{R_1 R_2} \left(2 - \frac{1}{n} - n\right) & 1 - \frac{d}{R_2} \left(\frac{1}{n} - 1\right)
\end{bmatrix}
\]  
(Thick lens)

Table 9.1 Summary of ABCD matrices for common optical elements.

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 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 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
\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}
= \begin{bmatrix}
1 & L \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-2/R_2 & 1
\end{bmatrix}
\begin{bmatrix}
1 & L \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-2/R_1 & 1
\end{bmatrix}
\] (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
\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}
= \begin{bmatrix}
1 & 2L_1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-1/f & 1
\end{bmatrix}
\begin{bmatrix}
1 & 2L_2 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-1/f & 1
\end{bmatrix}
\] (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:
\[
\begin{bmatrix}
y_{N+1} \\
\theta_{N+1}
\end{bmatrix}
= \begin{bmatrix}
A & B \\
C & D
\end{bmatrix}^N
\begin{bmatrix}
y_1 \\
\theta_1
\end{bmatrix}
\] (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 0.5, which states that if
\[
\begin{vmatrix}
A & B \\
C & D
\end{vmatrix}
= 1
\] (9.70)
then
\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}^N
= \frac{1}{\sin \theta}
\begin{bmatrix}
A \sin N\theta - \sin (N - 1) \theta & B \sin N\theta \\
C \sin N\theta & D \sin N\theta - \sin (N - 1) \theta
\end{bmatrix}
\] (9.71)
where
\[
\cos \theta = \frac{1}{2} (A + D).
\] (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 \( \theta \) 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 \( \theta \) must exist for (9.72), or in other words we need

\[
-1 < \frac{1}{2} (A + D) < 1 \quad \text{(condition for a stable cavity)} \quad (9.73)
\]

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

*Figure 9.12* Ray tracing through a simple lens.
Chromatic aberration causes lenses to have different focal lengths for different wavelengths. It can be corrected using an achromatic doublet lens.

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 aberration 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 aberration 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
9.9 Aberrations and Ray Tracing

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.

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 aberration at two wavelengths and do a good job for a wide range of others. Achromatic doublets can also be designed to minimize spherical aberration (see below), so they are often a good choice when you need a high quality lens.

*Monochromatic aberrations* 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 aberrations had to be analyzed primarily with analytic techniques. The 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 aberration terms added in (shown conceptually in Fig. 9.14(b)). There are five aberration terms in this second-order analysis, and these represent a convenient basis for discussing abberation.

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 \(\rho_b\) in the plane of the figure are imaged to point \(b\). Rays that travel through the lens at radius \(\rho_b\), but outside the plane of the figure are imaged to other points on the circle (in the image plane) containing point \(b\). Rays that travel through the lens at other radii on the lens (e.g. \(\rho_c\)) also form circles in the image plane with radius proportional to \(\rho^2\) with the center offset from point \(a\) a distance proportional to \(\rho^2\). When light from each of these circles combines on the screen it produces an imaged point with a “comet tail.”

The first aberration term is known as spherical aberration. This type of aberration 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 aberations from one lens are corrected by the other.

The aberration term referred to as astigmatism occurs when an off-axis object point 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 aberration can be corrected by inserting a cylindrical lens at the correct orientation (this is a common correction needed in eyeglasses).

A third aberration 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 aberration got its name.) This aberration 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 aberration 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 aberration 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 aberration term is referred to as distortion. This aberration 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 aberrations listed above (i.e. chromatic aberration plus the five second-order aberration terms). In addition to the five named monochromatic aberrations, there are many other higher order aberrations that also have to be considered. Aberrations 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 aberration, so designing a system always involves some degree of compromise in choosing which aberrations to minimize and which ones you can live with.
Exercises

Exercises for 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 \( n(y) = n_0 \sqrt{1 + y^2 / h^2} \). Show that \( R(x, y) = n_0 x - n_0 y^2 / 2h \) is a solution of the eikonal equation (9.9).

(b) Give an expression for \( \mathbf{s} \) as a function of \( y \).

(c) Compute \( \mathbf{s} \) 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 \( \nabla R(\mathbf{r}) \) or \( \mathbf{s} \). First work through the partial solution provided below (write your work down), then finish as directed.

Solution: (partial)

First, from Faraday’s law (1.44) we have

\[
\mathbf{B}(\mathbf{r}, t) = \frac{i}{\omega} \mathbf{E}_0(\mathbf{r}) e^{ik_{\text{vac}} R(\mathbf{r}, t)}
\]

Applying the identity \( \nabla \times (a \mathbf{v}) = \mathbf{v} \times \nabla a + \nabla \times \mathbf{v} \) to this equation, we obtain:

\[
\begin{align*}
\mathbf{B}(\mathbf{r}, t) &= \frac{i}{\omega} \left( e^{ik_{\text{vac}} R(\mathbf{r}, t)} \left[ \nabla \times \mathbf{E}_0(\mathbf{r}) + i k_{\text{vac}} e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)} \left[ \nabla R(\mathbf{r}) \times \mathbf{E}_0(\mathbf{r}) \right] \right] \right) \\
&= \frac{i k_{\text{vac}}}{2\pi c} e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)} \left[ \nabla \times \mathbf{E}_0(\mathbf{r}) \right] - \frac{1}{c} e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)} \left[ \nabla R(\mathbf{r}) \times \mathbf{E}_0(\mathbf{r}) \right]
\end{align*}
\]

The first term vanishes in the limit of very short wavelength, and we have:

\[
\mathbf{B}(\mathbf{r}, t) = -\frac{1}{c} \nabla R(\mathbf{r}) \times \mathbf{E}_0(\mathbf{r}) e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)}. \tag{9.74}
\]

Next, from Gauss’s law (1.42) and the constitutive relation (2.16) we have

\[
\nabla \cdot \left[ (1 + \chi(\mathbf{r})) \mathbf{E}_0(\mathbf{r}) e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)} \right] = 0
\]

Applying the identity \( \nabla \cdot (a \mathbf{v}) = a \nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla a \) to this expression yields:

\[
e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)} \nabla \cdot \left[ (1 + \chi(\mathbf{r})) \mathbf{E}_0(\mathbf{r}) \right] + i k_{\text{vac}} e^{i(k_{\text{vac}} R(\mathbf{r}, t) - \omega t)} \left[ 1 + \chi(\mathbf{r}) \right] \nabla R(\mathbf{r}) \cdot \mathbf{E}_0(\mathbf{r}) = 0
\]

Canceling the common exponential term, using \( k_{\text{vac}} = 2\pi / \lambda_{\text{vac}} \), and a little algebra then gives

\[
-ik_{\text{vac}} \nabla \cdot \left[ (1 + \chi(\mathbf{r})) \mathbf{E}_0(\mathbf{r}) \right] + \nabla R(\mathbf{r}) \cdot \mathbf{E}_0(\mathbf{r}) = 0
\]

In the limit of very short wavelength, this becomes

\[
\nabla R(\mathbf{r}) \cdot \mathbf{E}_0(\mathbf{r}) \rightarrow 0 \tag{9.75}
\]
Finally, compute the time average of the Poynting vector

\[ S = \frac{1}{\mu_0} \text{Re}\{E(r, t)\} \times \text{Re}\{B(r, t)\} \]

\[ = \frac{1}{4\mu_0} [E(r, t) + E^*(r, t)] \times [B(r, t) + B^*(r, t)] \]

You will need to employ expressions (9.74) and (9.75), as well as the BAC-CAB rule (see P 0.4).

---

**Exercises for 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](image)

**Exercises for 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:

\[
\begin{bmatrix}
A & B \\
C & D \\
\end{bmatrix}
= \begin{bmatrix}
\frac{1}{n_1} & \frac{d}{R_1} \\
\frac{n_2}{n_1} - 1 & \frac{1}{n_2} \\
\end{bmatrix}
+ \begin{bmatrix}
\frac{1}{n_1} & \frac{d}{R_2} \\
\frac{n_2}{n_1} - 1 & \frac{1}{n_2} \\
\end{bmatrix}
\]

\[
\frac{1}{n_1} & \frac{d}{R_1} \\
\frac{n_2}{n_1} - 1 & \frac{1}{n_2} \\
\end{bmatrix}
+ \begin{bmatrix}
\frac{1}{n_1} & \frac{d}{R_2} \\
\frac{n_2}{n_1} - 1 & \frac{1}{n_2} \\
\end{bmatrix}
\]

Exercises for 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 \( d_o = R \), \( d_o = R/2 \), \( d_o = R/4 \), and \( d_o = -R/2 \) (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 \( d_o = 2f \), \( d_o = f \), \( d_o = f/2 \), and \( d_o = -f \) (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 \( R/2 \).

Exercises for 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](image-url)
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 \( d = 5 \text{ cm} \), \( R_1 = 5 \text{ cm} \), \( R_2 = -10 \text{ cm} \), \( n = 1.5 \). 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 \( f_{\text{eff}} \) for this system?

![Figure 9.20](image)

**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](image)

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 compound system shown in Fig. 9.22, known as the “Tessar lens.”
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, \( R_1 = \infty \), \( R_2 = 1.920 \), \( n = 1.5123 \)), 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.

