SIMG-455 Physical Optics
SPSP-455 Optical Physics

Roger L. Easton, Jr.
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology
54 Lomb Memorial Drive
Rochester, NY 14623-5604
585-475-5969
585-475-5988 (FAX)
easton@cis.rit.edu

5 April 2008
# Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 Sources:</td>
<td>1</td>
</tr>
<tr>
<td>0.1.1 Optics Texts</td>
<td>2</td>
</tr>
<tr>
<td>0.1.2 Physics Texts</td>
<td>2</td>
</tr>
<tr>
<td>1 Wave Optics = Physical Optics</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The Wave Equation</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Traveling Waves</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Notation and Dimensions for Traveling Sinusoidal Waves in a Medium</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Electric and Magnetic Fields</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1 A Note on Units</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 Magnetic Fields</td>
<td>7</td>
</tr>
<tr>
<td>1.3 Maxwell’s Equations</td>
<td>9</td>
</tr>
<tr>
<td>1.3.1 Vector Operations</td>
<td>9</td>
</tr>
<tr>
<td>1.3.2 Vector Calculus</td>
<td>15</td>
</tr>
<tr>
<td>1.3.3 Curl of Curl</td>
<td>22</td>
</tr>
<tr>
<td>1.4 1. Gauss’ Law for Electric Fields</td>
<td>22</td>
</tr>
<tr>
<td>1.5 2. Gauss’ Law for Magnetic Fields</td>
<td>23</td>
</tr>
<tr>
<td>1.6 3. Faraday’s Law of Magnetic Induction</td>
<td>23</td>
</tr>
<tr>
<td>1.7 4. Ampère’s Law</td>
<td>23</td>
</tr>
<tr>
<td>1.8 Solution of Maxwell’s Equations</td>
<td>24</td>
</tr>
<tr>
<td>1.8.1 Wave Equation for EM Waves</td>
<td>25</td>
</tr>
<tr>
<td>1.8.2 Electromagnetic Waves from Maxwell’s Equations</td>
<td>28</td>
</tr>
<tr>
<td>1.8.3 Phase Velocity of Electromagnetic Waves</td>
<td>30</td>
</tr>
<tr>
<td>1.9 Consequences of Maxwell’s Equations</td>
<td>31</td>
</tr>
<tr>
<td>1.9.1 “Shape” of the Wavefronts</td>
<td>32</td>
</tr>
<tr>
<td>1.10 Optical Frequencies – Detector Response</td>
<td>33</td>
</tr>
<tr>
<td>1.11 Dual Nature of Light: Photons</td>
<td>34</td>
</tr>
<tr>
<td>1.11.1 Momentum of Photons</td>
<td>36</td>
</tr>
<tr>
<td>2 Index of Refraction</td>
<td>37</td>
</tr>
<tr>
<td>2.0.2 Refractive Indices of Different Materials</td>
<td>39</td>
</tr>
<tr>
<td>2.0.3 Optical Path Length</td>
<td>39</td>
</tr>
<tr>
<td>2.1 Lorentz’s Model of Refractive Index</td>
<td>40</td>
</tr>
<tr>
<td>2.1.1 Refractive Index in terms of Phase</td>
<td>41</td>
</tr>
<tr>
<td>2.1.2 Refractive Index in Terms of Forces</td>
<td>43</td>
</tr>
<tr>
<td>2.1.3 Damping Forces and Complex Refractive Index</td>
<td>47</td>
</tr>
<tr>
<td>2.1.4 Dispersion Curves</td>
<td>51</td>
</tr>
<tr>
<td>2.1.5 Empirical Expressions for Dispersion</td>
<td>53</td>
</tr>
<tr>
<td>2.1.6 Negative Refractive Index</td>
<td>56</td>
</tr>
</tbody>
</table>
3 Polarization

3.1 Plane Polarization = Linear Polarization

3.2 Circular Polarization

3.2.1 Nomenclature for Circular Polarization

3.2.2 Elliptical Polarization, Reflections

3.2.3 Change of Handedness on Reflection

3.2.4 Natural Light

3.3 Description of Polarization States

3.3.1 Jones Vector

3.3.2 Actions of Optical Elements on Jones Vectors

3.3.3 Coherency Matrix \( J \)

3.4 Generation of Polarized Light

3.4.1 Selective Emission

3.4.2 Selective Transmission or Absorption

3.4.3 Generating Polarized Light by Reflection – Brewster’s Angle

3.4.4 Polarization by Scattering

3.5 Birefringence – Double Refraction

3.5.1 Examples

3.5.2 Phase Delays in Birefringent Materials; Wave Plates

3.5.3 Critical Angle – Total Internal Reflection

4 Electromagnetic Waves at an Interface

4.1 Fresnel Equations

4.1.1 Definitions of Vectors

4.1.2 Snell’s Law for Reflection and Refraction of Waves

4.1.3 Boundary Conditions for Electric and Magnetic Fields

4.1.4 Transverse Electric (TE) Waves = “s” = “⊥” Polarization

4.1.5 Transverse Magnetic (TM) Waves = “p” = “‖” polarization

4.1.6 Comparison of Coefficients for TE and TM Waves

4.1.7 Angular Dependence of Amplitude Coefficients at “ Rare-to-Dense” Interface

4.1.8 Reflectance and Transmittance at “ Rare-to-Dense” Interface

4.1.9 Reflection and Transmittance at “ Dense-to-Rare” Interface, Critical Angle

4.1.10 Practical Applications for Fresnel’s Equations

5 Optical Interference and Coherence

5.1 Addition of Oscillations

5.1.1 Addition of 1-D Traveling Waves

5.2 Division-of-Wavefront Interference

5.2.1 Addition of 3-D Traveling Waves with same \( \omega \)

5.2.2 Superposition of Two Plane Waves with \( \omega_1 \neq \omega_2 \)

5.3 Fringe Visibility – Temporal Coherence

5.3.1 Coherence of Light

5.3.2 Coherence Time and Coherence Length

5.3.3 Polarization and Fringe Visibility

5.4 Van Cittert - Zernike Theorem

5.5 Division-of-Amplitude Interferometers

5.5.1 Fizeau Interferometer

5.5.2 Michelson Interferometer

5.5.3 Mach-Zehnder Interferometer

5.5.4 Sagnac Interferometer

5.6 Interference by Multiple Reflections

5.6.1Thin Films

5.6.2 Fabry-Perot Interferometer
6 Optical Diffraction and Imaging

6.1 Huygens' Principle ............................................................... 140
6.2 Diffraction Integrals ............................................................... 141
6.3 Fresnel Diffraction ................................................................. 142
  6.3.1 Fresnel Diffraction as a Convolution Integral ................. 144
  6.3.2 Characteristics of Fresnel Diffraction ......................... 154
  6.3.3 Transfer Function of Fresnel Diffraction ..................... 155
6.4 Fraunhofer Diffraction ......................................................... 156
  6.4.1 Examples of Fraunhofer Diffraction ......................... 158
6.5 Coherence of Light and Imaging ............................................. 160
6.6 Transmissive Optical Elements ............................................ 161
  6.6.1 Optical Elements with Constant or Linear Phase ...... 162
  6.6.2 Lenses with Spherical Surfaces ................................. 163
6.7 Fraunhofer Diffraction and Imaging .................................... 166
  6.7.1 “Diffraction-Limited” and “Detector-Limited” Imaging Systems 171
6.8 Depth of Field and Depth of Focus ....................................... 171
The science of optics is often divided into three classifications based on the scale of the phenomena considered.

I. Geometrical Optics (Ray Optics): Macroscopic-scale Phenomena

(a) considers light to be a RAY that travels in a straight line until it encounters an interface between media. The wavelength $\lambda$ and temporal frequency $\nu$ of the light are assumed to be zero and infinity, respectively: $\lambda \to 0$, $\nu \to \infty$;

(b) explains reflection and refraction;

(c) calculations are simple;

(d) useful for designing imaging systems (locate the images and their magnifications);

(e) more difficult to assess the quality of the resulting image

II. Physical Optics (Wave Optics): Microscopic-scale Phenomena

(a) considers light (electromagnetic radiation) to be a WAVE;

(b) action of light is described by Maxwell’s equations;

(c) “complicated” mathematical calculations;

(d) light has wavelength $\lambda$, frequency $\nu$, velocity $c$;

(e) leads to explanations of reflection, refraction, diffraction, interference, polarization, dispersion;

(f) Useful for assessing the quality of the images.

III. Quantum Optics: Atomic-scale Phenomena

(a) light is a PHOTON with both wave-like and particle-like characteristics;

(b) used to analyze the interaction of light and matter on a sub-microscopic level;

(c) explains the photoelectric effect, lasers.

Phenomena in the first two categories are most relevant to imaging. You may have already explored the concepts of ray optics in an earlier course. This course considers the principles within the second category. The first several chapters are reviews of physical principles that are most relevant to wave optics, starting with oscillations, complex numbers, traveling waves, and the Doppler effect. From there, we will consider electromagnetic waves and Maxwell’s equations, leading to the interaction of light within matter as demonstrated by the phenomenon of dispersion. The interaction of light at an interface between two media is considered next, including some discussion of Fresnel’s equations for reflectivity and transmittance. Then follows a discussion of polarization of light and optical birefringence. Finally we will consider the aspects of wave optics that are most relevant to optical imaging systems: interference and diffraction. As we shall see, these are really just two aspects of the same phenomenon that arises from interactions of light from different locations as the light propagates. This will lead to the discussion of the fundamental limitation to the image created by an optical system; the constraint to the ability of the image to distinguish point sources that are close together.

0.1 Sources:

Many references exist for the subject of wave optics, some from the point of view of physics and many others from the subdiscipline of optics. Unfortunately, relatively few from either camp concentrate on the aspects that are most relevant to imaging.
0.1.1 Optics Texts:


0.1.2 Physics Texts:


Chapter 1

Wave Optics = Physical Optics

Ray (geometrical) optics allows determination of locations and magnifications of images created by optical systems. The study of “wave optics” (also called physical optics) uses the more sophisticated model of electromagnetic radiation as a wave “disturbance” that can exert a force on a charged particle (typically an electron because of its small mass). The wave model allows explanation and modeling of microscopic-scale phenomena, including refractive index and dispersion, polarization, reflectance and transmittance at interfaces (Fresnel equations), diffraction, and interference. The last two items (which arguably are the same phenomenon) are very useful in imaging, because they allow assessment of the quality of images created by imaging systems, which is only somewhat feasible in geometrical optics.

The discussion of wave optics begins with brief reviews of the wave equation, Maxwell’s equations (which demonstrate the simultaneous propagation of electric and magnetic fields), and refractive index. We then use Maxwell’s equation to derive the Fresnel equations for the reflectance and transmittance of light at an interface between two media. This naturally leads to the study of polarization. The final subject to consider is optical diffraction, which begins with the discussion of interference and coherence and derives the effect of diffraction on imaging system performance. The entire discussion will depend on the previous study of linear systems and Fourier methods, especially the discussions of convolution, impulse responses, and transfer functions.

The need for wave optics in imaging may be illustrated by considering the observation of the “size” of the spot of light projected through an aperture, as shown in the figure. If the diameter $d$ of the aperture is “large” (based on some criterion to be determined), then the diameter $D$ of the “image” satisfies $D \propto d$; this is the domain where ray (or geometric) optics applies. If $d$ is decreased below some limit, then $D \propto d^{-1}$ and thus $D$ increases as $d$ decreases. This phenomenon of diffraction results in the ultimate limit to the ability of an imaging system to “resolve” detail in the object.

\[ D_1 \equiv d_1 \frac{z_1 + z_2}{z_1} \propto d_1 \]

\[ D_2 \propto (d_2)^{-1} \]
Light from a point source through aperture in opaque screen: (a) diameter $d_1$ of aperture is “large” $\Rightarrow$ diameter of projection is $D_1 \propto d_1$ (larger hole $\Rightarrow$ larger “image”); (b) if aperture diameter decreased, eventually diameter of projection is $D_2 \propto (d_2)^{-1}$ (smaller hole $\Rightarrow$ larger “image”).

1.1 The Wave Equation

A wave is a “self-replicating pattern” of energy that travels through space and time (and therefore is a function of both $\mathbf{r} = [x, y, z]$ and $t$). In our usual experience with sound, some attribute of a system (e.g., the air pressure) simultaneously oscillates and propagates. The reciprocating nature of an oscillation clearly requires the joint presence of two forces:

1. A force of motion (inertia) that displaces the physical attribute (e.g., the position angle of the pendulum or the voltage in the LC oscillator) from its equilibrium value; and
2. A static force (the restoring or return force) that acts in opposition to the force of motion so to return the attribute to equilibrium. The restoring force acts as negative feedback; the greater the deviation from equilibrium, the larger the restoring force.

In many cases in our experience (acoustic or water waves), the restoring force is supplied by a characteristic of a medium, such as air or water that exerts a pressure on the oscillating attribute. This observation led to the belief that a stationary medium had to exist for light to propagate through a vacuum. Waves generated by a moving oscillation in a stationary medium that travel to an observer exhibit changes in the observed oscillation frequency as Doppler shifts. However, it is easy to show that the sizes of Doppler shifts depend on whether the source or the observer is moving relative to the medium. If a medium for light propagation exists (the “aether”), then it should be possible to use the Doppler shift to determine whether the source or the observer is moving. This led to a quandary that lasted until 1864 when James Clerk Maxwell explained how the “aether” is not necessary because light acts as its “carries its medium with it” and so is its own propagation medium. Maxwell showed that light consists of two wave fields that are both necessary for light to propagate. The fields are electric and magnetic, and thus light is called an electromagnetic wave.

This observation leads to the interpretation that the magnetic field is the “medium” for the electric field to propagate and vice versa. The final blows against the aether theory of light propagation came in the 1880s when Michelson and Morley demonstrated that the velocity of light propagation was the same regardless of the velocity of the measuring device and in 1905 with the publication of Einstein’s special theory of relativity.