Exercises for 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 9.22

Exercises for 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 equations1 (see P 10.2). Near $R = 0$, this type of wave

1For 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
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 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

$$E(x, y, z = d) = -\frac{i}{\lambda} \int \int_{\text{aperture}} E(x', y', z = 0) \frac{e^{i k R}}{R} d x' d y'$$

where

$$R = \sqrt{(x - x')^2 + (y - y')^2 + d^2}$$

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

$$E(x, y, z = d) = -\frac{i}{\lambda} \int \int_{\text{aperture}} E(x', y', z = 0) \frac{e^{i k R}}{R} \left[ \frac{1 + \cos(r, \hat{z})}{2} \right] d x' d y'$$
The additional factor in square brackets is known as \textit{obliquity factor} \((\cos(r, \hat{z}))\) indicates the cosine of the angle between \(r\) and \(\hat{z}\). 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 \(-\hat{z}\) 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 \(\omega\). The light field can be represented by \(\mathbf{E}(\mathbf{r}) e^{-i\omega t}\) which must obey the wave equation

\[
\nabla^2 \mathbf{E}(\mathbf{r}) e^{-i\omega t} - \frac{n^2}{c^2} \mathbf{E}(\mathbf{r}) \frac{\partial^2 e^{-i\omega t}}{\partial t^2} = 0
\]

(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

\[
\nabla^2 \mathbf{E}(\mathbf{r}) + k^2 \mathbf{E}(\mathbf{r}) = 0
\]

(10.5)

where \(k \equiv n\omega/c\) is the magnitude of the usual wave vector. Equation (10.5) is called the \textit{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 \(\mathbf{E}(\mathbf{r})\).

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 \textit{scalar Helmholtz equation}:

\[
\nabla^2 E(\mathbf{r}) + k^2 E(\mathbf{r}) = 0
\]

(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

\[
\left( \frac{\partial^2 \tilde{E}}{\partial x^2} + \frac{\partial^2 \tilde{E}}{\partial y^2} + 2ik \frac{\partial \tilde{E}}{\partial z} \right) \tilde{E} \approx 0 \tag{10.8}
\]

the solution to which is (see P10.5)

\[
\tilde{E}(x, y, z) \equiv -\frac{i}{\lambda z} \int_{-\infty}^{\infty} \tilde{E}(x', y', 0)e^{i \frac{k}{2} \left[ (x-x')^2 + (y-y')^2 \right]} dx' dy' \tag{10.9}
\]

The field is then given by

\[
E(x, y, z) = \tilde{E}(x, y, z)e^{ikz} \approx -\frac{i}{\lambda z} \int_{-\infty}^{\infty} \tilde{E}(x', y', 0)e^{i \frac{k}{2} \left[ z + (x-x')^2 + (y-y')^2 \right]} dx' dy' \tag{10.10}
\]

This equation agrees with the Fresnel-Kirchhoff formula (10.1) to the extent that \( R \equiv z \) in the denominator of the integral and

\[
R \equiv z + \frac{(x-x')^2 + (y-y')^2}{2z}
\]
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 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.

Figure 10.3 Aperture comprised of the region between a circle and a square.
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 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 $[1 + \cos(r, \hat{z})]/2$ 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:

\[ R \approx d \quad \text{(denominator only; paraxial approximation)} \quad (10.11) \]

This approximation is wholly inappropriate in the exponent of (10.3) since small changes in \( R \) can result in dramatic variations in \( e^{i k R} \). 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

\[
R = \sqrt{(x - x')^2 + (y - y')^2 + d^2} \\
= d \sqrt{1 + \frac{(x - x')^2 + (y - y')^2}{d^2}} \\
\approx d \left[ 1 + \frac{(x - x')^2 + (y - y')^2}{2d^2} + \ldots \right] \quad \text{(paraxial approximation)} \quad (10.12)
\]

Substitution of (10.11) and (10.12) into the Fresnel-Kirchhoff diffraction formula (10.3) and (10.2) yields

\[
E(x, y, d) \equiv -\frac{i e^{i k d}}{\lambda d} e^{i \frac{k}{2d} (x'^2 + y'^2)} \int_{\text{aperture}} E(x', y', 0) e^{i \frac{k}{2d} (x'^2 + y'^2)} e^{-i \frac{k}{d} (xx' + yy')} \, dx' \, dy' \\
\quad \quad (10.13)
\]

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 \( E(x', y', 0) e^{i \frac{k}{2d} (x'^2 + y'^2)} \). 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 \textit{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 \textit{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 propagate (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...
Figure 10.5 “The Fraunhofer Approximation” by Sterling Cornaby
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:

$$e^{i \frac{k}{2d} (x'^2 + y'^2)} \approx 1 \quad \text{(far field)} \quad (10.14)$$

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

$$d \gg \frac{k}{2} (\text{aperture radius})^2 \quad \text{(condition for far field)} \quad (10.15)$$

By substituting (10.14) into (10.13), the Fraunhofer approximation yields

$$E(x, y, d) \equiv -i e^{i k d} e^{i \frac{k}{2d} (x^2 + y^2)} \int \int_{\text{aperture}} E(x', y', 0) e^{-i \frac{k}{2d} (xx' + yy')} \, dx' \, dy' \quad (10.16)$$

As students will no doubt appreciate, the removal of \(e^{i \frac{k}{2d} (x'^2 + y'^2)}\) 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 \(E(x', y', 0)\).

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

$$E(x', y', z = 0) = E(\rho', z = 0) \quad (10.17)$$

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 \(\rho'\). The Fresnel diffraction integral (10.13) in this situation is given by

$$E(\rho, z = d) = -i e^{i k d} e^{i \frac{k}{2d} \rho^2} \int_{0}^{2\pi} d\theta' \int_{\text{aperture}} \rho' \, d\rho' \, E(\rho', z = 0) e^{i \frac{k}{2d} \rho^2} e^{-i \frac{k}{2d} [(\rho \cos \theta)(\rho' \cos \theta') + (\rho \sin \theta)(\rho' \sin \theta')]} \quad (10.18)$$
where
\[
x = \rho \cos \theta \\
y = \rho \sin \theta \\
x' = \rho' \cos \theta' \\
y' = \rho' \sin \theta'
\]
(10.19)

Notice that in the exponent of (10.18) we have
\[
\rho' \rho [\cos \theta' \cos \theta + \sin \theta' \sin \theta] = \rho' \rho \cos (\theta' - \theta)
\]
(10.20)

With this simplification, the diffraction formula (10.18) can be written as
\[
E(\rho, z = d) = -i \frac{e^{ikd}}{\lambda d} \int_{\text{aperture}} \rho' d\rho' E(\rho', z = 0) e^{i \frac{k\rho'^2}{2\sigma}} \int_0^{2\pi} d\theta' e^{-i \frac{k\rho \rho'}{d} \cos(\theta - \theta')}
\]
(10.21)

We are able to perform the integration over \( \theta \) with the help of the formula (0.54)
\[
\int_0^{2\pi} e^{-i \frac{k\rho \rho'}{d} \cos(\theta - \theta')} d\theta' = 2\pi J_0 \left( \frac{k\rho \rho'}{d} \right)
\]
(10.22)

where \( J_0 \) is called the zero-order Bessel function. Equation (10.21) then reduces to
\[
E(\rho, z = d) = -\frac{2\pi i e^{ikd} e^{i \frac{k\rho'^2}{2\sigma}}}{\lambda d} \int_{\text{aperture}} \rho' d\rho' E(\rho', z = 0) e^{i \frac{k\rho'^2}{2\sigma}} J_0 \left( \frac{k\rho \rho'}{d} \right)
\]
(10.23)

(Fresnel approximation with cylindrical symmetry)

The integral in (10.23) is called a Hankel transform on \( E(\rho', z = 0) e^{i \frac{k\rho'^2}{2\sigma}} \).

In the case of the Fraunhofer approximation, the diffraction integral becomes a Hankel transform on just the field \( E(\rho', z = 0) \) since \( \exp \left( i \frac{k\rho \rho'}{2\sigma} \right) \) goes to one. Under cylindrical symmetry, the Fraunhofer approximation is
\[
E(\rho, z = d) = -\frac{2\pi i e^{ikd} e^{i \frac{k\rho'^2}{2\sigma}}}{\lambda d} \int_{\text{aperture}} \rho' d\rho' E(\rho', z = 0) J_0 \left( \frac{k\rho \rho'}{d} \right)
\]
(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.

© 2004-2009 Peatross and Ware
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

\[
\mathbf{E}(\mathbf{r}) = \mathbf{r} \times \nabla E_{\text{scalar}}(\mathbf{r})
\]  

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

\[
E_{\text{scalar}}(\mathbf{r}) = r_0 E_0 \left(1 - \frac{i}{kr} \right) e^{ikr/r} \cos \theta \tag{10.27}
\]

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

\[
\mathbf{E}(\mathbf{r}) = -\hat{\phi} r_0 E_0 \left(1 - \frac{i}{kr} \right) e^{ikr/r} \sin \theta \tag{10.28}
\]

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 \( \theta \equiv \pi/2 \) (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):

\[
\oint_S \left[ U \frac{\partial V}{\partial n} - V \frac{\partial U}{\partial n} \right] d\mathbf{a} = \int_V \left[ U \nabla^2 V - V \nabla^2 U \right] d\mathbf{v} \tag{10.29}
\]

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

\[
V \equiv e^{ikr/r} \\
U \equiv E(\mathbf{r})
\]  

© 2004-2009 Peatross and Ware
where $E(r)$ is assumed to satisfy the scalar Helmholtz equation, (10.6). When these functions are used in Green’s theorem (10.29), we obtain

$$\oint_S \left[ E \frac{\partial}{\partial n} e^{ikr} - \frac{e^{ikr}}{r} \partial E \right] d\alpha = \int_V \left[ E \nabla^2 e^{ikr} - \frac{e^{ikr}}{r} \nabla^2 E \right] d\nu$$

(10.31)

The right-hand side of this equation vanishes (as long as we exclude the point $r = 0$; see P 0.5 and P 0.6) since we have

$$E \nabla^2 e^{ikr} - \frac{e^{ikr}}{r} \nabla^2 E = k^2 E e^{ikr} + \frac{e^{ikr}}{r} k^2 E = 0$$

(10.32)

where we have taken advantage of the fact that $E(r)$ 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

$$\oint_S \left[ E \frac{\partial}{\partial n} e^{ikr} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] d\alpha = 0$$

(10.33)

Now consider a volume between a small sphere of radius $\epsilon$ 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

$$\oint_{S_2} \left[ E \frac{\partial}{\partial n} e^{ikr} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] d\alpha = -\oint_{S_1} \left[ E \frac{\partial}{\partial n} e^{ikr} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] d\alpha$$

(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 \( \epsilon \) is small so that \( E(r) \) is approximately the same everywhere on the surface \( S_1 \). Then the integral over \( S_1 \) becomes

\[
\oint_{S_1} \left[ \frac{E}{r} \frac{\partial}{\partial n} e^{ikr} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] \, da = \lim_{r=\epsilon \to 0} 2\pi \int_0^\pi \int_0^{2\pi} \left[ E \left( \frac{\partial}{\partial r} e^{ikr} \frac{\partial r}{\partial n} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial r} \frac{\partial r}{\partial n} \right) \right] r^2 \sin \theta \, d\theta 
\]

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 \( \hat{n} \), the normal derivative \( \partial r / \partial n \) is always equal to \(-1\). (From the definition of the normal derivative we have \( \partial r / \partial n = \nabla r \cdot \hat{n} = -\hat{n} \cdot \hat{n} = -1\).) We can now perform the integration in (10.35) as well as take the limit as \( \epsilon \to 0 \) to obtain

\[
\lim_{\epsilon \to 0} \oint_{S_1} \left[ \frac{E}{r} \frac{\partial}{\partial n} e^{ikr} - \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} \right] \, da = -4\pi \lim_{\epsilon \to 0} \left[ r^2 \left( -\frac{e^{ikr}}{r^2} + ike^{ikr} \right) E - r^2 \frac{e^{ikr}}{r} \left( \frac{\partial E}{\partial r} \right) \right]_{r=\epsilon} 
\]

\[
= -4\pi \lim_{\epsilon \to 0} \left[ (-e^{ikr} + ike^{ikr} - e^{ikr}) r^2 \frac{e^{ikr}}{r} \frac{\partial E}{\partial r} \right]_{r=\epsilon} 
\]

\[
= 4\pi E(0) \quad \text{(10.36)}
\]

With the aid of (10.36), Green’s theorem applied to our specific geometry (10.34) reduces to

\[
E(0) = \frac{1}{4\pi} \oint_{S_2} \left[ \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} - E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} \right] \, da \quad \text{(10.37)}
\]

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 \( \hat{n} \) 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 saves a lot of work.) Thus, we are left with only the integration over the open aperture:

\[
E(0) = \frac{1}{2\pi} \int_{\text{aperture}} \left[ \frac{e^{ikr}}{r} \frac{\partial E}{\partial n} - E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} \right] \, da
\]  \hspace{1cm} (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:

\[
\frac{\partial}{\partial n} \frac{e^{ikr}}{r} = \left( \frac{\partial}{\partial r} \frac{e^{ikr}}{r} \right) \frac{\partial r}{\partial n} = \frac{ik}{r} \left( \frac{1}{r^2} - 1 \right) e^{ikr} \cos(\mathbf{r}, \hat{n}) \rightarrow \frac{ik e^{ikr}}{r} \cos(\mathbf{r}, \hat{n}) \quad \text{as} \quad r \gg \lambda
\]  \hspace{1cm} (10.39)

where \( \partial r / \partial n = \cos(\mathbf{r}, \hat{n}) \) indicates the cosine of the angle between \( \mathbf{r} \) and \( \hat{n} \). 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 \( E \equiv \tilde{E}(x, y) e^{ikz} \). 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

\[
\frac{\partial E}{\partial n} = \frac{\partial E}{\partial z} \frac{\partial z}{\partial n} = ik \tilde{E}(x, y) e^{ikz} (-1) = -ik \tilde{E}
\]  \hspace{1cm} (10.40)

Substituting (10.39) and (10.40) into (10.38) yields

\[
E(0) = -\frac{i}{\lambda} \int_{\text{aperture}} \frac{E e^{ikr}}{r} \left[ \frac{1 + \cos(\mathbf{r}, \hat{n})}{2} \right] \, da
\]  \hspace{1cm} (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

$$E(x, y, z = d) = -\frac{i}{\lambda} \iint_{\text{aperture}} E(x', y', z = 0) e^{ikR/R} \left[ \frac{1 + \cos(r, \hat{z})}{2} \right] dx' dy'$$  \hspace{1cm} (10.42)

where

$$R = \sqrt{(x - x')^2 + (y - y')^2 + d^2}$$  \hspace{1cm} (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 $[1 + \cos(r, \hat{z})]/2$.

### Appendix 10.C  Green's Theorem

To derive Green's theorem, we begin with the divergence theorem (see (0.11)):

$$\oint_S \mathbf{f} \cdot \hat{n} \, da = \int_V \nabla \cdot \mathbf{f} \, dv$$  \hspace{1cm} (10.44)

The unit vector $\hat{n}$ 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

$$\oint_S (U \nabla V) \cdot \hat{n} \, da = \int_V \nabla \cdot (U \nabla V) \, dv$$  \hspace{1cm} (10.45)

We recognize $\nabla V \cdot \hat{n}$ as the directional derivative of $V$ directed along the surface normal $\hat{n}$. This is often represented in shorthand notation as

$$\nabla V \cdot \hat{n} = \frac{\partial V}{\partial n}$$  \hspace{1cm} (10.46)

The argument of the integral on the right-hand side of (10.45) can be expanded with the chain rule:

$$\nabla \cdot (U \nabla V) = \nabla U \cdot \nabla V + U \nabla^2 V$$  \hspace{1cm} (10.47)

With these substitutions, (10.45) becomes

$$\oint_S U \frac{\partial V}{\partial n} \, da = \int_V \left[ \nabla U \cdot \nabla V + U \nabla^2 V \right] \, dv$$  \hspace{1cm} (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:

\[
\oint_{S} V \frac{\partial U}{\partial n} \, da = \int_{V} \left[ \nabla V \cdot \nabla U + V \nabla^2 U \right] \, dv \quad (10.49)
\]

We simply subtract (10.49) from (10.48), and this leads to (10.29) known as Green’s theorem.
Exercises

Exercises for 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 \( n_i = 1 \) and \( n_t = 2 \).

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 \( \lambda = \frac{\lambda_{\text{vac}}}{n} \). 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 \( \omega \).

(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:

\[
E(r) = \frac{A}{r} \cos(kr - \omega t) \hat{\phi}
\]

Show that this choice for \( E \) 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

\[
E(r, \theta) = \frac{A \sin \theta}{r} \left[ \cos(kr - \omega t) - \frac{1}{kr} \sin(kr - \omega t) \right] \hat{\phi}
\]
does satisfy Maxwell’s equations. Describe how this wave behaves as a function of \( r \) and \( \theta \). What conditions need to be satisfied for this equation to reduce to the spherical wave formula used in the diffraction formulas?

**Exercises for 10.2 Scalar Diffraction**

**P10.3** Show that \( E(r) = E_0 r e^{ikr} / r \) is a solution to the scalar Helmholtz equation (10.6).

HINT:

\[
\nabla^2 \psi = \frac{1}{r} \frac{\partial^2 (r \psi)}{\partial r^2} + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \psi}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 \psi}{\partial \phi^2}
\]

**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 \( \ell \). Find the intensity of the light on axis (i.e. \( x, y = 0 \)).

HINT: The integral takes on the following form:

\[
E(0, 0, d) = -\frac{i}{\lambda} \int_\text{aperture} \int E(x', y', 0) \frac{e^{ik \sqrt{x'^2 + y'^2 + d^2}}}{\sqrt{x'^2 + y'^2 + d^2}} \, dx' \, dy'
\]

\[
= -\frac{i E_0}{\lambda} \int_0^{\ell/2} d\theta' \int_0^{\sqrt{\rho'^2 + d^2}} e^{ik \sqrt{\rho'^2 + d^2}} \, d\rho' \, d\rho
\]

Then you will want to make the following change of variables: \( \xi = \sqrt{\rho'^2 + d^2} \). This will make it easier to accomplish the integration.

Answer: \( I(0, 0, d) = 2I_0 \left[ 1 - \cos \left( k \sqrt{(\ell/2)^2 + d^2} - kd \right) \right] \).

**Exercises for 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](image_url)

**Exercises for 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.

**Exercises for 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 \( \ell = 10 \mu m \) and \( \lambda = 500 \text{ nm} \). To see the result better, use a log scale on the \( z \)-axis.

Answer:
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.