The simplest oscillations, and thus the simplest waves, are harmonic, which means that the oscillation of the attribute is composed of a single sinusoidal frequency (usually defined as a cosine rather than a sine because $\cos \theta$ is a symmetric function and is more compatible with complex notation).

The wave equation was first considered by d’Alembert in 1747.

1.1.1 Traveling Waves

Recall the differential equation for the motion of a 1-D simple harmonic oscillator:

$$m \frac{d^2 y}{dt^2} = -k \cdot y[t]$$

which says that the motion $y[t]$ has the same “form” (“shape”) as its second derivative. The motion of a traveling wave also satisfies a particular differential equation, but the position $y$ is a function of both spatial and temporal coordinates. To avoid confusion between the displacement and the y-coordinate, we will rename the displacement as $\psi$, so in the case of a single spatial dimension, the position is $\psi[z, t]$ and in the 3-D case, the position is $\psi[x, y, z, t] = \psi[\mathbf{r}, t]$. For simplicity, we derive the 1-D spatial case in this section and then extend the interpretation to 3-D.

Because the waveform $\psi[z, t]$ “travels,” it must evaluate to the same value at different locations and at different times:

$$\psi[z_0, t_0] = \psi[z_1, t_1]$$
In other words, the traveling wave must have the same “shape” at a different location and at specific different times, and thus that the derivative of the function $\psi [z, t]$ must evaluate to zero at appropriate pairs $[z_n, t_n]$. The derivation makes use of the chain rule of differentiation:

$$\frac{d\psi}{dt} = 0$$
$$= \frac{\partial \psi}{\partial z} \frac{dz}{dt} + \frac{\partial \psi}{\partial t}$$
$$= \frac{\partial \psi}{\partial (z-v_\phi t)} \frac{dz}{dt} + \frac{\partial \psi}{\partial (z-v_\phi t)} \frac{d(z-v_\phi t)}{dt}$$
$$= \frac{\partial \psi}{\partial (z-v_\phi t)} \left( \frac{d(z-v_\phi t)}{dt} + \frac{\partial (z-v_\phi t)}{dt} \right)$$
$$= \frac{\partial \psi}{\partial (z-v_\phi t)} \left( 1 \cdot \frac{dz}{dt} + (-v_\phi) \right)$$
$$= \frac{\partial \psi}{\partial (z-v_\phi t)} \left( 1 \cdot \frac{dz}{dt} - v_\phi \right) = 0 \Rightarrow \frac{dz}{dt} = v_\phi$$

The amplitude $\psi$ is identical at coordinate pairs $[z_1, t_1]$ and $[z_2, t_2]$ that satisfy $z_2 = z_1 - v_\phi (t_2 - t_1)$, where $v_\phi$ is the velocity of the wave. This means that we can generate a traveling wave from a sinusoidal function of time by substituting $t \rightarrow t - \frac{z}{v_\phi}$. We can use any function for $\psi$, but we also already know that sinusoidal functions are convenient because we can decompose any $\psi$ into sinusoidal component functions. The decomposition is discrete or continuous depending on whether $\psi$ is periodic or not.

A sinusoidal wave traveling toward $z = +\infty$ is:

$$\psi [z, t] = A_0 \cos \left[ \omega_0 \left( t - \frac{z}{v_\phi} \right) + \phi_0 \right] = A_0 \cos \left[ \omega_0 \left( \frac{z}{v_\phi} - t \right) + \phi_0 \right]$$
$$= A_0 \cos \left[ \omega_0 \left( \frac{z}{v_\phi} \right) z - \omega_0 t + \phi_0 \right]$$

where the quantity $\frac{\omega_0}{v_\phi}$ has dimensions of $\frac{\text{radians}}{s} = \frac{\text{radians}}{\text{m}} = \frac{\text{ radians}}{\text{m}}$, which is an angular spatial frequency, i.e., the rate of change of the phase along the direction of propagation $z$. Designate this by the scalar $k_0$:

$$\psi [z, t] = A_0 \cos [k_0 z - \omega_0 t + \phi_0]$$

The wave equation for $\psi [z, t]$ may be confirmed by evaluating its second partial derivatives with respect to the two coordinates $z$ and $t$:

$$\frac{\partial \psi}{\partial z} = -k_0 A_0 \sin [k_0 z - \omega_0 t + \phi_0]$$
$$= -k_0 A_0 \cos \left[ (k_0 z - \omega_0 t) - \frac{\pi}{2} + \phi_0 \right]$$
$$= +k_0 A_0 \cos \left[ (k_0 z - \omega_0 t) + \frac{\pi}{2} + \phi_0 \right]$$

$$\frac{\partial^2 \psi}{\partial z^2} = -{(k_0)}^2 A_0 \cos [k_0 z - \omega_0 t + \phi_0] = -k_0^2 \psi [z, t]$$
$$\Rightarrow \psi [z, t] = -\frac{1}{k_0^2} \frac{\partial^2 \psi}{\partial z^2}$$
\[ \frac{\partial \psi}{\partial t} = - (\omega_0) A_0 \sin [k_0 z - \omega_0 t + \phi_0] = +\omega_0 A_0 \sin [k_0 z - \omega_0 t + \phi_0] = +\omega_0 A_0 \cos \left( \frac{k_0 z - \omega_0 t}{2} + \phi_0 \right) \]

\[ \frac{\partial^2 \psi}{\partial t^2} = - (\omega_0)^2 A_0 \cos [k_0 z - \omega_0 t + \phi_0] = -\omega_0^2 \psi [z, t] \]

\[ \Longrightarrow \psi [z, t] = - \frac{1}{\omega_0^2} \frac{\partial^2 \psi}{\partial t^2} \]

The 1-D wave equation is obtained by equating these two expressions:

\[ \frac{1}{k_0^2} \frac{\partial^2 \psi}{\partial z^2} = \frac{1}{\omega_0^2} \frac{\partial^2 \psi}{\partial t^2} \]

\[ \Longrightarrow \frac{\partial^2 \psi}{\partial z^2} = \frac{k_0^2}{\omega_0^2} \cdot \frac{\partial^2 \psi}{\partial t^2} = \frac{1}{(\frac{\omega_0}{k_0})^2} \cdot \frac{\partial^2 \psi}{\partial t^2} \]

The dimensions of the scale factor may be evaluated to see that:

\[ \frac{\omega_0}{k_0} = \frac{\text{radians}}{s} = \frac{\text{mm}}{s} \]

This is the velocity of the same point on the wave or the velocity of a point of constant phase on the traveling wave which is commonly called the phase velocity \( v_\phi \equiv \frac{\omega_0}{k_0} \). This confirms the 1-D wave equation:

\[ \text{1-D } \psi [z, t] : \frac{\partial^2 \psi}{\partial z^2} = \frac{1}{v_\phi^2} \frac{\partial^2 \psi}{\partial t^2} \]

The corresponding equations for waves defined over two and three spatial dimensions are:

\[ \text{2-D } \psi [x, y, t] : \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = \frac{1}{v_\phi^2} \frac{\partial^2 \psi}{\partial t^2} \]

\[ \text{3-D } \psi [x, y, z, t] : \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} = \frac{1}{v_\phi^2} \frac{\partial^2 \psi}{\partial t^2} \]

As an aside, the sums of the second-order partial derivatives with respect to the space coordinates in the 2-D and 3-D cases are both called the Laplacian, commonly denoted \( \nabla^2 \):

\[ \text{2-D: } \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = \nabla^2 \psi = \frac{1}{v_\phi^2} \frac{\partial^2 \psi}{\partial t^2} \]

\[ \text{3-D: } \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} = \nabla^2 \psi = \frac{1}{v_\phi^2} \frac{\partial^2 \psi}{\partial t^2} \]

The Laplacian appears in many contexts in imaging science, including image processing (in its discrete form).

### 1.1.2 Notation and Dimensions for Traveling Sinusoidal Waves in a Medium

For a 1-D oscillation in the form of a sinusoid that travels down the \( z \)-axis. It may be written using either trigonometric or complex notation; the former is perhaps more obvious, while the latter is easier to use in calculations.
1.1 THE WAVE EQUATION

**Trigonometric Notation:** \( y[z,t] - y_0 = A_0 \cos \{ \Phi[z,t] \} = A_0 \cos [k_0 z + \omega_0 t + \phi_0] \)

**Complex Notation:** \( y[z,t] = A_0 e^{+i\Phi[z,t]} = \text{Re} \{ A_0 \exp [+i (k_0 z + \omega_0 t + \phi_0)] \} \)

The variables and parameters in the equations are:

1. \( y \) is the coordinate (or position) of the oscillating attribute of the system, (e.g., angle, voltage, etc.).
2. \( y_0 \) is the equilibrium value of the attribute.
3. \( A_0 \) is the amplitude of the oscillation, i.e., the maximum displacement from equilibrium. The units of \( A_0 \) are the same as those of \( y \).
4. \( z, t \) are the spatial and temporal coordinates, \( [z] = \text{length} \) (e.g., mm), \( [t] = \text{s} \).
5. \( \omega_0 \) is the angular temporal frequency of the oscillation, units of \( [\omega_0] = \text{radians/second} \).
6. \( \nu_0 \) is the temporal frequency of the oscillation, units of \( [\nu_0] = \text{cycles/second} = \text{Hertz (Hz)} \), \( \nu_0 = \frac{\omega_0}{2\pi} \).
7. \( T_0 = \frac{1}{\nu_0} \) is the temporal period of the oscillation, units of \( [T_0] = \text{s} \).
8. \( \lambda_0 \) is the wavelength measured in mm.
9. \( k_0 = \frac{2\pi}{\lambda_0} \) is the angular spatial frequency of the wave, \( k_0 = \frac{2\pi}{\lambda_0}, [k_0] = \text{radians/mm} \).
10. \( \Phi[z,t] \) is the phase angle of the oscillation (the argument of the sinusoid), measured in radians at position \( z \) and time \( t \).
11. \( \phi_0 \) = initial phase angle of the wave, i.e., phase angle \( \Phi \) evaluated at \( t = 0 \) AND \( z = 0 \), \( [\phi_0] = \text{radians} \).
12. \( \sigma_0 = \frac{1}{\lambda_0} \) is the wavenumber, which is a “spatial frequency” (the number of wavelengths per unit length), \( [\sigma_0] = \text{mm}^{-1} \).

The temporal frequencies \( \nu_0 \) and \( \omega_0 = 2\pi \nu_0 \) and the spatial frequencies \( \sigma_0 \) and \( k_0 = 2\pi \sigma_0 \) of the oscillation are partial derivatives of the phase \( \Phi[z,t] \):

\[
\begin{align*}
\omega_0 &= - \frac{\partial \Phi}{\partial t} \\
\nu_0 &= \frac{\omega_0}{2\pi} = - \frac{1}{2\pi} \frac{\partial \Phi}{\partial t} \\
k_0 &= + \frac{\partial \Phi}{\partial z} \\
\sigma_0 &= + \frac{1}{2\pi} \frac{\partial \Phi}{\partial z}
\end{align*}
\]

The extension to a 3-D wave requires substitution of a wavevector \( \mathbf{k}_0 = [k_x, k_y, k_z] \) that specifies the direction of propagation and the spatial wavelength \( \lambda_0 \) of the wave. The equation for the 3-D wave is:

\[
\psi[x,y,z,t] = A_0 \cos [k_x x + k_y y + k_z z \pm \omega_0 t + \phi_0] \\
= \psi[r,t] = A_0 \cos [\mathbf{k}_0 \cdot \mathbf{r} \pm \omega_0 t + \phi_0] \\
= A_0 \cos [\mathbf{k}_0 \cdot \mathbf{r} \pm \omega_0 t + \phi_0] \\
\text{where: } |\mathbf{k}_0| = \sqrt{k_x^2 + k_y^2 + k_z^2} = \frac{2\pi}{\lambda_0}
\]

where “\( \cdot \)” denotes the scalar product of the two vectors \( \mathbf{k}_0 \) and \( \mathbf{r} \).
CHAPTER 1 WAVE OPTICS = PHYSICAL OPTICS

3-D wavevector $\mathbf{k}_0$ defines the direction of propagation of the wave; length of $|\mathbf{k}_0|$ is proportional to the reciprocal of the wavelength $\lambda_0$.

1.2 Electric and Magnetic Fields

By 1864, much was known about electric and magnetic effects on materials. Faraday had discovered that a time-varying magnetic field (as produced by a moving magnet) can generate an electric field, and Ampère demonstrated the corresponding effect that a time-varying electric field (as from a moving electric charge) produces a magnetic field. Both electric and magnetic fields were known to be vector quantities that could vary in time and space: the amplitudes of the electric and magnetic fields as functions of position. Both quantities are spatial 3-D vectors that vary over time, and may be denoted by $\mathbf{E}[x, y, z, t]$ and $\mathbf{B}[x, y, z, t]$, respectively. The electric field $\mathbf{E}$ is measured by the force it exerts on a “test” electric charge $Q$ (measured in coulombs). The force is determined by:

$$\mathbf{F} = \frac{Q}{|Q|} \mathbf{E}$$

where the force is measured in newtons $\frac{kg \cdot m}{s^2}$ as the product of the charge $Q$ and the electric field $\mathbf{E}$, which has units of volts per meter (equivalent to joules per coulomb).