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

$$E(x', y', 0) = E_0 e^{-\frac{x'^2 + y'^2}{w_0^2}} \quad (11.1)$$

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):

$$I(x', y', 0) = I_0 e^{-2\rho'^2/w_0^2} \quad (11.2)$$

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):

$$E(x, y, z) = -i e^{ikz} e^{i \frac{k}{i2} (x^2 + y^2)} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \left[ E_0 e^{-\left(x'^2 + y'^2\right)/w_0^2}\right] e^{i \frac{k}{2z} (x'^2 + y'^2)} e^{-i \frac{k}{2z} (xx' + yy')} \quad (11.3)$$

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

---

**Figure 11.1** Diffraction of a Gaussian field profile.
region to a radius on the scale of \( w_0 \). Equation (11.3) can be rewritten as
\[
E(x, y, z) = -i \frac{E_0 e^{ikz} e^{i \frac{k}{2w_0} (x^2 + y^2)}}{\lambda z} \int_{-\infty}^{\infty} dx' e^{-\left(\frac{x^2}{w_0^2} - i \frac{k}{2z} x'\right)} \int_{-\infty}^{\infty} dy' e^{-\left(\frac{y^2}{w_0^2} + i \frac{k}{2z} y'\right)}
\]  
(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
\[
\int_{-\infty}^{\infty} dx' e^{-\left(\frac{x^2}{w_0^2} - i \frac{k}{2z} x'\right)} = \frac{\pi}{\lambda z} \exp \left(\frac{-k x^2}{2z} \right) \left(\frac{1}{1 + \left(\frac{2\pi}{kw_0^2}\right)^2 e^{\frac{2z}{kw_0^2}}}\right)^{\frac{1}{2}}
\]
(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
\[
E(x, y, z) = E_0 e^{ikz} e^{i \frac{k}{2w_0} (x^2 + y^2)} \frac{1}{\sqrt{1 + \left(\frac{2\pi}{kw_0^2}\right)^2}} e^{i \tan^{-1} \frac{2\pi}{kw_0^2}}
\]
(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
\[
E(\rho, z) = -\frac{2\pi i e^{ikz} e^{i \frac{2\pi}{kw_0^2} \rho^2}}{\lambda z} \int_0^\infty \rho' d\rho' E_0 e^{-\rho'^2/\rho_0^2} e^{i \frac{2\pi}{\rho_0^2} \rho'^2} J_0 \left(\frac{k \rho'}{\rho_0} \frac{\rho'}{z}\right)
\]
(11.7)
We can use the integral formula (0.56) to obtain
\[
E(\rho, z) = -i \frac{2\pi E_0 e^{ikz} e^{i \frac{2\pi}{kw_0^2} \rho^2}}{\lambda z} e^{-\left(\frac{\rho^2}{\rho_0^2} - i \frac{k}{2z}\right)}
\]
(11.8)
which is identical to (11.6).

### 11.3 Gaussian Laser Beams

The rather complicated Gaussian field expression (11.6) can be cleaned up through the judicious introduction of new quantities:

\[
E(\rho, z) = E_0 \frac{w_0}{w(z)} e^{-\frac{\rho^2}{w^2(z)}} e^{ikz+i\frac{2k\rho^2}{2R(z)} - i\tan^{-1}\frac{z}{z_0}}
\]

(11.9)

where

\[
\rho^2 \equiv x^2 + y^2, \quad (11.10)
\]

\[
w(z) \equiv w_0\sqrt{1 + z^2/z_0^2}, \quad (11.11)
\]

\[
R(z) \equiv z + z_0^2/z, \quad (11.12)
\]

\[
z_0 \equiv \frac{kw_0^2}{2} \quad (11.13)
\]

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:

\[
I(\rho, z) = I_0 \frac{w_0^2}{w^2(z)} e^{-\frac{2\rho^2}{w^2(z)}} = \frac{I_0}{1 + z^2/z_0^2} e^{-\frac{2\rho^2}{w^2(z)}}
\]

(11.14)

![Figure 11.2](image-url) A Gaussian laser field profile in the vicinity of its beam waist.

© 2004-2009 Peatross and Ware
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 \( z_0 \).

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 \( \exp\{ikz + ikp^2/2R(z)\} \) 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 \( R(z) \approx z \) (see (11.12)). In this case, we may write these phase terms as \( kz + k\rho^2/2R(z) \approx k\sqrt{\rho^2 + z^2} \). This describes a spherical wave front emanating from the origin out to point \((\rho, z)\). 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 \( \exp(-i\tan^{-1}z/z_0) \) 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 \( \pi \).

### 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

\[
I(x, y, z) = \frac{1}{2}c\epsilon_0 \left| \int E(x', y', 0) e^{-ik(x'^2 + y'^2)/2z} dx' dy' \right|^2
\]  

(11.15)

Notice that the dependence of the diffraction on \( x, y, \) and \( z \) comes only through the combinations \( \theta_x \equiv x/z \) and \( \theta_y \equiv y/z \). Therefore, the diffraction pattern in the Fraunhofer limit is governed by the two angles \( \theta_x \) and \( \theta_y \), 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.

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))

\[
\frac{1}{f} = \frac{1}{z-L} + \frac{1}{d_i}.
\] (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. \( z \rightarrow \infty \)) we see that the image of the Fraunhofer pattern must appear at the focus of the lens:

\[
d_i \equiv f.
\] (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.
11.4 Fraunhofer Diffraction Through a Lens

Figure 11.4 Imaging of the Fraunhofer diffraction pattern to the focus of a lens.

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

\[ M = \frac{d_i}{-(z-L)} \]  

(11.18)

Taking the limit of very large \( z \) and employing (11.17), the magnification becomes

\[ M \to \frac{f}{z} \]  

(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

\[ I(x, y, L+f) \approx \frac{1}{2} \varepsilon_0 \frac{1}{f} \int \int_{\text{aperture}} E(x', y', 0) e^{-i\frac{\lambda}{2}(xx'+yy')} dxdy' \]  

(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 \( \ell_1 \) and \( \ell_2 \) 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 \((1 - n)(\ell_1 + \ell_2)\) (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

\[
\Delta \phi = -k(n-1)(\ell_1 + \ell_2). \tag{11.21}
\]

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 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 \( \ell_1 \) and \( \ell_2 \) from the equations describing the spherical surfaces of the lens:

\[
\begin{align*}
(R_1 - \ell_1)^2 + x^2 + y^2 &= R_1^2 \\
(R_2 + \ell_2)^2 + x^2 + y^2 &= R_2^2
\end{align*}
\tag{11.22}
\]

In the Fresnel approximation, the light propagation takes place in the paraxial limit. It is therefore appropriate to neglect the terms \( \ell_1^2 \) and \( \ell_2^2 \) in comparison with the other terms present. Within this approximation, equations (11.22) can
be solved, and they render

\[ \ell_1 \approx \frac{x^2 + y^2}{2R_1} \]
\[ \ell_2 \approx -\frac{x^2 + y^2}{2R_2} \]

(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

\[ \Delta \phi = -k(n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\frac{(x^2 + y^2)}{2} \]

(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

\[ \Delta \phi = -\frac{k}{2f}(x^2 + y^2) \]

(11.25)

In summary, the light traversing a lens experiences a relative phase shift given by

\[ E(x, y, z) = E(x, y, z) \text{ before lens} e^{i \frac{k}{2f}(x^2 + y^2)} \]

(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.

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 \( E(x', y', 0) \) at the aperture is known. We use the Fresnel approximation to compute the field incident on the lens:

\[ E(x'', y'', L) = -i \frac{e^{ikL} e^{i \frac{k}{\lambda L} (x''^2 + y''^2)}}{\lambda L} \int \int E(x', y', 0) e^{i \frac{k}{\lambda L} (x'^2 + y'^2)} e^{-i \frac{k}{\lambda L} (x''x' + y''y')} dx' dy' \]

(11.27)
Chapter 11 Diffraction Applications

(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:

\[
E(x, y, L + f) = -i \frac{e^{ikf} e^{i \frac{k}{2\lambda f} (x^2 + y^2)}}{\lambda f} \int \int \left[ E(x'', y'', L) e^{-i \frac{k}{2f} (x''^2 + y''^2)} \right] \times e^{i \frac{k}{2f} (x''^2 + y''^2)} e^{-i \frac{k}{f} (xx'' + yy'')} \, dx'' \, dy''
\]

(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

\[
E(x, y, L + f) = -i \frac{e^{ik(L+f)} e^{i \frac{k}{2\lambda f} (x^2 + y^2)}}{\lambda f} \int \int E(x', y', 0) e^{-i \frac{k}{f} (xx' + yy')} \, dx' \, dy'
\]

(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 carries 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:

\[
E(x', y', L + f) \big|_{L=f} = -i \frac{e^{2ikf}}{\lambda f} \int \int E(x', y', 0) e^{-i \frac{k}{f} (xx' + yy')} \, dx' \, dy'
\]

(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 iris 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

\[ I(\rho, f) = I_0 \left( \frac{\pi \ell^2}{4 \lambda f} \right)^2 \left[ 2 \frac{J_1 \left( \frac{k \ell \rho}{2f} \right)}{\left( k \ell \rho / 2f \right)} \right]^2 \]  

(11.31)

where \( f \) is the focal length of the lens and \( \ell \) 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 \( 1.22 \pi \).

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...
2\(J_1(x)/x\) 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. 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, 1801–1892) 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 \(\ell\) 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 1.22\(\pi\), the location of the first zero:

\[
\frac{k\ell\rho}{2f} = 1.22\pi \tag{11.32}
\]
With a little rearranging we have

$$\theta_{\text{min}} \approx \frac{\rho}{f} = \frac{1.22 \lambda}{\ell}$$  \hspace{1cm} (11.33)$$

Here we have associated the ratio $\rho/f$ (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 $\theta_{\text{min}}$.

### 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

$$E_{\text{aperture}}(x', y', 0)$$  \hspace{1cm} (11.34)$$

Each identical aperture has a unique location on the mask. Let the location of the $n^{\text{th}}$ aperture be designated by the coordinates $(x'_n, y'_n)$. Since each aperture is
identical, we can conveniently write the total field by summing over the same individual pattern displaced repeatedly according to the locations of the individual apertures:

\[ E(x', y', 0) = \sum_{n=1}^{N} E_{\text{aperture}}(x' - x'_n, y' - y'_n, 0) \]  

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

\[ E(x, y, z) = -i \frac{e^{i k z} e^{i \frac{1}{2} \pi (x^2 + y^2)}}{\lambda z} \sum_{n=1}^{N} \int_{-\infty}^{\infty} d x' \int_{-\infty}^{\infty} d y' E_{\text{aperture}}(x' - x'_n, y' - y'_n, 0) e^{-i \frac{1}{k z} (x x' + y y')} \]  

where we have taken the summation out in front of the integral.

To proceed further, let us make the following change of variables:

\[ x'' \equiv x' - x'_n \]
\[ y'' \equiv y' - y'_n \]

With the use of these new variables, (11.36) becomes

\[ E(x, y, z) = -i \frac{e^{i k z} e^{i \frac{1}{2} \pi (x^2 + y^2)}}{\lambda z} \sum_{n=1}^{N} \int_{-\infty}^{\infty} d x'' \int_{-\infty}^{\infty} d y'' E_{\text{aperture}}(x'', y'', 0) e^{-i \frac{1}{k z} [x(x'' + x'_n) + y(y'' + y'_n)]} \]  

© 2004-2009 Peatross and Ware
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

$$E(x, y, z) = \left[ \sum_{n=1}^{N} e^{-i k z (x x'_{n} + y y'_{n})} \right]$$

$$\times \left[ -i \frac{e^{ikz} e^{i \frac{z}{\lambda}} (x^{2} + y^{2})}{\lambda z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_{aperture}(x', y', 0) e^{-i \frac{z}{\lambda} (x x' + y y')} \right]$$

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 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

$$E_{\text{aperture}}(x, y, z) = -i E_0 \Delta x \Delta y e^{i k z} \frac{\Delta x \Delta y}{\lambda z} e^{i \frac{k z}{2z} (x^2 + y^2)} \sin\left(\frac{\pi \Delta x}{\lambda z} x\right) \sin\left(\frac{\pi \Delta y}{\lambda z} y\right)$$  \hspace{1cm} (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

$$x'_n = \left(n - \frac{N + 1}{2}\right) h, \quad y'_n = 0$$  \hspace{1cm} (11.41)

where $N$ is the total number of slits. Then the summation in the array theorem, (11.39), becomes

$$\sum_{n=1}^{N} e^{-i \frac{k z}{2z} (xx'_n + yy'_n)} = e^{i \frac{k h z}{2z} (\frac{N+1}{2})} \sum_{n=1}^{N} e^{-i \frac{k h z}{2z} n}$$  \hspace{1cm} (11.42)

This summation is recognized as a geometric sum, which can be performed using formula (0.59).

Equation (11.42) then simplifies to

$$\sum_{n=1}^{N} e^{-i \frac{k z}{2z} (xx'_n + yy'_n)} = e^{i \frac{k h z}{2z} (\frac{N+1}{2})} \frac{e^{-i \frac{k h z}{2z} N} - 1}{e^{-i \frac{k h z}{2z}} - 1}$$  

$$= \frac{\sin N \frac{k h z}{2z}}{\sin \frac{k h z}{2z}}$$  \hspace{1cm} (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

$$E(x, y, z) = \frac{\sin N \frac{k h z}{2z}}{\sin \frac{k h z}{2z}} \left[-i E_0 \frac{\Delta x \Delta y e^{i k z}}{\lambda z} e^{i \frac{k z}{2z} (x^2 + y^2)} \sin\left(\frac{\pi \Delta x}{\lambda z} x\right) \sin\left(\frac{\pi \Delta y}{\lambda z} y\right)\right]$$  \hspace{1cm} (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:

$$I(x) = I_{\text{peak}} \sin^2\left(\frac{\pi \Delta x}{\lambda z} x\right) \sin^2\left(\frac{\pi \Delta y}{\lambda z} y\right) \hspace{1cm} (11.45)$$

\(^1\)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 $\sin \left[ \left(\frac{\pi \Delta y}{\lambda z} y\right) \right]$ 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.
Note that \( \lim_{\alpha \to 0} \frac{\sin Na}{\sin \alpha} = N \) so we have placed \( N^2 \) in the denominator when introducing our definition of \( I_{\text{peak}} \), which represents the intensity on the screen at \( x = 0 \). In principle, the intensity \( I_{\text{peak}} \) 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 \( \Delta x \) 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.

### 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 \( \sin^2 \left( \frac{2 \pi h x}{\lambda z} \right)/\sin^2 \left( \frac{\pi h x}{\lambda z} \right) = 4 \cos^2 \left( \frac{\pi h x}{\lambda z} \right) \).

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 \( \sin(\pi h x/\lambda z) \) 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 \( N = 100 \).
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.
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

\[
\frac{\pi hx}{\lambda_0 z} = m \pi \Rightarrow \lambda_0 = \frac{hx}{mz}
\]  

(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 Fig. 11.12d) occurs when

\[
N \frac{k'hx}{2z} = N m \pi + \pi \Rightarrow \lambda_0 - \Delta \lambda = N \frac{hx}{(Nm + 1)z}
\]  

(11.47)

The wavelength difference that shifts the peak by this amount (from peak center to the adjacent zero) is then

\[
\Delta \lambda = \frac{hx}{mz} - \frac{Nhx}{(Nm + 1)z} = \frac{(Nm + 1) - Nm}{(Nm + 1)} \frac{hx}{mz} \equiv \frac{hx}{Nm^2z} = \frac{\lambda_0}{Nm}
\]  

(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

\[
RP \equiv \frac{\lambda}{\Delta \lambda} = mN
\]  

(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.

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 matrices. First, consider the matrix for propagation through a distance \( d \):

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix} =
\begin{bmatrix}
1 & d \\
0 & 1
\end{bmatrix}
\] (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 \( z' = z_0 \). 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):

\[
z' - iz'_0 = \frac{1(z - iz_0) + d}{0(z - iz_0) + 1} = z + d - iz_0 \quad \text{(propagation through distance } d)\]

(11.52)

Thus, the law holds in this case.

Next we consider the ABCD matrix of a thin lens (or a curved mirror):

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix} =
\begin{bmatrix}
1 & 0 \\
-1/f & 1
\end{bmatrix}
\] (11.53)

A beam that traverses a thin lens undergoes the phase shift \(-kp^2/2f\), according to (11.26). This modifies the original phase of the wave front \( kp^2/2R(z) \), seen in (11.9). The phase of the exiting beam is therefore

\[
\frac{kp^2}{2R(z')} = \frac{kp^2}{2R(z)} - \frac{kp^2}{2f}
\]

(11.54)

where we do not keep track of unimportant overall phases such as \( kz \) or \( k z' \). With (11.12) this relationship reduces to

\[
\frac{1}{R(z')} = \frac{1}{R(z)} - \frac{1}{f} \Rightarrow \frac{1}{z' + z'^2/2z'} = \frac{1}{z + z_0^2/2z_0} - \frac{1}{f}
\]

(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,

\[
w(z') = w(z) \Rightarrow z'_0\left(1 + \frac{z'^2}{2z'^2_0}\right) = z_0\left(1 + \frac{z^2}{2z_0^2}\right)
\]

(11.56)

On the other hand, the ABCD law for the thin lens gives

\[
z' - iz'_0 = \frac{1(z - iz_0) + 0}{-\left(1/f\right)(z - iz_0) + 1} \quad \text{(traversing a thin lens with focal length } f)\]

(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

\[
\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} = \begin{bmatrix} A_2A_1 + B_2C_1 & A_2B_1 + B_2D_1 \\ C_2A_1 + D_2C_1 & C_2B_1 + D_2D_1 \end{bmatrix}
\]

(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

\[
z'' - iz''_0 = \frac{A_2 \left( z' - iz'_0 \right) + B_2}{C_2 \left( z' - iz'_0 \right) + D_2}
\]

\[
= \frac{A_2 \left[ A_1 \left( z - iz_0 \right) + B_1 \right] + B_2}{C_2 \left[ A_1 \left( z - iz_0 \right) + B_1 \right] + D_2}
\]

\[
= \frac{A_2 \left[ A_1 \left( z - iz_0 \right) + B_1 \right] + B_2 \left[ C_1 \left( z - iz_0 \right) + D_1 \right]}{C_2 \left[ A_1 \left( z - iz_0 \right) + B_1 \right] + D_2 \left[ C_1 \left( z - iz_0 \right) + D_1 \right]}
\]

(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

Exercises for 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 $\phi$ is the relative phase between the two beams. The net intensity is then $I_f(\rho, z) \propto |E_1(\rho, z) + E_2 e^{ikz+i\phi}|^2$ or

$$I_f(\rho, z) = \left[ I_2 + I_1(\rho, z) + 2\sqrt{I_2 I_1(\rho, z)} \cos \left( \frac{kp^2}{2R(z)} \tan^{-1} \frac{z}{z_0} - \phi \right) \right]$$
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 \( \pi \), 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 \( \rho \), the only interference term that varies is \( k\rho^2/2R(z) \). If you count \( N \) fringes out to a radius \( \rho \), 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:
Exercises for 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 $\Delta x$ by $\Delta y$) illuminated by a plane wave $E_0$. 
HINT: Use (10.16)
\[
E(x, y, z) = -i E_0 \frac{e^{ikz}}{\lambda z} e^{i \frac{k}{2}(x^2 + y^2)} \int_{-\Delta x/2}^{\Delta x/2} dx' e^{-i \frac{k}{2} x'} \int_{-\Delta y/2}^{\Delta y/2} dy' e^{-i \frac{k}{2} y'}
\]

Answer: \[I(x, y, z) = I_0 \frac{\Delta x^2 \Delta y^2}{\lambda^2 z^2} \text{sinc}^2 \left( \frac{\pi \Delta x}{\lambda z x} \right) \text{sinc}^2 \left( \frac{\pi \Delta y}{\lambda z y} \right)\]

**P11.6** Calculate the Fraunhofer diffraction intensity pattern for a circular aperture (diameter \( \ell \)) illuminated by a plane wave \( E_0 \).