1.2.1 A Note on Units

If you consult other books, you will likely see many differences in the details of these equations due to the different systems of units used in electromagnetics (and thus in optics); many students (including the author!) find it more difficult than probably necessary to cut through the seeming morass of differences. For example, two of the well-known physics texts on the subject, by Lorrain and Corson and by Jackson, use different systems; the former uses the rationalized MKS system (meter, kilogram, second, which engineers tend to use), while the latter uses CGS units (centimeter, gram, second, more useful for physicists). The MKS system includes many factors of $4\pi$. The systems evolved from Coulomb’s law that evaluates the force between two electrical charges $q_1$ and $q_2$:

$$\mathbf{F} \propto \frac{q_1 q_2}{r_{12}^2} \hat{r}_{12}$$

The constant of proportionality may be called $b$:

$$\mathbf{F} = b \frac{q_1 q_2}{r_{12}^2} \hat{r}_{12}$$

If the charges are measured in electrostatic units (esu) (also called statcoulombs), the distance $r_{12}$ in cm, and the force in dynes (g cm$^2$ s$^{-2}$), then $k = 1$. This means that two charges of 1 esu separated by 1 cm produces a force of 1 dyn. But what if the charges are measured in coulombs...
1.2 ELECTRIC AND MAGNETIC FIELDS

[C], the distance in meters [m], and the force in newtons (1 N = 1 kg m s⁻²)? The value of \( k \) is determined from the knowledge that there are \( 10^9 \) dyn per N, \( 2.998 \cdot 10^9 \) electrostatic units (esu) per Coulomb, and 100 cm per m:

\[
k \approx \frac{(2.998 \cdot 10^9 \text{ cm} \text{ dyn})}{(10^2 \text{ cm m})} \cdot 10^5 \text{ dyn N}^{-1} \approx 8.988 \times 10^9 \text{ N m}^2 \text{ C}^{-2}
\]

In words, the force between two charges of 1 C separated by 1 m is nearly \( 10^{10} \) N \( \approx 4.5 \cdot 10^{10} \) pounds of force [lbf], or about 1,100,000 tons!

The constant \( k \) generally is normalized by a factor of \( 4\pi \):

\[
k = \frac{1}{4\pi \epsilon_0} \implies F = \frac{1}{4\pi \epsilon_0} \frac{q_1 q_2}{r_{12}^2}
\]

where \( \epsilon_0 = \frac{1}{4\pi k} \approx 8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^2 = \frac{8.854 \times 10^{-12} \text{ F}}{\text{m}} = \epsilon_0
\]

where 1 Farad [F] is the unit of capacitance equivalent to:

\[
1 \text{ F} = 1 \frac{\text{C}^2}{\text{N m}} = 1 \frac{\text{C}}{\text{V}}
\]

(one coulomb per volt). This last expression is a restatement of the equation for electrical capacitance:

\[
Q \text{ [coulombs]} = C \text{ [farads]} \cdot V \text{ [volts]}
\]

which also means that one volt is equivalent to one newton-meter per coulomb:

\[
1 \text{ V} = 1 \frac{\text{N m}}{\text{C}}
\]

This “new” normalization constant \( \epsilon_0 \) is called the “dielectric constant” or “permittivity” of free space.

A similar procedure for the magnetic force between two current-carrying wires leads to the exact value of a proportionality constant \( \mu_0 \):

\[
\mu_0 = 4\pi \cdot 10^7 \frac{\text{N}}{\text{A}^2}
\]

where 1 ampere is one coulomb per second. The magnetic field in free space \( B \) (the so-called magnetic induction, measured in tesla) is then related to the magnetic field intensity \( H \) (also called the auxiliary field) in free space (measured in amperes per meter) by

\[
\mathbf{B} = \mu_0 \mathbf{H}
\]

1.2.2 Magnetic Fields

The concept of a magnetic field is seemingly somewhat less intuitive than that of an electric field, so we’ll consider it in somewhat more detail. The magnetic field is measured in terms of the “flux” (often labelled by \( \phi \)), which is a term arising from the original concept of “lines” of magnetic flux emanating from the magnetic “poles.” In fact, the original CGS unit for magnetic flux was called the “line” (now called the “maxwell,” Mx). The flux emanating from a unit field of 1 gauss is \( 4\pi \) lines because the area of the sphere is \( 4\pi r^2 \). The MKS unit of magnetic flux is the “weber” (1 Wb = 10⁸ Mx), which is the flux which, when changing uniformly in one second, induces 1 volt in 1 turn of a conductor. In electromagnetism, the more important quantity is the magnetic flux density (\( \frac{\text{flux}}{\text{unit area}} \)) labeled \( B \) and measured in gauss (CGS, \( 1 \text{ G} = 1 \frac{\text{linx}}{\text{cm}^2} = 10^{-2} \frac{\text{line}}{\text{mm}^2} \)) or tesla (MKS,
1 T = 1 \frac{\text{Wb}}{\text{m}^2}):

1 \text{ G} = 1 \frac{\text{Mx}}{\text{cm}^2}

1 \text{ T} = 1 \frac{\text{Wb}}{\text{m}^2} = 1 \frac{\text{N}}{\text{A} \cdot \text{m}} = 10^4 \text{ G}

Just like electric fields, magnetic flux will “flow” only under the influence of a force or “pressure.” The pressure of an electric field is the electromotive force (EMF) measured in volts. The magnetic “pressure” is the magnetomotive force (MMF), which is measured in “ampere-turns,” the force radiated when one ampere of current is pushed through one loop of a conductor. The resulting force field is measured in \( \frac{\text{A} \cdot \text{turn}}{\text{m}} \) (MKS) or oersteds (CGS), where 1 oersted = \( \frac{1000 \text{ A} \cdot \text{turn}}{4\pi \text{ m}} \).

The flux within a material varies with the property that the material can “resist” the flux, which is called reluctance and is measured as the induced magnetic “pressure” per unit of magnetic flux, i.e. as \( \frac{\text{MMF}}{\text{Wb}} \) (CGS) or \( \frac{\text{A} \cdot \text{turn}}{\text{Wb}} \) (MKS). This property of a material is more commonly expressed as the reciprocal, which is called the permeability \( \mu \). The permeability is analogous to electrical conductivity, the ability of a material to conduct electricity, so it would be appropriate to think of the permeability as the magnetic conductivity.

The electric field \( \mathbf{E} \) and the magnetic flux density \( \mathbf{B} \) refer to the “strength” of the force that exists in a vacuum. Two other vector fields are required when describing propagation of electromagnetism through matter: the electric displacement \( \mathbf{D} \) and the magnetic field intensity \( \mathbf{H} \) (also called the “magnetizing force” or the “auxiliary field”). In this discussion, we assume that any material is linear, isotropic, and homogeneous. “Linearity” means that the response of the medium to an incident field varies in proportion to the field. The response of “isotropic” media does not change with orientation of the field, while the characteristics of a “homogeneous” medium do not vary with position in the medium. The electric displacement \( \mathbf{D} \) defines the total electric field within a material that is due to the external field \( \mathbf{E} \). \( \mathbf{D} \) is the sum of \( \mathbf{E} \) and any local field \( \mathbf{P} \) generated within the matter due to the changes in positions of positive and negative electric charges within the atoms (and thus within the material) due to the presence of the \( \mathbf{E} \) field; this induced field \( \mathbf{P} \) is called the “polarization” of the material (which may be called the “polarizability” to avoid confusion with the “polarization” of the electric field vector, which we will mention later). \( \mathbf{H} \) is a similar construct for the magnetic field in a medium generated by the incident magnetic field (“induction”) \( \mathbf{B}, \mathbf{E} \) and \( \mathbf{D} \), and \( \mathbf{B} \) and \( \mathbf{H} \) are related by the so-called constitutive equations that are determined by constants of the medium:

\[
\begin{align*}
\mathbf{D} &= \varepsilon \mathbf{E} \\
\mathbf{B} &= \mu \mathbf{H}
\end{align*}
\]

where \( \varepsilon \) and \( \mu \) are the electric permittivity and magnetic permeability of the material, respectively, which we have already seen to be measures of the ability of the electric and magnetic fields to “permeate” the medium. If \( \varepsilon \) is increased, then a larger percentage of the external electric field exists within the material. Similarly, for a larger permeability \( \mu \), more of the magnetic field is present in the material. Since we will consider propagation of light only in vacuum, \( \mathbf{D} = \mathbf{E} \) and \( \mathbf{B} = \mathbf{H} \). In vacuum, \( \mu \) and \( \varepsilon \) are denoted \( \mu_0 \) and \( \varepsilon_0 \) and both are set to unity in CGS units. In MKS units, the quantities are:

\[
\begin{align*}
\mu_0 &= 4\pi \cdot 10^7 \frac{\text{N}}{\text{A}^2} \quad (\text{newton per ampere}^2) \\
\varepsilon_0 &= 8.85 \cdot 10^{-12} \frac{\text{F}}{\text{m}} \quad (\text{farads per meter}) = 8.85 \cdot 10^{-12} \frac{\text{C}}{\text{V} \cdot \text{m}}.
\end{align*}
\]

The permittivity and permeability are larger in matter than in vacuum: \( \varepsilon > \varepsilon_0 \), and \( \mu > \mu_0 \). In fact (though we won’t discuss it in detail), \( \varepsilon \) and \( \mu \) determine the phase velocity \( v \) and the refractive
1.3 Maxwell’s Equations

In 1864, James Clerk Maxwell published a paper on the dynamics of electromagnetic fields in which he collected four previously described equations that relate electric and magnetic forces, modified one (by adding a term to remove an inconsistency), and combined them to demonstrate the true nature of light waves. He demonstrated that the amplitudes of the electric and magnetic fields would vary as the reciprocal of the distance (rather than the square of the reciprocal of the distance, as is true for static electric fields). In this way, an electric current in one location has a much larger effect on a distant electric charge than a static electric charge at the same location as the current.

As we shall see, Maxwell’s equations demonstrate that energy can propagate from a source through vacuum to a sensor as traveling waves formed of two component vector fields that are oriented transverse (perpendicular) to the direction of propagation. Both of these fields, the electric and magnetic fields, must be present for propagation. Maxwell’s interpretation is a revolutionary concept, as it negated the need for a “medium” (the so-called aether) that had been hypothesized to provide the restoring force for oscillations to propagate. Maxwell’s equations show that the electric field provides the restoring force for the magnetic field, and vice versa, so it is reasonable to interpret that energy is conveyed by the copropagation of the two fields.

1.3.1 Vector Operations

Any physical or mathematical quantity whose amplitude may be decomposed into “directional” components often is represented conveniently as a vector. In this discussion, vectors are denoted by bold-faced underscored lower-case letters, e.g., $\mathbf{x}$. The usual notation for a vector with $N$ elements is a column of $N$ individual numerical scalars, where $N$ is the dimensionality of the vector. For example, the 3-D vector $\mathbf{x}$ is specified by a vertical column of the three ordered numerical components:

$$
\mathbf{x} = 
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{bmatrix}
$$

Both real- and complex-valued scalars will be used as the components $x_n$ with the same notation. If the $x_n$ are real, then the vector $\mathbf{x}$ specifies a location in 3-D Cartesian space. The individual scalar components $x_1$, $x_2$, and $x_3$ are equivalent to the distances along the three axial directions (commonly labeled $x$, $y$, and $z$, respectively, in the space domain). In common situations, the components of the vector $\mathbf{x}$ have dimensions of length, but other representations are possible. For example, we shall often use a convenient representation of a sinusoid in the $x–y$ plane that is specified by a vector whose components have the dimensions of spatial frequency (e.g., cycles per mm).

To minimize any confusion resulting from the use of the symbol “$x$” to represent both a vector and a particular component of a vector, a normal-faced “$x_i$” with a subscript will be used to indicate the $i^{th}$ component of the vector $\mathbf{x}$, while the bold-faced subscripted symbol “$\mathbf{x}_i$” denotes the $i^{th}$ member...
of a set of vectors. Other notations also will be employed during certain aspects of the discussion, but these cases will be explicitly noted.

Definitions of the algebraic operations of vectors will be essential to this discussion. For example, the sum of two N-D vectors \( \mathbf{x} \) and \( \mathbf{y} \) is generated by summing the pairs of corresponding components:

\[
\mathbf{x} + \mathbf{y} = \begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{bmatrix} + \begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_N
\end{bmatrix} = \begin{bmatrix}
    x_1 + y_1 \\
    x_2 + y_2 \\
    \vdots \\
    x_N + y_N
\end{bmatrix}
\]

The notation "\( x \)" and "\( y \)" used here merely distinguish between the two vectors and their components; they are not references to the \( x \)- and \( y \)-coordinates of 2-D or 3-D space. Note that this definition implies that two vectors must have the same dimension for their sum to exist.

The definition of the difference of two vectors is evident from the equation for the sum:

\[
\mathbf{x} - \mathbf{y} = \begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{bmatrix} - \begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_N
\end{bmatrix} = \begin{bmatrix}
    x_1 - y_1 \\
    x_2 - y_2 \\
    \vdots \\
    x_N - y_N
\end{bmatrix}
\]

Obviously, if the number of dimensions \( N \) of the vector is 1, 2, or 3, then the corresponding vector \( \mathbf{x} \) specifies a location on a line, on a plane, or within a volume, respectively. This interpretation of a vector as the location of a point in space is so pervasive and intuitive that it may obscure other useful and perhaps more general interpretations of vectors and vector components. For example, we can use the vector notation to represent a two-dimensional (2-D) sampled object. Such an object formed from an \( N \times N \) array of samples or by "stacking" the \( N \) columns to create a 1-D vector with \( N^2 \) components. This stacking process is known as lexicographic ordering of the matrix. Such a representation often is used when constructing computer algorithms for processing digital images, but will not be considered further here.

The transpose of the column vector \( \mathbf{x} \) is the same set of scalar components arrayed as a horizontal row, and is denoted in this discussion by a superscript \( T \); another common notation uses an overscored tilde:

\[
\mathbf{x}^T = \begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{bmatrix} = \mathbf{x}
\]

By analogy with the usual interpretation of a vector in Cartesian space, the length of a vector with real-valued components is a real-valued scalar computed from the 2-D or 3-D "Pythagorean" sum of the components:

\[
\sum_{n=1}^{N} (x_n)^2 = |\mathbf{x}|^2
\]

The result is the squared magnitude of the vector. The vector’s length, or norm, is the square root of Eq.(3.5), as shown in the figure and thus also is real valued.
Length, or “norm”, of 2-D vector with real-valued components.