HINT: Use (10.24) and (0.55).

Answer: \[I(p, z) = I_0 \left( \frac{\pi \ell^2}{\lambda z} \right)^2 \frac{2 J_1(k \ell p/2z)}{k \ell p/2z} \]

The function \( \frac{2 J_1(x)}{x} \) (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 \( \pi \).

Note that \( \lim_{x \to 0} \frac{2 J_1(x)}{x} = 1 \).

**L11.7** Set up a collimated “plane wave” in the laboratory using a HeNe laser (\( \lambda = 633 \text{ nm} \)) and appropriate lenses.

(a) Choose a rectangular aperture (\( \Delta x \) by \( \Delta y \)) and place it in the plane wave. Observe the Fraunhofer diffraction on a very far away screen (i.e., where \( z \gg k \Delta x^2 / 2 \) 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 \( \ell \). Check the position of the first “zero.”

**Exercises for 11.5 Resolution of a Telescope**

**P11.8**

(a) What minimum telescope diameter would be required to distinguish a Jupiter-like planet (orbital radius \( 8 \times 10^8 \text{ km} \)) from its star if they are 10 light-years away? Take the wavelength to be \( \lambda = 500 \text{ nm} \). 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 $\lambda$ is incident on a circular aperture of diameter $\ell$ 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](image)

**Exercises for 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 “$\pi$” the light coming through is shifted to be $180^\circ$ out of phase with light coming through the other regions. (Thus the square is $180^\circ$ out of phase with the rest of the upper circle, and the left circle is $180^\circ$ out of phase with the right circle.) The distances $c$ and $L$ as indicated below are also given.
Exercises

(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.

Exercises for 11.7 Diffraction Grating

P11.13 Consider Fraunhofer diffraction from a grating of $N$ slits having widths $\Delta x$ 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 $h = 2\Delta x$, $\Delta x = 5 \mu m$, 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 $I_{\text{peak}}$ 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 $k h x / (2f) = \pm \pi$). (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 $\Delta x$ by $\Delta y$ (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.

---

**Exercises for 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 $\phi$ 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.

(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 500 nm to 700 nm?
of 500 nm? Do not worry about aberration such as astigmatism from using the mirrors off axis.

![Diagram of monochromator](image)

**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 \( \Delta x \) of the entrance and exit slits be to obtain this resolving power? Assume \( \lambda = 500 \text{ nm} \).

HINT: Set \( \Delta x \) to be the distance between the peak and the first zero of the diffraction pattern at the exit slit for monochromatic light.

**Exercises for 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

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}
= \begin{bmatrix}
1 & d/n \\
0 & 1
\end{bmatrix}
\]
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 $1/f = 1/d_o + 1/d_i$ 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 $\lambda$.

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.
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 $\ell$ 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 $d_i$.

Write a vector equation relating $\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$ to $\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$. 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 \(d_0\) 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 \(d_0\) 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 \(\theta_1\). Angular magnification \(M_0\) quantifies the telescope's purpose of enlarging the apparent angle between points in the field of view.

\[
\begin{bmatrix}
y_1 \\
\theta_1
\end{bmatrix}
\]

\[
\begin{bmatrix}
\begin{array}{c}
y_2 \\
\theta_2
\end{array}
\end{bmatrix}
\]

\[
f_1 \quad f_2
\]

**Figure 11.25**

Give a sensible definition for angular magnification in terms of \(\theta_1\) and \(\theta_2\). 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:

\[ E(x, y, d) \approx -\frac{ie^{ikd}e^{\frac{i}{\lambda d} (x^2 + y^2)}}{\lambda d} \int \int E(x', y', 0) e^{i\frac{k}{\lambda d} (x'^2 + y'^2)} e^{-i\frac{k}{\lambda d} (xx' + yy')} \, dx' \, dy' \]

\[ \int_{-\infty}^{\infty} e^{-Ax^2 + Bx + C} \, dx = \sqrt{\frac{\pi}{A}} e^{B^2/4A + C}. \]

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

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

where \( \cos \theta = \frac{1}{2} (A + D) \).

(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 \( R(z) \) 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.

\[
E(\rho, z) = E_0 \frac{w_0}{w(z)} e^{-\frac{\rho^2}{w_0^2(z)}} e^{ikz} e^{\frac{k^2z^2}{2R(z)}} e^{-i\tan^{-1}\frac{z}{z_0}}
\]

\[
\rho^2 \equiv x^2 + y^2
\]

\[
w(z) \equiv w_0 \sqrt{1 + z^2 / z_0^2}
\]

\[
R(z) \equiv z + z_0^2 / z
\]

\[
z_0 \equiv \frac{k w_0^2}{2}
\]

R73

(a) Compute the Fraunhofer diffraction intensity pattern for a uniformly illuminated circular aperture with diameter \( \ell \).

HINT:

\[
E(x, y, d) \approx -\frac{ie^{ikd}e^{i\frac{k}{\lambda d} (x^2 + y^2)}}{\lambda d} \int \int E(x', y', 0) e^{-i\frac{k}{\lambda d} (xx' + yy')} \, dx' \, dy'
\]

\[
J_0(\alpha) = \frac{1}{2\pi} \int_0^{2\pi} e^{\pm i\alpha \cos(\theta - \theta')} \, d\theta'
\]

© 2004-2009 Peatross and Ware
\[
\int_0^a J_0(bx) x \, dx = \frac{a}{b} J_1(ab)
\]

\[
J_1(1.22\pi) = 0
\]

\[
\lim_{x \to 0} \frac{2J_1(x)}{x} = 1
\]

(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 \(\lambda = 500 \text{ nm}\)? Compare with the radius of Earth’s orbit, \(1.5 \times 10^8 \text{ km}\).

R74

(a) Derive the Fraunhofer diffraction pattern for the field from a uniformly illuminated single slit of width \(\Delta x\). (Don’t worry about the \(y\)-dimension.)

(b) Find the Fraunhofer intensity pattern for a grating of \(N\) slits of width \(\Delta x\) 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 \(\sum_{n=1}^{N} e^{-i\frac{\pi}{\lambda} xx'_n}\) 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 \(\lambda = 500 \text{ nm}\).

Selected Answers

R72: (a) 100 cm (b) 0.32 mm.
R73: (b) \(4.8 \times 10^8 \text{ km}\).
R74: (c) \(5 \mu \text{m}\).
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.

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. 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 reconstruct 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 beamsplitter divides the fields evenly, so that the overall output intensity is given by

\[ I = \frac{1}{2} I_0 \]

where \( I_0 \) is the intensity of the incident light. The interferogram is then given by

\[ I = \frac{1}{2} I_0 (1 + 
\]
where \( \tau \) 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 \( \tau \) is varied.

What happens if one of the retro-reflecting mirrors is misaligned by a small angle \( \theta \)? 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 \( 2\theta \). This puts a relative phase term of

\[
\phi = k x \sin (2\theta_x) + k y \sin (2\theta_y)
\]

on the misaligned beam (in addition to \( \omega \tau \)). Here \( \theta_x \) represents the tilt of the mirror in the \( x \)-dimension and \( \theta_y \) represents the amount of tilt in the \( y \)-dimension.

When the two plane waves join, the resulting intensity pattern is

\[
I_{\text{det}} = 2I_0 [1 + \cos (\phi + \omega \tau)]
\]

Of course, the phase term \( \phi \) 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 \( \tau \) 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 the longer path is increased by the added amount of delay $\tau / c$. 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

$$\phi = k\rho^2 / 2R_1 - k\rho^2 / 2R_2$$

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.
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.

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 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.
Figure 12.4 Twyman-Green setup for testing lenses.
12.4 Generating Holograms

Consider a coherent monochromatic beam of light that is split in half by a beam-splitter, 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 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

$$E_{\text{film}}(\mathbf{r}) e^{-i\omega t} = E_{\text{object}}(\mathbf{r}) e^{-i\omega t} + E_{\text{ref}}(\mathbf{r}) e^{-i\omega t}$$  \hspace{1cm} (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.

![Exposure of holographic film](image_url)
The intensity of the field (12.5) is given by

\[ I_{\text{film}}(\mathbf{r}) = \frac{1}{2} c \varepsilon_0 \left[ |E_{\text{object}}(\mathbf{r})|^2 + |E_{\text{ref}}(\mathbf{r})|^2 + E_{\text{ref}}^*(\mathbf{r}) E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}(\mathbf{r}) E_{\text{object}}^*(\mathbf{r}) \right] \]  

(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 \( I_{\text{film}} \). 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

\[ T(\mathbf{r}) \propto I_{\text{film}}^2(\mathbf{r}) \]  

(12.7)

This means that a field that is later incident on the film has its amplitude modified by

\[ E_{\text{transmitted}}(\mathbf{r}) = t(\mathbf{r}) E_{\text{incident}}(\mathbf{r}) \propto I_{\text{film}}(\mathbf{r}) E_{\text{incident}}(\mathbf{r}) \]  

(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

\[ E_{\text{incident}}(\mathbf{r}) = E_{\text{ref}}(\mathbf{r}) \]  

(12.9)

and view the light that is transmitted. According to (12.6) and (12.8), the transmitted field is proportional to

\[ E_{\text{transmitted}}(\mathbf{r}) \propto I_{\text{film}}(\mathbf{r}) E_{\text{ref}}(\mathbf{r}) \]

\[ = \left[ |E_{\text{object}}(\mathbf{r})|^2 + |E_{\text{ref}}(\mathbf{r})|^2 \right] E_{\text{ref}}(\mathbf{r}) + |E_{\text{ref}}(\mathbf{r})|^2 E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}^2(\mathbf{r}) E_{\text{object}}^*(\mathbf{r}) \]  

(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.