$|\mathbf{x}| = \sqrt{x_1^2 + x_2^2}$

From this definition, it is evident that the norm of a vector must be nonnegative ($|\mathbf{x}| \geq 0$) and that it is zero only if all scalar components of the vector are zero.

Vectors with unit length will be essential in the discussion of transformations into alternate representations. Such a unit vector often is indicated by an overscored caret. The unit vector pointing in the direction of any vector $\mathbf{x}$ may be generated by dividing each component of $\mathbf{x}$ by the scalar length $|\mathbf{x}|$ of the vector:

$\hat{\mathbf{x}} = \frac{\mathbf{x}}{|\mathbf{x}|} = \left[ \left( \frac{x_1}{|\mathbf{x}|} \right) \left( \frac{x_2}{|\mathbf{x}|} \right) \ldots \left( \frac{x_N}{|\mathbf{x}|} \right) \right]$

The squared-magnitude operation is the first example of the vector scalar product (also called the dot product), which defines a “product” of two vectors of the same dimension that generates a scalar. Following common mathematical notation, the scalar-product operation will be denoted by a “dot” (●) between the symbols for the vectors. The process also may be written as the transpose of $\hat{\mathbf{x}}$ multiplied from the right by $\hat{\mathbf{x}}$. Therefore, the scalar product of a vector $\mathbf{x}$ with itself may be written in equivalent ways.

$|\mathbf{x}|^2 = (\mathbf{x} \bullet \mathbf{x}) \equiv \mathbf{x}^T \mathbf{x} = \sum_{n=1}^{N} x_n^2$
Scalar Product of Two Vectors

It is easy to generalize the squared magnitude operation to apply to distinct vectors $\mathbf{a}$ and $\mathbf{x}$ that have real-valued components and that have the same dimension $N$:

$$\mathbf{a} \cdot \mathbf{x} \equiv \mathbf{a}^T \mathbf{x} = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = a_1 x_1 + a_2 x_2 + \cdots + a_N x_N = \sum_{n=1}^{N} a_n x_n$$

In words, the scalar product of two vectors is obtained by multiplying pairs of vector components with the same indices and summing these products. Note that the scalar product of two distinct vectors may be positive, negative, or zero, whereas that the squared magnitude of a vector must be nonnegative. From these equivalent mathematical expressions, it is apparent that the scalar product of vectors with real-valued components in either order are identical:

$$\mathbf{a} \cdot \mathbf{x} = \mathbf{x} \cdot \mathbf{a}$$

Any process that performs an action between two entities and that may be performed in either order is commutative. The simple concept of the scalar product is the basis (future pun intended) for some very powerful tools for describing vectors and, after appropriate generalization, for functions of continuous variables. The features of the various forms of scalar product are the subject of much of the remainder of this chapter.

The scalar product of an arbitrary “input” vector $\mathbf{x}$ with a “reference” vector $\mathbf{a}$ has the form of an operator acting on $\mathbf{x}$ to produce a scalar $g$: The appropriate process was just defined:

$$\mathcal{O} \{ \mathbf{x} \} = \mathbf{a} \cdot \mathbf{x} = \sum_{n=1}^{N} a_n x_n = g$$

It is apparent that a multiplicative scale factor $k$ applied to each component of the real-valued input vector $\mathbf{x}$ results in the same scaling of the output scalar:

$$\mathcal{O} \{ k \mathbf{x} \} = \sum_{n=1}^{N} a_n (k x_n) = k \sum_{n=1}^{N} a_n x_n = k g$$

which demonstrates that the scalar product “operator” satisfies the linearity condition.

The geometrical interpretation of a 2-D vector as the endpoint of a line drawn from the origin on the 2-D plane leads to an alternate expression for the scalar product of two vectors. It is convenient to use 2-D vectors denoted by $\mathbf{f}_n$ with Cartesian components $[x_n, y_n]$, or represented in polar coordinates by the length $|\mathbf{f}_n|$ and the azimuth angles $\theta_n$. The geometric picture of the vector establishes the relationship between the polar and Cartesian representations to be:

$$\mathbf{f}_n = [x_n, y_n] = [|\mathbf{f}_n| \cos \theta_n], [|\mathbf{f}_n| \sin \theta_n]$$

where, in this case, $x_n$ and $y_n$ represent $x$- and $y$-coordinates of the vector $f_n$. The scalar product of two such vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ is obtained by applying the definition and casting into a different form
by using the well-known trigonometric identity for the cosine of the difference of two angles:

\[
\mathbf{f}_1 \cdot \mathbf{f}_2 = x_1 x_2 + y_1 y_2 = (|\mathbf{f}_1| \cos [\theta_1]) (|\mathbf{f}_2| \cos [\theta_2]) + (|\mathbf{f}_1| \sin [\theta_1]) (|\mathbf{f}_2| \cos [\theta_2])
\]

\[
= |\mathbf{f}_1| \mathbf{f}_2 \cdot (\cos [\theta_1] \cos [\theta_2] + \sin [\theta_1] \sin [\theta_2])
\]

\[
= |\mathbf{f}_1| \mathbf{f}_2 \cdot \cos [\theta_1 - \theta_2] = |\mathbf{f}_1| |\mathbf{f}_2| \cos [\theta_1 - \theta_2]
\]

where the symmetry of the cosine function has been used in the last step. In words, the scalar product of two 2-D vectors is equal to the product of the lengths of the vectors and the cosine of the included angle \(\theta_1 - \theta_2\). The knowledgeable reader is aware that this result has been obtained by circular reasoning; we are defining the scalar product form by using the Cartesian components of polar vectors, which were themselves determined by scalar products with the Cartesian basis vectors. This quandary is due in part to the familiarity of these concepts. Rather than resolve the issue from first principles, we will instead “sweep it under the rug” while continuing to use our existing intuition as a springboard to generalize these concepts to other applications. For example, it is easy now to generalize the scalar product to real-valued vectors \(\mathbf{a}\) and \(\mathbf{x}\) with arbitrary dimension \(N\):

\[
\mathbf{a} \cdot \mathbf{x} = |\mathbf{a}| |\mathbf{x}| \cos [\theta_a - \theta_x] = |\mathbf{a}| |\mathbf{x}| \cos [\theta]
\]

where \(\theta\) represents the “included” angle between the two \(N\)-D vectors.

This last definition for the scalar product may be used to derive the Schwarz inequality for vectors by recognizing that \(\cos [\theta] \leq 1\):

\[
\mathbf{a} \cdot \mathbf{x} \leq |\mathbf{a}| |\mathbf{x}|
\]

The equality is satisfied only for vectors \(\mathbf{a}\) and \(\mathbf{x}\) that “point” in the same direction, which means that the ratios of the corresponding components of \(\mathbf{a}\) and \(\mathbf{x}\) are equal, and that the included angle \(\theta = 0\) radians, which means that the vectors are scaled replicas. Note both the similarity and difference between the Schwarz inequality and triangle inequality for vectors:

\[
|\mathbf{a} + \mathbf{x}| \leq |\mathbf{a}| + |\mathbf{x}|
\]

In words, the Schwarz inequality says that the scalar product of two vectors can be no larger than the product of their lengths, while the triangle inequality establishes that one side of a triangle can be no longer than the sum of the other two sides. Both relations are illustrated in the figure.
Cross Product

Consider the area of the parallelogram formed by two vectors \( \mathbf{a} \) and \( \mathbf{b} \), as shown:

The cross product of the two vectors \( \mathbf{a} \) and \( \mathbf{b} \) yields a third vector orthogonal to those two and with length equal to \( | \mathbf{a} | | \mathbf{b} | \sin[\theta] \), which is equal to the area of the parallelogram formed by those two vectors.

The area of the parallelogram is \( | \mathbf{a} | | \mathbf{b} | \sin[\theta] \), as shown, and may be computed as a 3-D vector that points perpendicular to the two component vectors with length equal to the area; the calculation is the “cross product” of the two 3-D vectors. Given the two vectors with components:

\[
\mathbf{a} = \hat{x}a_x + \hat{y}a_y + \hat{z}a_z \\
\mathbf{b} = \hat{x}b_x + \hat{y}b_y + \hat{z}b_z
\]

the cross product may be defined as the determinant of the specific 3 \( \times \) 3 matrix:

\[
\mathbf{a} \times \mathbf{b} = \det\begin{bmatrix}
\hat{x} & \hat{y} & \hat{z} \\
\hat{a}_x & \hat{a}_y & \hat{a}_z \\
\hat{b}_x & \hat{b}_y & \hat{b}_z
\end{bmatrix}
= \hat{x} (\hat{a}_yb_z - \hat{a}_zb_y) + \hat{y} (\hat{a}_xb_z - \hat{a}_zb_x) + \hat{z} (\hat{a}_xb_y - \hat{a}_yb_x)
\]

In the example shown, the two vectors are \( \mathbf{a} = \hat{x} | \mathbf{a} | \) and \( \mathbf{b} = \hat{x} (| \mathbf{b} | \cos[\theta]) + \hat{y} (| \mathbf{b} | \sin[\theta]) \), so that \( a_z = b_z = 0 \)

\[
\det\begin{bmatrix}
\hat{x} & \hat{y} & \hat{z} \\
| \mathbf{a} | & 0 & 0 \\
| \mathbf{b} | \cos[\theta] & | \mathbf{b} | \sin[\theta] & 0
\end{bmatrix}
= \hat{z} (| \mathbf{a} | | \mathbf{b} | \sin[\theta])
\]

It is easy to see that:

\( \mathbf{b} \times \mathbf{a} = -\mathbf{a} \times \mathbf{b} \)

Note that the cross product is defined for 3-D vectors ONLY!

Triple Vector Product

The “triple vector product” is the cross product of two 3-D vectors \( \mathbf{a} \) and \( \mathbf{b} \) crossed with a third vector \( \mathbf{c} \). The result may be evaluated by straightforward (yet tedious!) calculation and produces the result:

\[
\mathbf{a} \times \mathbf{b} \times \mathbf{c} = \mathbf{b} (\mathbf{c} \cdot \mathbf{a}) - \mathbf{a} (\mathbf{c} \cdot \mathbf{b})
= \mathbf{b} (\mathbf{a} \cdot \mathbf{c}) - \mathbf{a} (\mathbf{b} \cdot \mathbf{c})
\]
where the fact that the scalar product commutes for vectors with real-valued components. The result is the difference of two scaled replicas of the first two vectors, where the scaling factors are the scalar products of $\mathbf{c}$ with $\mathbf{a}$ and $\mathbf{b}$. The "output" is a vector, as it must be.

### 1.3.2 Vector Calculus

The four equations are now collected into a group that bears his name. To interpret the four Maxwell equations, we must first understand some concepts of differential vector calculus, which may seem intimidating but is really just an extension of normal differentiation applied to scalar and vector fields. For our purposes, a scalar field is a description of scalar values in space (one or more spatial dimensions). One example of a scalar field is the temperature distribution in the air throughout the atmosphere. Obviously, a single number is assigned to each point in the space. On the other hand, a vector field defines the values of a vector quantity throughout a volume. For example, the vector field of wind velocity in the atmosphere assigns a three-dimensional vector to each point in the 3-D space. Scalar quantities are denoted by normal-face type and vectors (usually) by underscored bold-face characters, e.g., $f[x,y,z]$ and $g[x,y,z]$ describe scalar and vector fields, respectively. Unit vectors (vectors with unit magnitude, also called unit length) are indicated by bold-faced characters topped by a caret, e.g., $\hat{x}$, $\hat{y}$, and $\hat{z}$.

In preparation of the discussion of vector calculus, we’ll review a few concepts of classical mechanics. Consider a force described by the vector $\mathbf{F} = \hat{x}F_x + \hat{y}F_y + \hat{z}F_z$. The force performs "work" if it acts to create a displacement (described by the vector $\mathbf{s}$).

$$W = \mathbf{F} \cdot \mathbf{s}$$

If the displacement is the differential element $d\mathbf{s} = \hat{x} \, dx + \hat{y} \, dy + \hat{z} \, dz$, then the scalar product yields a differential element of work

$$dW = \mathbf{F} \cdot d\mathbf{s}$$

and the work resulting by the action of the force from point $a$ to point $b$ is:

$$W = \int_a^b \mathbf{F} \cdot d\mathbf{s}$$

No work is performed if the force acts at right angles to the displacement so that $\mathbf{F} \cdot d\mathbf{s} = 0$; the work is “positive” if the force acts in the direction of the displacement (e.g., a weight dropping in a gravitational field); the work is “negative” if the force acts in opposition to the displacement (a spring resisting motion). If the work is independent of the path from $a$ to $b$, then the force is conservative.

The work can be evaluated via:

$$W = \int_a^b \mathbf{F} \cdot d\mathbf{s} = \int_a^b (\hat{x} F_x + \hat{y} F_y + \hat{z} F_z) \cdot (\hat{x} \, dx + \hat{y} \, dy + \hat{z} \, dz)$$

$$= \int_a^b F_x \, dx + \int_a^b F_y \, dy + \int_a^b F_z \, dz = T + c$$

where $T$ is the kinetic energy and $c$ is a constant determined from boundary conditions.