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 $E_{\text{ref}}$. The field from the point object can be treated as a spherical wave:

$$E_{\text{object}}(\rho) = \frac{E_{\text{ref}} L}{\sqrt{L^2 + \rho^2}} e^{i k \sqrt{L^2 + \rho^2}} \quad \text{(point source example)} \quad (12.11)$$

Here $\rho$ 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 $E_{\text{ref}}$ 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),...
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.

becomes

\[
E_{\text{transmitted}}(\rho) \propto \left( \frac{E_{\text{ref}}^2 L^2}{L^2 + \rho^2} + E_{\text{ref}}^2 \right) E_{\text{ref}} + E_{\text{ref}}^2 \frac{E_{\text{ref}} L}{\sqrt{L^2 + \rho^2}} e^{i k \sqrt{L^2 + \rho^2}} + E_{\text{ref}}^2 \frac{E_{\text{ref}} L}{\sqrt{L^2 + \rho^2}} e^{-i k \sqrt{L^2 + \rho^2}} \quad (\text{point source example})
\]

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

Exercises for 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 \( \lambda = 633 \text{ nm} \)? What would happen if 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 \( \lambda = 633 \text{ nm} \)?

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 \( \theta \). Let the reference beam strike the film at normal incidence, and take the wavelength to be \( \lambda \).

(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 \( \rho \) and superimpose a similar graph for a “best-fit” mask that has regions of either 100% or 0% transmission. Use \( \lambda = 633 \text{ nm} \) 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 $I$ of radiation (including all frequencies) that flows outward from a blackbody radiator is given by

$$I = e \sigma T^4,$$  \hspace{2cm} (13.1)

where $\sigma$ 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 $\sigma = 5.6696 \times 10^{-8} \text{ W/m}^2\cdot\text{K}^4$. 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 $u_{\text{field}}$ (given by (2.51) in units of energy per volume). This connection between outward-going intensity and energy density of the field is given by

$$I = \frac{c u_{\text{field}}}{4} \Rightarrow u_{\text{field}} = e \frac{4\sigma T^4}{c}$$  \hspace{2cm} (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 $\hat{n}$. The average over the hemisphere is carried out as follows:

$$\frac{2\pi}{0} \int d\phi \int_0^{\pi/2} r \cdot \hat{n} \sin \theta d\theta = \frac{2\pi}{0} \int d\phi \int_0^{\pi/2} r \cos \theta \sin \theta d\theta = \frac{1}{2}$$  \hspace{2cm} (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 black-body 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, experimentalists 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 $k_B T/2$ for each degree of freedom, where $k_B = 1.38 \times 10^{-23} \text{J/K}$ is Boltzmann’s constant. Rayleigh and Jeans supposed that each unique mode of the electromagnetic field should carry energy $k_B T$ just as each mechanical spring in thermal equilibrium carries energy $k_B T$ ($k_B T/2$ as kinetic and $k_B T/2$ 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 $k_B T$ 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 $\nu$). 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 (0.33) 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

$$\text{Re} \left\{ \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} E_{n,m,\ell} e^{i(nk_0 x + mk_0 y + \ell k_0 z)} \right\}$$

where each component of the wave number in any of the three dimensions is
always an integer times

$$k_0 = \frac{2\pi}{L}$$  

(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 $\ell$:

$$n^2 + m^2 + \ell^2 = \left( \frac{k}{k_0} \right)^2$$  

(13.6)

We need to know how many more ways there are to choose $n$, $m$, and $\ell$ 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:

$$\text{# modes in } (k,k+dk) = \frac{4\pi k^2}{k_0^3} \frac{dk}{k_0}$$  

(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 $\ell$ 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 $\ell$) the field may travel in the forwards or the backwards direction.

Since according to the Rayleigh-Jeans assumption each mode carries energy of $k_0 T$, the energy density (energy per volume) associated with a specified range

\[\text{energy density} = \frac{\text{energy}}{\text{volume}}\]
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

\[
u_{\text{field}} = \int_0^\infty 2 \times \frac{k_B T}{L^3} \times \frac{4\pi k^2}{k^3} dk = k_B T \int_0^\infty \frac{k^2}{\pi^2} dk
\]  

(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

\[
u_{\text{field}} = k_B T \int_0^\infty \frac{8\pi \nu^2}{c^3} d\nu
\]  

(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

\[
u_{\text{field}} = \int_0^\infty \rho(\nu) d\nu
\]  

(13.10)

where

\[
\rho_{\text{Rayleigh-Jeans}}(\nu) = k_B T \frac{8\pi \nu^2}{c^3}
\]

(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

\[
\rho_{\text{Wien}}(\nu) = \frac{8\pi h\nu^3 e^{-\hbar \nu / k_B T}}{c^3}
\]  

(13.12)
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.

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 $\rho_{\text{Wien}}(\nu) = a\nu^3 e^{-b\nu/T}$, where $a$ and $b$ were parameters used to fit the data.

Wien’s formula did a good job of fitting experimental data. However, in 1900 Lummer and Pringsheim 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:

$$\rho_{\text{Planck}}(\nu) = \frac{8\pi h\nu^3}{c^3 \left[ e^{h\nu/k_B T} - 1 \right]}$$

(13.13)

where $h = 6.626 \times 10^{-34}$ J ⋅ 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 Pringsheim’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, \( k_B T \) 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 \( e^{-nh\nu/k_B T} \). 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:

\[
P_n = \frac{e^{-nh\nu/k_B T}}{\sum_{m=0}^{\infty} e^{-mh\nu/k_B T}} = e^{-nh\nu/k_B T} \left[ 1 - e^{-h\nu/k_B T} \right] \quad (13.14)
\]

Then, the energy in each mode of the field is expected to be

\[
\sum_{n=0}^{\infty} h\nu n P_n = h\nu \left[ 1 - e^{-h\nu/k_B T} \right] \sum_{n=0}^{\infty} n e^{-nh\nu/k_B T} = h\nu \left[ e^{-h\nu/k_B T} - 1 \right] \frac{\partial}{\partial (h\nu/k_B T)} \sum_{n=0}^{\infty} e^{-nh\nu/k_B T} = \frac{h\nu}{e^{h\nu/k_B T} - 1} \quad (13.15)
\]

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 \( \frac{8\pi\nu^2}{c^3} \), obtained in the derivation of the Rayleigh-Jeans formula. In other words, we substitute (13.15) for \( k_B T \) 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 field energy density, which is in thermal equilibrium with the blackbody radiator:

\[
u_{\text{field}} = \int_0^{\infty} \rho_{\text{Planck}}(v) dv = \frac{4}{c} \frac{2\pi^5 k_B^4}{15 c^2 h^3} T^4 = \frac{4}{c} \sigma T^4 \quad (13.16)
\]

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.
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.

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 \( \nu \), 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 \( h \nu \):

\[
\begin{align*}
\dot{N}_1 &= A_{21} N_2 - B_{12} \rho(\nu) N_1 + B_{21} \rho(\nu) N_2, \\
\dot{N}_2 &= -A_{21} N_2 + B_{12} \rho(\nu) N_1 - B_{21} \rho(\nu) N_2 \\
\end{align*}
\]

(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

\[
\rho(\nu) = \frac{A_{21}}{N_1/N_2 B_{12} - B_{21}}
\]

(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:

\[
\frac{N_1}{N_2} = \frac{e^{-E_1/k_B T}}{e^{-E_2/k_B T}} = e^{(E_2 - E_1)/k_B T} = e^{h \nu/k_B T}
\]

(13.19)

Then when equating (13.18) to the Planck blackbody spectral density (13.13) we get

\[
\frac{A_{21}}{e^{h \nu/k_B T} B_{12} - B_{21}} = \frac{8 \pi h \nu^3}{c^3 \left[ e^{h \nu/k_B T} - 1 \right]}
\]

(13.20)

From this expression we deduce that

\[
B_{12} = B_{21}
\]

(13.21)

and

\[
A_{21} = \frac{8 \pi h \nu^3}{c^3} B_{21}
\]

(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 spontaneous 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.