It the vector force is a function only of the distance from some reference point, it may be written in terms of a scalar function of that distance, called the 3-D “potential” (or “potential energy”) $V$.
that satisfies the conditions:

\[
\begin{align*}
F_x &= -\frac{\partial V}{\partial x} \\
F_y &= -\frac{\partial V}{\partial y} \\
F_z &= -\frac{\partial V}{\partial z}
\end{align*}
\]

We can substitute these differential expressions into the integral equation for the work:

\[
\int \mathbf{F} \cdot d\mathbf{s} = \int \left( \mathbf{\hat{x}} F_x + \mathbf{\hat{y}} F_x + \mathbf{\hat{z}} F_z \right) \cdot \left( \mathbf{\hat{x}} \, dx + \mathbf{\hat{y}} \, dy + \mathbf{\hat{z}} \, dz \right)
\]

\[
= \int \left( -\frac{\partial V}{\partial x} \right) \, dx + \int \left( -\frac{\partial V}{\partial y} \right) \, dy + \int \left( -\frac{\partial V}{\partial z} \right) \, dz
\]

\[
= -\int dV = -V = T + c
\]

\[
\implies T + V \equiv E = \text{constant}
\]

The sum of the potential and kinetic energies is the total energy, a constant under these conditions of a “conservative system.”

For a simple illustration, consider the force of gravity near the earth’s surface; the vector force is:

\[
\mathbf{F} = \mathbf{\hat{x}} F_x + \mathbf{\hat{y}} F_x + \mathbf{\hat{z}} F_z
\]

\[
= 0\mathbf{\hat{x}} + 0\mathbf{\hat{y}} + \mathbf{\hat{z}} (-mg)
\]

so that:

\[
\begin{align*}
-\frac{\partial V}{\partial x} &= 0 \implies V = c_1 \\
-\frac{\partial V}{\partial y} &= 0 \implies V = c_2 \\
-\frac{\partial V}{\partial z} &= -mg \implies V = mg \int dz = mgz + c_3
\end{align*}
\]

\[
\implies V \mid_{x,y,z} = mgz + (c_1 + c_2 + c_3) = mgz + \text{constant}
\]

\[
E = mgz + \frac{1}{2}mv^2
\]

Under the conditions of a conservative force, we can write differentiate the first two expressions with respect to the “other” variable and equate them:

\[
\begin{align*}
\frac{\partial}{\partial y} F_x &= \frac{\partial}{\partial y} \left( -\frac{\partial V}{\partial x} \right) = -\frac{\partial^2 V}{\partial x \partial y} \\
\frac{\partial}{\partial x} F_y &= \frac{\partial}{\partial x} \left( -\frac{\partial V}{\partial y} \right) = -\frac{\partial^2 V}{\partial y \partial x} = \frac{\partial}{\partial y} F_x \\
\implies \frac{\partial}{\partial y} F_x &= \frac{\partial}{\partial x} F_y
\end{align*}
\]
1.3 MAXWELL’S EQUATIONS

The same pattern of operations leads to two other relations:

\[
\frac{\partial F_x}{\partial z} = \frac{\partial F_z}{\partial x} \\
\frac{\partial F_y}{\partial z} = \frac{\partial F_z}{\partial y}
\]

These three are necessary and sufficient conditions that a force is conservative.

We can then write:

\[
\mathbf{F} = -\left(\hat{x}\frac{\partial V}{\partial x} + \hat{y}\frac{\partial V}{\partial y} + \hat{z}\frac{\partial V}{\partial z}\right)
\]

which can be written in a shorthand form by defining the first-order differential vector operator \( \nabla \) (called “del”) with three components:

\[
\nabla = \left[ \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right]
\]

It also may be written in explicit vector form as:

\[
\nabla = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}
\]

where \( \hat{x}, \hat{y}, \) and \( \hat{z} \) are the unit vectors along the \( x, y, \) and \( z \) axes respectively. Thus we can write that the force is proportional to the gradient of a potential:

\[
\mathbf{F} = -\nabla V
\]

It is easy to show that \( \nabla \) satisfies the requirements for a linear operator:

\[
\nabla (A + B) = \nabla A + \nabla B \\
\nabla (\alpha A) = \alpha \nabla A
\]

where \( A, B \) are scalar functions and \( \alpha \) is a numerical constant.

The del operator \( \nabla \) may be applied in the same manner as a vector, though the result is a description of the rate of change of the entity to which it is applied. The operator may be applied to a 3-D “field” of scalars (such as \( f[x, y, z] \), where \( f \) is a scalar “weight”); an example is the measurement of temperature at each point in \( [x, y, z] \). The result \( \nabla f[x, y, z] \) assigns a 3-D vector to each point in space. (the gradient). The operator may be applied to a field of vectors (e.g. \( \mathbf{g}[x, y, z] \)) via a scalar product to create a scalar field \( \nabla \bullet \mathbf{g}[x, y, z] \); this is the divergence of the vector field. Finally, it may be applied to a field of 3-D vectors to create a different 3-D vector field \( \nabla \times \mathbf{g}[x, y, z] \) (the curl of the vector field). The first two operations may be generalized to operate on or generate 2-D vectors, whereas the curl is defined only for 3-D vector fields.

**Gradient of a Scalar Field**

Derives a Vector Field \( \nabla f \) from a Scalar Field \( f \)

Application of the del operator \( \nabla \) to a scalar field \( f[x, y, z] \) with three dimensions (such as the temperature of air at all points in the atmosphere) generates a field of 3-D vectors which describes the spatial rate-of-change of the scalar field, i.e., the gradient of the temperature at each point in the atmosphere is a vector that describes the direction and magnitude of the change in air temperature. In the 2-D case where the scalar field describes the altitude of landform topography, the gradient vector is the size and direction of the maximum slope of the landform.

\[
\nabla f[x, y, z] = \left[ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right] = \hat{x}\frac{\partial f}{\partial x} + \hat{y}\frac{\partial f}{\partial y} + \hat{z}\frac{\partial f}{\partial z} = \text{a vector}
\]
As implied by its name, the gradient vector at \([x, y, z]\) points “uphill” in the direction of maximum rate-of-change of the field; the magnitude of the gradient \(|\nabla f|\) is the slope of the scalar field:

\[
\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}
\]

The “gradient” of the scalar field \(f[x, y]\) specifies the vector \(\nabla f\) at each point in the field. The vector points in the direction of maximum increase in amplitude and its length is equal to this “slope.”

Divergence of a Vector field

Derives a Scalar Field \(\nabla \cdot g\) from a Vector Field \(g\):

\[
\nabla \cdot g[x, y, z] = \frac{\partial g_x}{\partial x} + \frac{\partial g_y}{\partial y} + \frac{\partial g_z}{\partial z} = a \text{ scalar}
\]

The divergence at each point in a vector field is a number that describes the total spatial rate-of-change, such as the total outgoing vector flux per unit volume (flux = net outward flow), and thus is equal to:

\[
\text{flux} = (\text{average normal vector component}) \times (\text{surface area})
\]

For a vector field \(g[x, y, z]\) and an infinitesimal surface “element” described by its normal differential vector element \(\text{d}a\) directed outward from the volume, the differential element of the “flux” \(F\) (a scalar) of \(g\) through the surface element \(\text{d}a\) is the scalar (“dot”) product of the vector that describes the field with the vector normal to the surface. Thus the total flux is the integral over the surface:

\[
dF = g \cdot \text{d}a
\]

\[
\Rightarrow F = \iint_{\text{surface area}} g \cdot \text{d}a
\]
The divergence of the vector field describes the total flux through the macroscopic surface area composed of differential surface elements \( da \) enclosing the volume. Unless the volume encloses a net “source” or “sink” of the vector field (a point from which the vector field “diverges” or “converges”), then the divergence over that surface must be zero:

\[
\nabla \cdot \mathbf{g} = \int \int_{\text{surface area}} \mathbf{g} \cdot da = 0 \text{ if no “source” or “sink” of vector field within volume enclosed by } S
\]

This is Gauss’ theorem (also called the divergence theorem).

The divergence of the scalar field \( f(x,y) \) is \( \nabla \cdot f(x,y) \) and calculates a scalar at each point in space. The divergence of vectors in a volume is zero if there are no sources or sinks of the vector field in that volume.

Of course, the flux of an electric field is not made up of a substance that “moves” through the surface, since the electric field is not the “velocity of anything” (in Feynman’s words)

Curl of a Vector Field

Derives a 3-D VectorField from a 3-D Vector Field \( \mathbf{g} \):

The curl of a vector field describes the state of spatial nonuniformity of the 3-D vector field \( \mathbf{g}[x,y,z] \). If the field describes the flow of a liquid (matter moving with a velocity), the curl determines whether the liquid is “circulating,” i.e., whether there is a net rotational motion about some location. The word definition of “circulation” is:

\[
\text{circulation} = (\text{average tangential component}) \cdot (\text{circumference})
\]

Rather than develop the measure from this equation, we again define an operator (the “curl”) and show that it measures the quantity in question. The “curl” of a 3-D vector field is the cross product of the differential operator \( \nabla \) with the field:

\[
\nabla \times \mathbf{g}[x,y,z] = \det \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ g_x & g_y & g_z \end{bmatrix} = \hat{x} \left( \frac{\partial g_z}{\partial y} - \frac{\partial g_y}{\partial z} \right) + \hat{y} \left( \frac{\partial g_z}{\partial x} - \frac{\partial g_x}{\partial z} \right) + \hat{z} \left( \frac{\partial g_y}{\partial x} - \frac{\partial g_x}{\partial y} \right) = \text{a 3-D vector}
\]

To visualize curl, imagine a vector field that describes the motion of a fluid (e.g., water or wind). If a paddle-wheel placed in the fluid does not revolve, then the field has no curl. If the wheel does revolve, the curl is nonzero. The direction of the curl vector is that of the axis of the paddlewheel when the rotation is maximized and its magnitude is that rotation rate. The algebraic sign of the curl is determined by the direction of rotation (clockwise \( \Rightarrow \) positive curl). The paddle rotates only if the vector field is spatially nonuniform. Note that the curl of the field typically varies from
point to point; some points in the field can have zero curl while others have nonvanishing curl. Both vector fields shown in the examples of divergence have zero curl, since a paddle wheel placed at any point in either field will not rotate. The vector field shown in the next figure have nonzero curl; the paddlewheel rotates. Two of Maxwell’s four equations relate the curl of one of the two fields (electric and magnetic) to the time derivative of the other field.

**Example of Function with Large Curl**  Consider the 3-D field composed of vectors defined by the equation:

\[
g[x, y, z] = (-y) \hat{x} + (+x) \hat{y} + 0 \hat{z}
\]

The vectors in this field lie in the \( x - y \) plane and those located on the \( x \) or \( y \) axes are oriented perpendicular to the axes and get longer with increasing distance from the origin.

The vectors in this field define a “flow” in the counterclockwise direction whose velocity increases with radial distance from the origin. This is a 2-D analogue of the complement of the well-known “bathtub drain vortex,” in which the vectors converge on the center of the drain and the velocity increases with decreasing distance from the center. The magnitudes and azimuth angles of the vectors in this field may be evaluated easily:

\[
|\mathbf{g}[x, y, z]| = \sqrt{(-y)^2 + (+x)^2 + 0^2} = \sqrt{x^2 + y^2}
\]

\[
\phi[x, y, 0] = \tan^{-1}\left(\frac{x}{-y}\right) = -\theta
\]

Now evaluate the partial derivatives of the vectors:
The curl of the field is obtained by direct substitution:

\[ \nabla \times \mathbf{g}[x,y,z] = \mathbf{\hat{x}} \left( \frac{\partial g_z}{\partial y} - \frac{\partial g_y}{\partial z} \right) + \mathbf{\hat{y}} \left( \frac{\partial g_x}{\partial z} - \frac{\partial g_z}{\partial x} \right) + \mathbf{\hat{z}} \left( \frac{\partial g_y}{\partial x} - \frac{\partial g_x}{\partial y} \right) \]

\[ = \mathbf{\hat{x}} (0 - 0) + \mathbf{\hat{y}} (0 - 0) + \mathbf{\hat{z}} (+1 - (-1)) \]

\[ = 0 \cdot \mathbf{\hat{x}} + 0 \cdot \mathbf{\hat{y}} + 2 \cdot \mathbf{\hat{z}} = \begin{bmatrix} 0 \\ 0 \\ +2 \end{bmatrix} \]

The curl vector points in the direction of the +z axis, i.e., out of the plane of the flow towards the viewer. The direction of the curl determines that the flow is in the x - y plane, and the magnitude of the curl is related to the “speed” of the flow, if the vectors describe a motion.

Laplacian of Scalar and Vector Fields

The divergence of the gradient of a scalar function often appears in problems in electromagnetic theory and in imaging; it is a measure of the “curvature” or the local variation of the function \( f[x,y,z] \):

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

\[ = 1 \cdot \frac{\partial^2 f}{\partial x^2} + 0 \cdot \frac{\partial^2 f}{\partial y \partial x} + 1 \cdot \frac{\partial^2 f}{\partial y^2} + 0 \cdot \frac{\partial^2 f}{\partial y \partial z} + 1 \cdot \frac{\partial^2 f}{\partial z^2} \]

Because it is a derivative, the Laplacian of a vector field is the sum of the Laplacians of the three
component functions:

\[(\nabla \cdot \nabla) \mathbf{g} = \nabla^2 \mathbf{g}\]

\[= \left( \frac{\partial^2 g_x}{\partial x^2} + \frac{\partial^2 g_y}{\partial y^2} + \frac{\partial^2 g_z}{\partial z^2} \right) \mathbf{g}\]

The Laplacian is the spatial derivative in the 3-D wave equation:

\[\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} = \frac{\mu}{\epsilon} \frac{\partial \psi}{\partial t} \]

\[\implies \nabla^2 \psi = \frac{1}{\sqrt{\epsilon}} \frac{\partial^2 \psi}{\partial t^2}\]

The discrete analogue of the Laplacian is a very useful operation in digital image processing.

### 1.3.3 Curl of Curl

Since the curl of a vector is a vector, it is possible to compute its curl (the “curl of the curl”). The shorthand solution uses the vector triple product that was presented earlier:

\[\mathbf{a} \times \mathbf{b} \times \mathbf{c} = \mathbf{b} (\mathbf{a} \cdot \mathbf{c}) - \mathbf{a} (\mathbf{b} \cdot \mathbf{c})\]

\[\implies \nabla \times (\nabla \times \mathbf{g}) = \nabla (\nabla \cdot \mathbf{g}) - \nabla^2 \mathbf{g}\]

\[= \nabla (\nabla \cdot \mathbf{g}) - (\nabla \cdot \nabla) \mathbf{g}\]

In words, it is the difference of the gradient of the divergence and the Laplacian of the vector field.

### 1.4 1. Gauss’ Law for Electric Fields

Gauss’ law relates the flux of the electric field through a closed surface to the total charge enclosed by the surface. In its simplest terms, Gauss’ law states that electrical charges within a volume produce an electric field that passes through the surface of the volume and that the flux \(\Phi\) of the field through the surface is proportional to the “amount” of charge within the volume. If the volume is enlarged, then so is the surface area, so the flux density through the surface must decrease at the same rate that the surface area increases. There still can be flux through the surface of a volume even if no charge is enclosed, but the ingoing and outgoing parts of the flux due to external charges will cancel.

A differential element of the closed surface is defined by its normal vector \(d\mathbf{a}\). The flux of the electric field \(\mathbf{E}\) through this surface element is its scalar product with \(d\mathbf{a}\):

\[d\Phi = \mathbf{E} \cdot d\mathbf{a}\]

where the symbol “\(\cdot\)” denotes the scalar product of the two vector quantities. According to Gauss’ law, the integral of the flux over the entire enclosing surface is proportional to the volume integral of the charge density \(\rho[x, y, z]\) (measured coulombs per unit volume).

\[\int_\text{surface} \mathbf{E} \cdot d\mathbf{a} = \frac{Q}{\epsilon} = \frac{1}{\epsilon} \int_\text{volume} \rho[x, y, z] \, dV\]

If the surface encloses no charges, then this integral evaluates to zero.

The surface integral may be equated to the divergence of the electric field via Gauss’ integral law:
\[ \iint_{\text{surface}} \mathbf{E} \cdot d\mathbf{a} = \iiint_{\text{volume}} \left( \frac{\partial}{\partial x} \mathbf{E} [x, y, z, t] + \frac{\partial}{\partial y} \mathbf{E} [x, y, z, t] + \frac{\partial}{\partial z} \mathbf{E} [x, y, z, t] \right) dV \]
\[ = \iiint_{\text{volume}} (\nabla \cdot \mathbf{E} [x, y, z, t]) dV \]

We can equate the volume integrals of the charge density to the volume integral of the divergence, which means that the integrands must be equal:
\[ \iiint_{\text{volume}} (\nabla \cdot \mathbf{E} [x, y, z, t]) dV = \frac{1}{\epsilon} \iiint_{\text{volume}} \rho [x, y, z] dV \]
\[ \Rightarrow \nabla \cdot \mathbf{E} [x, y, z, t] = \frac{\rho [x, y, z]}{\epsilon} \]

### 1.5 2. Gauss’ Law for Magnetic Fields

Since there are no magnetic analogues for “charges”, the volume cannot enclose a magnetic analogue of \( \rho \), which leads to the particularly simple equivalent forms for Gauss’s law for the magnetic flux density:
\[ \iint_{\text{surface}} \mathbf{B} \cdot d\mathbf{a} = 0 \]
\[ \Rightarrow \nabla \cdot \mathbf{B} [x, y, z, t] = 0 \]

In other words, the flux of the magnetic field through any enclosed surface ALWAYS is zero. This is often interpreted by the statement that there are no magnetic “monopoles.”

### 1.6 3. Faraday’s Law of Magnetic Induction

Michael Faraday observed in 1831 the phenomenon that he called “electromagnetic induction,” that generates (“induces”) electricity in a wire by means of the electromagnetic effect of a current in another wire. In other words, he discovered the basis for the electric transformer. Shortly thereafter, Faraday discovered magneto-electric induction: the production of a steady electric current by mechanical manipulation of a magnet. He attached two wires to a copper disc through a sliding contact. He rotated the disc between the poles of a horseshoe magnet and generated a continuous direct current; in short, this was the first generator.

The mathematical formulation of Faraday’s magneto-electric induction is called Faraday’s law, which states that the rate of change of a magnetic field through a surface is equivalent to the circulation of the electric field around the perimeter of the surface. In mathematical terms, the time derivative of the magnetic field is proportional to the particular spatial derivative (the curl) of the electric field:
\[ \frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E} \]

Thus a time-varying magnetic field produces a spatially varying electric field, and vice versa.

### 1.7 4. Ampère’s Law

The analogue of Faraday’s law relates the rate of change of the flux of an electric field through a surface to the circulation of the magnetic field around the perimeter of the surface. Maxwell added a “correction term” due to the flux of electric current (due to moving electric charges) through the
The corrected form of Ampère’s law is:

$$+\varepsilon \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J} = \nabla \times \frac{\mathbf{B}}{\mu}$$

where the additional source term $\mathbf{J}$ is the “current density” of the electric field (measured in amperes per unit volume, or coulombs per second per unit volume). Note the change of sign in the two analogues, Faraday’s law and Ampère’s law.

We have already seen that:

$$\mu \varepsilon = \frac{1}{c^2}$$

where $c$ is the velocity of light, $c \approx 2.99792458 \times 10^8 \text{ m s}^{-1}$, which shows that the effect of the spatial variation of the magnetic field produces a much smaller temporal change in the electric field than vice versa.

There are two “source” terms in Maxwell’s equations: the “static” charge density $\rho$ and the “dynamic” current density $\mathbf{J}$. These can only be nonzero within media (such as copper wire) and thus both vanish in vacuum without sources. If we consider the propagation of light only in a vacuum, neither electric charges nor conductors are present and both source terms vanish.

### 1.8 Solution of Maxwell’s Equations

(Jackson, Classical Electrodynamics, §6)

In 1864, James Clerk Maxwell collected these four equations and derived the form of the fields that simultaneously satisfy them in some simple cases. In rationalized MKS units, the differential forms of the equations in the presence of charges and currents are:

\[
\begin{align*}
\nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon} \quad \text{Gauss’ Law for Electric Fields, Coulomb’s Law} \\
\nabla \cdot \mathbf{B} &= 0 \quad \text{Gauss’ Law for Magnetic Fields} \\
-\frac{\partial \mathbf{B}}{\partial t} &= \nabla \times \mathbf{E} \quad \text{Faraday’s Law of Magnetic Induction} \\
+\varepsilon \frac{\partial \mathbf{E}}{\partial t} &= \nabla \times \frac{\mathbf{B}}{\mu} - \mathbf{J} \quad \text{Ampère’s Law (including correction for moving charges)}
\end{align*}
\]

The definition of curl may be used to rewrite the four vector equations of Maxwell as eight scalar equations:
1.8 SOLUTION OF MAXWELL’S EQUATIONS

\[
\begin{align*}
\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} &= \frac{\rho}{\epsilon} \text{ Gauss' Law for } E \\
\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} &= 0 \text{ Gauss’ Law for } B \\
-\frac{\partial B_x}{\partial t} &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \text{ Faraday’s Law} \\
-\frac{\partial B_y}{\partial t} &= \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \\
-\frac{\partial B_z}{\partial t} &= \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \\
+\mu \frac{\partial E_x}{\partial t} &= \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} - J_x \text{ Ampère’s Law} \\
+\mu \frac{\partial E_y}{\partial t} &= \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} - J_y \\
+\mu \frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} - J_z
\end{align*}
\]

These four coupled first-order differential equations can be solved directly in many cases. If we assume that the volume surrounding the waves that we analyze includes no sources or sinks (as would be the case for waves emitted by sources at a large distance away), then the charge density \( \rho = 0 \) and the current density \( \mathbf{J} = 0 \) \( \Rightarrow J_x = J_y = J_z = 0 \). The corresponding equations are:

\[
\begin{align*}
\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} &= 0 \\
\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} &= 0 \\
-\frac{\partial B_x}{\partial t} &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \\
-\frac{\partial B_y}{\partial t} &= \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \\
-\frac{\partial B_z}{\partial t} &= \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \\
+\mu \frac{\partial E_x}{\partial t} &= \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} \\
+\mu \frac{\partial E_y}{\partial t} &= \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} \\
+\mu \frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}
\end{align*}
\]

### 1.8.1 Wave Equation for EM Waves

Take the curl of both sides of Faraday’s law (Eq.(B) above) and apply the expression for the “curl of the curl” previously mentioned (though not derived). In a region containing no electric charges, then Gauss’ law for the electric field simplifies to \( \nabla \cdot \mathbf{E} = 0 \) and the result for the left-hand side of
The equation is:
\[ \nabla \times (\nabla \times \mathbf{E}) = \nabla (\nabla \cdot \mathbf{E}) - (\nabla \cdot \nabla) \mathbf{E} \]
\[ = \nabla (\nabla \cdot \mathbf{E}) - \nabla \cdot \nabla \mathbf{E} \]
\[ = 0 - \nabla^2 \mathbf{E} = -\nabla^2 \mathbf{E} \]

The right side of the equation may be rewritten by applying Ampère’s law:
\[ \nabla \times \left( -\frac{\partial \mathbf{B}}{\partial t} \right) = -\frac{\partial}{\partial t} (\nabla \times \mathbf{B}) \]
\[ = -\varepsilon \mu \frac{\partial}{\partial t} \left( \frac{\partial \mathbf{E}}{\partial t} \right) \]
\[ = -\frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} \]

Equate these two equations to obtain:
\[ \nabla^2 \mathbf{E} = \frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} \]
which relates the spatial and temporal second derivatives of the electric field; this again is the wave equation, but for 3-D functions. It assumes that no energy of the wave is lost to friction or damping forces.

We already inferred that the wave equation for 3-D spatial waves is written:
\[ \nabla^2 \psi [x, y, z, t] = \frac{1}{v^2} \frac{\partial^2 \psi [x, y, z, t]}{\partial t^2} \]
where \( v_\phi \) is the velocity of a point of constant phase: this is our old friend, the phase velocity.

The wave equation for electric fields confirms our earlier observation:
\[ \varepsilon \mu = \frac{1}{c^2} \implies c = \sqrt{\frac{1}{\varepsilon \mu}} \]
Think about this result for a second; it says that the phase velocity of the wave in a medium is a simple function of two properties of the medium; the electrical permittivity and the magnetic permeability.

The 1-D equation may be written in the form of a “second-order homogeneous” differential equation:
\[ \left( \frac{\partial^2}{\partial z^2} - \frac{1}{v^2_\phi} \frac{\partial^2}{\partial t^2} \right) \psi [z, t] = 0 \]
Any differential equation is linear, so that if \( \psi_1 [z, t] \) and \( \psi_2 [z, t] \) are solutions to the equation, so is \( a\psi_1 [z, t] + b\psi_2 [z, t] \), where a and b are arbitrary constants. The linearity property means that light beams can pass “through” each other and that waves can constructively or destructively interfere.

This wave equation has the simple solution:
\[ \psi [z, t] = f [z \pm v_\phi t] \]
where \( f [u] \) is any function that may be differentiated twice.
1.8 SOLUTION OF MAXWELL’S EQUATIONS

Proof.

Define \( u \equiv z \pm v_\phi t \) \( \Rightarrow \ \frac{\partial u}{\partial z} = 1, \frac{\partial u}{\partial t} = \pm v_\phi \)

Apply chain rule: \( \frac{\partial f}{\partial z} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial z} = \frac{\partial f}{\partial u} \cdot 1 \ \Rightarrow \ \frac{\partial^2 f}{\partial z^2} = \frac{\partial^2 f}{\partial u^2} \cdot v_\phi^2 \)

Apply chain rule: \( \frac{\partial f}{\partial t} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial t} = \frac{\partial f}{\partial u} \cdot (\pm v_\phi) \ \Rightarrow \ \frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial u^2} \cdot v_\phi^2 \)

Substitute into wave equation:

\[
\frac{\partial^2 f}{\partial z^2} = \frac{1}{v_\phi^2} \frac{\partial^2 f}{\partial t^2} = \frac{1}{v_\phi^2} \left( \frac{\partial^2 f}{\partial u^2} \cdot v_\phi^2 \right) = \frac{\partial^2 f}{\partial u^2}
\]

The expressions for sinusoidal waves derived in the last section satisfy the wave equation:

\[
\frac{\partial^2}{\partial z^2} \left( \hat{x} E_0 \cos (kz - \omega t) \right) = \hat{x} E_0 \left( -k^2 \right) \cos (kz - \omega t)
\]

\[
\frac{1}{v_\phi^2} \frac{\partial^2}{\partial t^2} \left( \hat{x} E_0 \cos (kz - \omega t) \right) = \frac{1}{v_\phi^2} \hat{x} E_0 \left( -\omega^2 \right) \cos (kz - \omega t)
\]

\[
= -\hat{x} E_0 \left( \frac{\omega^2}{v_\phi^2} \right) \cos (kz - \omega t)
\]

\[
\Rightarrow \ \frac{\omega^2}{k^2}
\]

If the general solution to the wave equation has the form:

\[
\psi [z, t] = f [z - v_\phi t]
\]

where the shape (“form”) of the function \( f \) is arbitrary, then the argument of the function \([z - v_\phi t]\) (the “phase”) remains constant if the location \( z \) increases with increasing time \( t \), so that the “shape” \( f \) moves towards \( z = +\infty \) with increasing time without changing its shape (i.e., without “dispersion”).

A second solution to the wave equation is:

\[
\psi [z, t] = g [z + v_\phi t]
\]

which moves towards \( z = -\infty \).