**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 $u_{\text{field}}$ (energy per volume) are equivalent to force per area, or in other words pressure. The radiation exerts a pressure of

$$P = u_{\text{field}}/3 \quad (13.23)$$

on each wall of the box. This can be derived from the fact that radiation of energy $\Delta E$ imparts a momentum

$$\Delta p = \frac{2\Delta E}{c} \cos \theta \quad (13.24)$$

when it is absorbed and reemitted from a wall at an angle $\theta$. 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

\[
P = \frac{\pi^{1/2} \int_0^{\pi/2} \frac{1}{A} \sin \theta \, d\theta}{\int_0^{\pi/2} \sin \theta \, d\theta}
\]  

(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

\[
dS = \frac{dQ}{T}
\]  

(13.26)

where \( dQ \) 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 \( dQ \) in terms of \( u_{\text{field}} \), \( V \), and \( T \). Then we may invoke the fact that \( S \) is a state variable, which implies

\[
\frac{\partial^2 S}{\partial T \partial V} = \frac{\partial^2 S}{\partial V \partial T}
\]  

(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 \( dQ \) in the form that we need, we can use the 1st law of thermodynamics, which is a statement of energy conservation:

\[
dQ = dU + PdV = d(u_{\text{field}}V) + PdV
\]

\[
= Vdu_{\text{field}} + u_{\text{field}}dV + \frac{1}{3}u_{\text{field}}dV
\]

\[
= V \frac{du_{\text{field}}}{dT}dT + \frac{4}{3}u_{\text{field}}dV
\]  

(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 \( dU = d(u_{\text{field}}V) \) can take place by the injection of heat \( dQ \) or by doing work \( dW = PdV \) as the volume increases. We can use (13.28) to rewrite (13.26):

\[
dS = \frac{V}{T} \frac{du_{\text{field}}}{dT}dT + \frac{4}{3}u_{\text{field}}dV
\]  

(13.29)

When we differentiate (13.29) with respect to temperature or volume we get

\[
\frac{\partial S}{\partial T} = \frac{V}{T} \frac{du_{\text{field}}}{dT}
\]

\[
\frac{\partial S}{\partial V} = \frac{4u_{\text{field}}}{3T}
\]  

(13.30)
We are now able to evaluate the partial derivatives in (13.27), which give

\[
\frac{\partial^2 S}{\partial T \partial V} = \frac{4}{3} \frac{\partial u_{\text{field}}}{\partial T} = \frac{4}{3} \frac{1}{T} \frac{\partial u_{\text{field}}}{\partial T} - \frac{4}{3} \frac{u_{\text{field}}}{T^2} \tag{13.31}
\]

Finally, (13.27) becomes a differential equation relating the internal energy of the system to the temperature:

\[
\frac{4}{3} \frac{1}{T} \frac{\partial u_{\text{field}}}{\partial T} - \frac{4}{3} \frac{u_{\text{field}}}{T^2} = \frac{1}{T} \frac{du_{\text{field}}}{dT} \Rightarrow \frac{\partial u_{\text{field}}}{\partial T} = \frac{4}{3} \frac{u_{\text{field}}}{T} \tag{13.32}
\]

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

Exercises for 13.1 Introduction

P13.1 The Sun has a radius of $R_s = 6.96 \times 10^8$ 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/cm³.

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.

Exercises for 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 $\sigma$.

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 $\sigma$. Compare results of (b) and (c).

HINT: $\int_0^\infty \frac{x^3}{e^{ax} - 1} dx = \frac{x^4}{15a^3}$.
P13.4  (a) Derive Wien's displacement law

\[ \lambda_{\text{max}} = \frac{0.00290 \text{ m} \cdot \text{K}}{T} \]

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 \( \lambda \) 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 \( \lambda_{\text{max}} \) the same as \( c / \nu_{\text{max}} \), where \( \nu_{\text{max}} \) corresponds to the peak of \( \rho_{\text{Planck}}(\nu) \)? Why would we be interested mainly in \( \lambda_{\text{max}} \)?
Bibliography


Index

A and B coefficients, 338
ABCD law for gaussian beams, 299
ABCD matrices, 238
reflection from a curved surface, 241
transmission through a curved surface, 242
aberrations, 251
Airy pattern, 292
Ampere's law, 25
derivation of, 25
array theorem, 293
astigmatism, 254
Babinet's Principle, 265
beam waist, 282
Biot-Savart law, 23
birefringence, 119
blackbody radiation, 331
boundary conditions for \( E \) and \( B \), 71
Brewster's angle, 67
decomposition of, 67
degree of coherence \( \gamma \), 202
depth of focus, 285
diffraction (scalar), 263
Fraunhofer approximation, 267
Fresnel approximation, 266
of a gaussian profile, 282
through a lens, 285
with cylindrical symmetry, 269
diffraction grating, 295
dispersion relation
in crystals, 110
vacuum, 39
displacement current, 27
distortion, 255
double boundary, 138
at sub-critical angles, 142
beyond critical angle, 145
eikonal equation, 231, 233
Einstein, Albert, 338
ellipsometry, 77, 93
elliptically polarized light, 79, 81
ellipticity, 82
energy density
electric fields, 51
magnetic fields, 52
equipartition principle (failing), 333
Euler's formula, 4
evanescence waves, 69
exchange energy, 189
Fabry-Perot
distinguishing wavelengths with, 152
etalon, 147
setup of, 150
Faraday's law, 25
Faraday, Michael, 25
Fermat's principle, 233
Fermat, Pierre, 235
finesse
coefficient of, 143
reflecting, 156
focal length
of a mirror, 244
Fourier spectroscopy, 206
fourier spectroscopy, 206
fourier theory, 7
Fraunhofer approximation, 267
Fraunhofer, Joseph, 269
free spectral range (Fabry-Perot), 154
Fresnel coefficients, 63
Fresnel's equation, 111
Fresnel, Augustin, 64
Fresnel-Kirchhoff formula, 271
fringe visibility, 204, 205
Gabor, Dennis, 324
Galileo, 246
Gauss' s law, 22
derivation of, 23
gaussian beams, 284
Green's theorem, 275
group delay, 180
generalized context, 185
group velocity, 172, 173
helicity, 82
Helmholtz equation
scalar, 263
vector, 263
holograms
generating, 324
reconstruction, 325
Huygens' elliptical construction, 123
Huygens' principle, 261
Huygens, Christian, 124
image formation, 244
complex optical system, 246
index of refraction, 42
instantaneous power spectrum, 192
integral table, 14
intensity
of a wave packet, 171
interferograms, 319
irradiance
of a plane wave, 49
Jones matrices, 82
polarizer at arbitrary angle, 85
wave plates, 88
Jones vector, 80
Jones vectors, 79
Jones, R. Clark, 81
Kirchhoff, Gustav, 332
laser cavity
stability of, 248
linear medium, 41
Lorentz force, 24
Lorentz model, 41, 43
Lorentz, Hendrik, 45
Maxwell, James, 27
Michelson interferometer, 200
Michelson, Albert, 201
multilayer coatings, 157
repeated stacks, 161
Newton, Isaac, 175
obliquity factor, 263, 266, 275
optic axes of a crystal, 115
optical activity, 100
optical path length, 234
oscillator strength, 45
p-polarized light, 60
paraxial approximation, 238
partially polarized light, 93
pellicle, 147
phase velocity, 172, 173
photometry, 53
Planck formula, 335
Planck, Max, 336
plane wave, 40
plane waves, 38
plasma frequency, 45
polarization (of a medium), 28
polarization (of light), 78
Polaroid, 82
Poynting's theorem, 47
principal axes, 107
quadratic dispersion, 182
radiometry, 53
ray tracing, 251
Rayleigh criterion, 293
Rayleigh range, 285
reflectance, 65
reflection
   at a single boundary, 60
   effect of polarization on, 90
   from metallic surface, 70
   total internal, 68
reflection, law of, 62
resolving power (Fabry-Perot), 156
Roemer, Ole, 36
rotation of coordinates, 121
s-polarized light, 61
Snell's law, 62
Snell, Willebrord, 62
spatial coherence, 208
   with a continuous source, 213
spatial filter, 307
spectrometers, 297
spectrum, 175
spherical aberration, 254
spherical wave, 261, 326
Stefen-Boltzmann law
   derivation of, 339
Strutt, John, 181
Sylvester's theorem, 11
telescope
   resolving power, 290
temporal coherence, 202
testing optical components, 321
transmittance, 65
uniaxial crystals, 115
   Poynting vector in, 119
van Cittert-Zernike theorem, 215
vector calculus, 1
wave equation, 30
Young's two slit, 208
Young, Thomas, 209

© 2004-2009 Peatross and Ware
<table>
<thead>
<tr>
<th>Constant</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permittivity</td>
<td>$\epsilon_0$</td>
<td>$8.854 \times 10^{-12} \text{ C}^2/\text{N} \cdot \text{m}^2$</td>
</tr>
<tr>
<td>Permeability</td>
<td>$\mu_0$</td>
<td>$4\pi \times 10^{-7} \text{ T} \cdot \text{m}/\text{A}$ (same as $\text{kg} \cdot \text{m}^2/\text{C}^2$)</td>
</tr>
<tr>
<td>Speed of light in vacuum</td>
<td>$c$</td>
<td>$2.9979 \times 10^8 \text{ m/s}$</td>
</tr>
<tr>
<td>Charge of an electron</td>
<td>$q_e$</td>
<td>$1.602 \times 10^{-19} \text{ C}$</td>
</tr>
<tr>
<td>Mass of an electron</td>
<td>$m_e$</td>
<td>$9.108 \times 10^{-31} \text{ kg}$</td>
</tr>
<tr>
<td>Boltzmann's constant</td>
<td>$k_b$</td>
<td>$1.380 \times 10^{-23} \text{ J}/\text{K}$</td>
</tr>
<tr>
<td>Plancks constant</td>
<td>$h$</td>
<td>$6.626 \times 10^{-34} \text{ J} \cdot \text{s}$</td>
</tr>
<tr>
<td>Planck's constant</td>
<td>$\hbar$</td>
<td>$1.054 \times 10^{-34} \text{ J} \cdot \text{s}$</td>
</tr>
<tr>
<td>Stefan-Boltzmann constant</td>
<td>$\sigma$</td>
<td>$5.670 \times 10^{-8} \text{ W}/\text{m}^2 \cdot \text{K}^4$</td>
</tr>
</tbody>
</table>