The spatial derivative of the corresponding 3-D wave equation is the sum of the three second partial derivatives:

\[
\frac{1}{v^2} \frac{\partial^2}{\partial t^2} \psi [x, y, z, t] = \frac{\partial^2}{\partial x^2} \psi [x, y, z, t] + \frac{\partial^2}{\partial y^2} \psi [x, y, z, t] + \frac{\partial^2}{\partial z^2} \psi [x, y, z, t]
\]

\[
= \nabla^2 \psi [x, y, z, t]
\]

The 3-D wave may still be a sinusoid with argument in radians, so we must be more careful about how the 3-D function becomes a 1-D function. The \( x, y, \) and \( z \) dependencies all have associated “wavelengths” that may be defined by the three corresponding “wavenumbers” \( k_x, k_y, k_z \), which may
be written as a “wavevector” \( \mathbf{k}_0 \):
\[
\psi [x, y, z, t] = A \cos \left[ \Phi [x, y, z, t] \right] \\
= A \cos \left[ k_x x + k_y y + k_z z \pm \omega_0 t - \phi_0 \right] \\
= A \cos \left[ \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} \pm \omega_0 t - \phi_0 \right] \\
= A \cos \left[ k_0 \cdot \mathbf{r} \pm \omega_0 t - \phi_0 \right]
\]

The length of the wavevector is proportional to the reciprocal of the wavelength:
\[
|k_0| \equiv k_0 = \sqrt{k_x^2 + k_y^2 + k_z^2} = \frac{2\pi}{\lambda_0}
\]

The components of the electric and magnetic fields \((E_x, E_y, E_z, B_x, B_y, B_z)\) all satisfy the wave equation.

### 1.8.2 Electromagnetic Waves from Maxwell’s Equations

In the general case, the electric field and magnetic fields can have the form:
\[
E [x, y, z, t] = \mathbf{x} E_x [x, y, z, t] + \mathbf{y} E_y [x, y, z, t] + \mathbf{z} E_z [x, y, z, t]
\]
\[
B [x, y, z, t] = \mathbf{x} B_x [x, y, z, t] + \mathbf{y} B_y [x, y, z, t] + \mathbf{z} B_z [x, y, z, t]
\]

We will now solve these equations for a single specific case: an infinite plane electric field wave propagating in vacuum toward \( z = +\infty \). The locus of points of constant phase (often called a wavefront) of a plane wave is (obviously) a plane. The electric field \( \mathbf{E} \) has no variation along \( x \) or \( y \) at a particular value of \( z \), but can vary with \( z \); this variation will be shown to be sinusoidal. This constraint affects the derivatives of the components of the electric field:
\[
\frac{\partial E_x}{\partial x} = \frac{\partial E_y}{\partial y} = \frac{\partial E_z}{\partial z} = 0
\]

and the 4-D vector field \( \mathbf{E} [x, y, z, t] \) can be written as:
\[
\mathbf{E} [x, y, z, t] = \mathbf{E} [z, t] = \mathbf{x} E_x [z, t] + \mathbf{y} E_y [z, t] + \mathbf{z} E_z [z, t]
\]

From (9) and Gauss’ law for electric fields (1), we find that:
\[
\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0 \implies \frac{\partial E_z}{\partial z} = 0
\]

Since the derivative of \( E_z \) with respect to \( z \) vanishes, then the \( z \)-component of the electric field \( E_z \) is an arbitrary constant, which we select to be 0:
\[
E_z [x, y, z] = constant \to 0
\]

Therefore, the electric field is now expressable in a much simpler form:
\[
\mathbf{E} [x, y, z] = E_x [z, t] + E_y [z, t]
\]

i.e., the only existing electric field is perpendicular (transverse) to \( z \)! This alone is a significant result. We can simplify eq.(13) by rotating the coordinate system about the \( z \) axis such that \( \mathbf{E} \) is aligned with the \( x \)-axis so that
\[
E_y [z, t] = 0 \text{ by assumption}
\]
1.8 SOLUTION OF MAXWELL’S EQUATIONS

The expression for the electric field is:

\[ \mathbf{E}[x, y, z] = \hat{x} E_x[z, t] \]  

Given the expression for \( \mathbf{E}[x, y, z, t] \), we can substitute these results into Faraday’s Law (eqs. 3, 4, 5) to find the magnetic field:

\[ -\frac{\partial B_z}{\partial t} = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = 0 \rightarrow \frac{\partial B_x}{\partial t} = \frac{\partial E_y}{\partial z} = 0 \rightarrow B_x[t] \text{ is constant} \]  

\[ -\frac{\partial B_y}{\partial t} = \frac{\partial E_z}{\partial z} - \frac{\partial E_z}{\partial y} = \frac{\partial E_x}{\partial z} \rightarrow \frac{\partial B_y}{\partial t} = -\frac{\partial E_x}{\partial z} \]  

\[ -\frac{\partial B_z}{\partial t} = \frac{\partial E_y}{\partial x} - \frac{\partial E_y}{\partial y} = 0 \rightarrow B_z[t] \text{ is constant} \]  

We can arbitrarily set the constant term \( B_z = 0 \), so the only remaining equation is:

\[ -\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} \]  

which says that the time derivative of the magnetic field \( B_y \) is equal to the negative of the space derivative of \( E_x \). We can now find a relation between \( B_y \) and \( E_x \) by standard solution techniques of differential equations. Assume that: \( \mathbf{E} \) is a vector field that varies sinusoidally with \( z \):

\[ \mathbf{E}[x, y, z, t] = \hat{x} E_x[z, t] = \hat{x} E_0 \cos[k_0 z - \omega_0 t] \]

\[ \implies -\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} = -k_0 E_0 \sin[k_0 z - \omega_0 t] \]

\[ \implies -B_y[z, t] = \int \frac{\partial E_x}{\partial z} dt = \int \sin[k_0 z - \omega_0 t] dt \]

\[ -B_y[z, t] = +k_0 E_0 \cdot \left[ \frac{-\cos[k_0 z - \omega_0 t]}{-\omega_0} \right] \]

\[ = E_0 \frac{k_0}{\omega_0} \cos[k_0 z - \omega_0 t] \]

\[ = E_0 \frac{k_0}{v_\phi} \cos[k_0 z - \omega_0 t] \]

\[ = E_0 \cos[k_0 z - \omega_0 t] \]

\[ \implies \mathbf{B}[z, t] = \hat{y} \left( -E_0 \frac{k_0}{v_\phi} \cos[k_0 z - \omega_0 t] \right) \]

where \( v_\phi \) is the phase velocity of the electromagnetic wave

\[ B_y = E_0 \frac{k_0}{v_\phi} \cos[k_0 z - \omega_0 t] = E_x \frac{k_0}{v_\phi} \implies E_x = v_\phi B_y \]

Note that the only existing component of \( \mathbf{B} \) is \( B_y \), which is perpendicular to the component \( E_x \) of \( \mathbf{E} \). Also note that the sinusoidal variations of \( \mathbf{E} \) and \( \mathbf{B} \) have the same arguments, which means that they oscillate “in phase”. The amplitude of the magnetic field is smaller by the factor of the phase velocity \( v_\phi = c \), so the effect of the magnetic field on observations is generally much smaller and often ignored.
1.8.3 Phase Velocity of Electromagnetic Waves

Given the form for the plane electromagnetic wave in a vacuum, we can now use the three Ampère relations to find something else useful:

\[
\begin{align*}
+\mu e & \frac{\partial E_x}{\partial t} = \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} \\
+\mu e & \frac{\partial E_y}{\partial t} = \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} \\
+\mu e & \frac{\partial E_z}{\partial t} = \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}
\end{align*}
\]

Because \( E = \hat{k}E_0 \), only (6) does not vanish:

\[
\mu e \frac{\partial}{\partial t} (E_0 \cos (k_0 z - \omega_0 t)) = -\frac{\partial}{\partial z} \left( \frac{E_0}{v_0} \cos (k_0 z - \omega_0 t) \right)
\]

\[
\Rightarrow \mu e E_0 (\omega_0 \sin (k_0 z - \omega_0 t)) = -\frac{E_0}{v_0} (-k_0 \sin (k_0 z - \omega_0 t))
\]

\[
\Rightarrow \mu e \omega_0 E_0 = \frac{E_0 k_0}{v_0}
\]

\[
\Rightarrow \mu e = \frac{k_0}{\omega_0 v_0} = \frac{1}{v_0^2} \Rightarrow v_0^2 = \left( \frac{\omega_0}{k_0} \right)^2 = \frac{1}{\mu e}
\]

\[
v_0 = \left( \frac{1}{\mu e} \right)^{-\frac{1}{2}}
\]

which we already knew from the wave equation. In vacuum, \( \mu \equiv \mu_0 \), \( \epsilon \equiv \epsilon_0 \), \( v_0 \equiv c \), and \( c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} \).

The permittivity and permeability of free space (vacuum) can be measured in laboratory experiments, thus allowing a calculation of the phase velocity of electromagnetic waves. The permeability and permittivity in vacuum are respectively:

\[
\mu_0 \equiv 4\pi \times 10^{-7} \text{ newton} \ampere^{-2} \equiv 1.26 \times 10^{-6} \frac{\text{N}}{\text{A}^2} = 1.26 \times 10^{-6} \frac{(\text{J} \text{m})}{\text{A}^2}
\]

\[
\epsilon_0 \equiv 8.85 \times 10^{-12} \frac{\text{F}}{\text{m}} = 8.85 \times 10^{-12} \frac{\text{C}}{\text{V} \text{m}}
\]

Their product is:

\[
\nu_0 \epsilon_0 = \left( 8.85 \times 10^{-12} \frac{\text{C}}{\text{V} \text{m}} \right) \left( 1.26 \times 10^{-6} \frac{\text{J}}{\text{A}^2 \text{m}} \right)
\]

\[
= 1.11 \times 10^{-17} \frac{\text{C}}{\text{V} \text{m}} \cdot \frac{\text{J}}{\left( \text{C} \cdot \text{sec} \right)^{2} \text{m}}
\]

\[
= 1.11 \times 10^{-17} \frac{1}{\text{V} \text{m}} \cdot \frac{\text{J} \text{sec}^2}{\text{C} \text{m}}
\]

\[
= 1.11 \times 10^{-17} \frac{\text{sec}^2}{\text{m}^2}
\]

\[
\Rightarrow c = \sqrt{\frac{1}{\nu_0 \epsilon_0}} \approx 2.99 \times 10^8 \frac{\text{m}}{\text{sec}}, \text{ which agrees with experiment.}
\]

In media (i.e., not in vacuum), the phase velocity is slower than \( c \). The same relation may be written using the permittivity and permeability of the medium. The permeability of most optical materials
is close to that of vacuum ($\mu \approx \mu_0$), while the permittivity $\epsilon$ of optical materials is larger than in vacuum:

$$\epsilon > \epsilon_0 \approx 8.85 \times 10^{-12} \frac{F}{m}$$

So therefore

$$v = \sqrt{\frac{1}{\mu \epsilon}} \approx \sqrt{\frac{1}{\epsilon}} \leq \sqrt{\frac{1}{\epsilon_0}} \approx \sqrt{\frac{1}{\mu_0 \epsilon_0}} = c$$

### 1.9 Consequences of Maxwell’s Equations

1. **Copropagation of $E$ and $B$:** The wave travels in a direction mutually perpendicular to both $E$ and $B$, and in fact the propagation direction $\mathbf{S}$ is defined by the direction:

   $$\mathbf{S} = (c^2 \epsilon) \cdot (\mathbf{E} \times \mathbf{B})$$

   *the Poynting vector*

2. The force induced by the electric field on an electric charge $q_0$ (measured in Coulombs) is specified by the Lorentz relation:

   $$\mathbf{F} = q_0 \left( \mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B} \right)$$

   which indicates that a force is induced on the charge by the magnetic field only if the charge is moving with velocity $\mathbf{v}$, and that the force is scaled by the velocity of light $c$. In short, the force due to the electric field dominates under most conditions. This observation, and the fact that the waves are transverse to the direction of propagation leads to the concept of $\circ$ of the electric field, which is merely its orientation measured from some radial reference line transverse to the direction of propagation.

3. Both electric and magnetic fields are required for energy and force to propagate. In other words, the two vector fields copropagate. I like to interpret this result as meaning that the magnetic field provides the medium for propagation of the electric field and vice versa.

4. In vacuum, $E$ and $B$ are in phase, which means that the phases of the sinusoidal variation of $E$ and $B$ are identical (the phases of the fields often are out of phase in some types of matter).

5. Both $E$ and $B$ travel at $c = \sqrt{\mu_0 \epsilon_0}$, the phase velocity of the wave.

6. Energy is carried by both the electric and magnetic fields, and the magnitude of the energy $E \propto E_0^2$.

7. There is no limitation on the possible frequencies of the waves, i.e., $[0 \leq \omega \leq \infty]$, which implies the allowed wavelengths are in the interval $[\infty \geq \lambda \geq 0]$

8. The average power of the light wave per unit area is the “irradiance,” which is a measure of the *average energy through one unit of area in one unit of time*. The irradiance is determined
from the Poynting vector. For a wave propagating down the z-axis, the irradiance is:

$$I[x, y, z; \Delta T] = \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} S[x, y, z, t'] \, dt'$$

$$= \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} c^2 \varepsilon_0 \left[ (\hat{x}E_0 \cos [k_z z - \omega_0 t]) \times \hat{z} \left( \frac{E_0}{c} \cos [k_z z - \omega_0 t] \right) \right] \, dt'$$

$$= \hat{z} (\varepsilon_0 E_0^2) \cdot \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} \cos^2 [k_z z - \omega_0 t] \, dt'$$

$$= \hat{z} (\varepsilon_0 E_0^2) \cdot \frac{1}{2} = \hat{z} \left( \frac{\varepsilon_0 E_0^2}{2} \right) \cdot 1[x, y]$$

In words, the unit vector \(\hat{z}\) indicates that the irradiance is propagating in this direction; \(1[x, y]\) indicates that the irradiance is uniform in the plane orthogonal to this direction, and the irradiance “value” is \(\varepsilon_0 E_0^2 / 2\), which has dimensions of:

$$\left[ c \cdot \varepsilon_0 \cdot (E_0)^2 \right] = \frac{m}{s} \cdot \frac{F}{m} \cdot \left( \frac{V}{m} \right)^2 = \frac{F \cdot V^2}{m^2 \cdot s}$$

But the relationship of capacitance \(Q = C \cdot V\) demonstrates that one farad is one coulomb per volt, which allows restatement:

$$\left[ \frac{c \cdot \varepsilon_0 \cdot (E_0)^2}{m^2} \right] = \frac{F \cdot V^2}{m^2 \cdot s} = \frac{C \cdot V^2}{m^2} = \frac{C \cdot V}{m} = \frac{V}{A}$$

because one ampere is one coulomb per volt. Now recall that the product of volts and amperes is power measured in watts, so we can rewrite this one more time

$$\left[ c \cdot \varepsilon_0 \cdot (E_0)^2 \right] = \frac{V}{m^2} = \frac{W}{m^2}$$

This measure of power through one square meter of area is an appropriate measure of irradiance.

Relationship between \(\mathbf{E}\) and \(\mathbf{B}\) for a linearly (plane) polarized wave traveling from left to right. The two fields are in phase and each oscillates in its own plane; the two planes are orthogonal.

1.9.1 “Shape” of the Wavefronts

Maxwell’s equations state that the electromagnetic field propagates away from the source, but the resulting description of the waves does not describe the “shape” of the wavefront, i.e., the locus of
points in the field that have the same phase because they were emitted by the source at the same instant. The shape of the wavefront depends on the shape of the source: a point source emits spherical wavefronts; a line source emits cylindrical wavefronts, and a planar source emits planar wavefronts.

### 1.10 Optical Frequencies – Detector Response

The general equation for a traveling electromagnetic wave is:

\[
y [z, t] = A_0 \cos \left( k_0 z + \omega_0 t + \phi_0 \right)
\]

\[
= A_0 \cos \left[ 2\pi \left( \frac{z}{\lambda_0} - \nu_0 t \right) + \phi_0 \right]
\]

\[
= A_0 \Re \{ \exp [ +i (k_0 z + \omega_0 t + \phi_0)] \}
\]

We “see” electromagnetic radiation with sensors that respond in some form to incident light by conversion to some other, measurable, attribute, e.g., to heat (bolometer), to a change in chemical form (photochemical reaction), or to a change in an electrical property (e.g., resistance in a photoconductor). The type of conversion determines the speed of the response of the sensor.

The human eye is sensitive only to light with wavelengths in the range \(400 \text{ nm} \lessgtr \lambda \lessgtr 700 \text{ nm}\). However, this is not the case for all lifeforms. The pit viper can see radiation emitted by humans at \(\lambda \approx 10 \text{ m}\) with special receptors on the sides of its head.

Clearly, the frequencies of visible wavelengths are quite large:

\[
\lambda = 400 \text{ nm} \implies \nu = \frac{c}{\lambda} \approx 7.5 \times 10^{14} \text{ Hz}
\]

\[
\lambda = 550 \text{ nm} \implies \nu \approx 5.5 \times 10^{14} \text{ Hz}
\]

\[
\lambda = 700 \text{ nm} \implies \nu \approx 4.3 \times 10^{14} \text{ Hz}
\]

The frequency bandwidth of visible light is:

\[
\Delta \nu = \nu_{\text{max}} - \nu_{\text{min}} \approx 3.2 \times 10^{14} \text{ Hz} = 320 \text{ THz}
\]

and the temporal period of an optical wave is therefore

\[
T = \nu^{-1} \approx 1.8 \times 10^{-15} \text{ s}
\]

Sensors, including the human eye, cannot respond at this speed and therefore cannot detect the periodic oscillation of wave amplitude: the optical phase. Human eyes and other sensors of visible light “see” a constant response to incident light. As asides, we mention that it is possible to measure relative phase of light using indirect techniques, such as optical interference used in holography, and that human ears also cannot sense phase in sound waves at frequencies above a few Hz. It is possible to use the relative time delay in sound arrival at your two ears to locate sound sources.

For other types of waves, e.g., waves in media (such as water waves) and lower-frequency electromagnetic waves (e.g., radio waves or microwaves), the amplitude and phase is measurable. This allows many more detection techniques to be applied, such as modulation (mixing) for phase measurement.

Now consider the effect of time averaging; the average amplitude of a sinusoidal wave measured
over the time interval $T_d$ is:

$$\langle y(z,t) \rangle = \frac{1}{T_d} \int_0^{T_d} y(z,t) \, dt$$

$$= \frac{1}{T_d} \int_0^{T_d} A_0 \cos [k_0 z - \omega_0 t + \phi_0] \, dt$$

$$= -\frac{A_0}{\omega_0 T_d} \sin [k_0 z - \omega_0 t + \phi_0] \bigg|_{t=0}^{t=T_d}$$

If $T_d$ is large, then $\langle y(z,t) \rangle \to 0$. If $T_d \leq \frac{1}{\delta_0} = \frac{2\pi}{\omega_0}$, then $\langle y(z,t) \rangle$ may be finite. However, the power of the sinusoidal wave (its squared magnitude) does not average to zero:

$$I(z) \propto \langle y^2(z,t) \rangle = \frac{1}{T_d} \int_0^{T_d} y^2(z,t) \, dt$$

$$= \frac{1}{T_d} \int_0^{T_d} A_0^2 \cos^2 [k_0 z - \omega_0 t] \, dt$$

$$= \frac{A_0^2}{T_d} \int_0^{T_d} \cos^2 [k_0 z - \omega_0 t] \, dt$$

$$= \frac{A_0^2}{2} \frac{T_d}{T_d} = \frac{A_0^2}{2}$$

$$\Rightarrow I(z) \propto \langle y^2(z,t) \rangle \Rightarrow \frac{A_0^2}{2} \quad \text{if } T_d >> \nu^{-1}$$

because the average value of $\cos^2 [x] = \frac{1}{2}$.

Detectors of visible light are sensitive to time-averaged power (called irradiance).

1.11 Dual Nature of Light: Photons

Though the wave picture of light is essential in analyzing the resolution of imaging systems, the “particle” picture of light is often more appropriate in the study of imaging sensors. Consider images created of light with different exposure times. Photographs taken in shorter exposures are made with fewer photons and generally look “grainier” because of statistical uncertainty in the number of counted photons is more apparent:
The “particle of light” is the photon, which conveys energy in discrete “packets” or “quanta.” Einstein demonstrated that the energy carried by a photon is proportional to its temporal oscillation frequency:

\[ E = h\nu = \frac{hc}{\lambda} = \hbar \omega \]

where \( h \) and \( \hbar \) are the various flavors of Planck’s constant; \( \hbar \) (called “h-bar”) is normalized by a factor of \( 2\pi \):

\[ h \cong 6.625 \times 10^{-34} \text{ J s} = 6.625 \times 10^{-27} \text{ erg s} \]

\[ \hbar = \frac{h}{2\pi} \cong 1.054 \times 10^{-34} \text{ J s} \]

For example, the energy per photon at \( \lambda = 550 \text{ nm} \) is:

\[ E = 6.625 \times 10^{-27} \text{ erg s} \cdot \frac{3 \times 10^8 \text{ m}}{550 \text{ nm}} \cong 3.6 \times 10^{-12} \text{ erg} \ @ \lambda = 550 \text{ nm} \]

Large swarms of photons may be analyzed only by their statistics, which requires the study of statistical mechanics and is beyond our scope. For our purposes, we will make only the briefest mention of the physics. Classical mechanics with distinguishable particles may be analyzed by Maxwell-Boltzmann statistics, where the number of particles with energy state \( E_n \) in an ensemble of many particles (\( N_0 \)) held at temperature \( T \) is a decaying exponential distribution:

\[ N_n = N_0 \exp \left(-\frac{E_n}{kT}\right) \]

\[ k = \text{Boltzmann’s constant} \cong 1.38 \times 10^{-23} \text{ J K}^{-1} \]

which clearly indicates that \( N_n \downarrow \) as \( E_n \uparrow \). Quantum mechanics involves indistinguishable particles whose statistics depend on the “spin” characteristics of the particles; particles with integer spins are called “bosons” and obey Bose-Einstein statistics, including photons, which have unit spin. Bosons tend to cluster in an energy state, which means that the electromagnetic field created by an ensemble of photons clustered in one energy (i.e., they have the same corresponding frequency) appears as a continuous electromagnetic wave. Electrons are fermions, which must each have a distinctive energy and thus cannot cluster in one state.

The “photon flux” \( \Phi \) is the number of photons per second in a light beam, which may be evaluated from the total power of the beam (measured in watts)

\[ \Phi = \frac{P \text{ [W]}}{h \text{ [J s]} \cdot \nu \text{ [Hz]} = I \frac{\text{W}}{\text{m}^2} \cdot A \frac{\text{m}^2}{\text{J s} \cdot \nu \text{ [Hz]}}} \]

Images created from increasing numbers of photons, showing increase in signal-to-noise ratio.
where $J$ is the irradiance of the light and $A$ is the cross-sectional area. Typical fluxes per unit area for some sources are shown in the table:

<table>
<thead>
<tr>
<th>Light Source</th>
<th>$\frac{\Phi}{A}$ photons sec mm$^{-2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>focused laser</td>
<td>$10^{20}$</td>
</tr>
<tr>
<td>unfocused laser</td>
<td>$10^{15}$</td>
</tr>
<tr>
<td>bright sunlight</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>indoor light</td>
<td>$10^{10}$</td>
</tr>
<tr>
<td>twilight</td>
<td>$10^8$</td>
</tr>
<tr>
<td>moonlight</td>
<td>$10^6$</td>
</tr>
<tr>
<td>starlight</td>
<td>$10^4$</td>
</tr>
</tbody>
</table>

The “pattern” of photon arrivals measured by a sensor tells something about the source. Random (incoherent) light sources (such as light bulbs) emit photons with a Bose-Einstein distribution of random arrival times. Coherent light sources, on the other hand, emit photons with a Poisson distribution, which is more uniform but still random.

1.11.1 Momentum of Photons

Atoms that emit photons “recoil” in the opposite direction, and surfaces that absorb photons also recoil. The momentum of a single photon is

$$p = \frac{h}{\lambda} = \hbar k$$

The pressure due to radiation is the force per unit area, which is equal to the energy per unit volume, or the energy density. Radiation pressures are often neglected, but cannot be if the mass is small or the flux is large, e.g., in the motion of comet tails or spacecraft, in stellar interiors, and in the light of lasers.
# Chapter 2

## Index of Refraction

The idea that the refractive index relates the velocity of light in a medium to that in vacuum is ubiquitous, but does not by itself address the reason “why” light slows down within a medium. This section attempts to address this issue.

The refractive index \( n \) is the dimensionless ratio of the velocity of light in vacuum to that in the medium:

\[
 n \equiv \frac{c}{v} \geq 1
\]

\[
 n = \sqrt{\frac{1}{\mu_0 \epsilon_0}} = \sqrt{\frac{\mu \epsilon}{\mu_0 \epsilon_0}} \approx \sqrt{\frac{\epsilon}{\epsilon_0}}
\]

in dielectric materials that have no free electrons to conduct electric current. The ratio \( \frac{\epsilon}{\epsilon_0} \) is the relative permittivity of the medium. In a dielectric medium with \( \epsilon > \epsilon_0 \), the refractive index \( n > 1 \).

We can relate refractive index to the wavelength of light in the medium after recognizing that \( E = h \nu \) and energy conservation requires that \( \nu \) not change in different media. This means that the wavelength must change. If the wavelength in vacuum is \( \lambda_0 \) the wavelength in a medium with index \( n \) is \( \lambda' \):

\[
 \frac{\lambda_0}{1} = \frac{\lambda'}{n}
\]

\[
 \implies k_0 = \frac{2\pi}{\lambda_0} = \frac{2\pi n \nu_0}{c} = \frac{n \omega_0}{c}
\]

In words, the wavenumber (length of the wavevector) in a medium is scaled by the index \( n \) of the medium.

In nondielectric materials such as metals or at wavelengths near dielectric absorption features, the index of refraction is complex valued and often is denoted by \( \tilde{n} \) and its imaginary part by \( \kappa \):

\[
 \tilde{n}^2 \equiv (n + i\kappa)^2 = \frac{\mu \epsilon}{\mu_0 \epsilon_0}
\]

(some authors write \( \tilde{n} = n \cdot (1 + i\kappa) \), where \( \kappa \) is the “attenuation index”). The length of the
propagation wavevector for a complex-valued index is:

\[ k_0 = \frac{n \omega_0}{c} = \left( n + i \kappa \right) \frac{\omega_0}{c} \]

\[ \text{Re} \{ k_0 \} = \frac{n \omega_0}{c} \]

\[ \text{Im} \{ k_0 \} = \kappa \frac{\omega_0}{c} \]

This means that the wavevector \( \mathbf{k} \) also is complex valued. If \( \mathbf{\hat{s}} \) is the unit vector in the direction of \( \mathbf{k} \), we can write the electric field:

\[ \mathbf{E}(x, y, z, t) = \mathbf{E}_0 \exp \left[ +i \left( k_0 \mathbf{s} \cdot \mathbf{r} - \omega_0 t \right) \right] \]

Note the distinct multiplicative exponential factors: the first has a complex-valued argument and represents a field that propagates sinusoidally, while the second has a real-valued negative argument that decays with increasing distance \( |\mathbf{r}| = r \).

If we assume that the direction of propagation \( \mathbf{\hat{s}} \) is in the direction of \( \mathbf{r} \) (as would be the case for a plane wave), then

\[ \mathbf{\hat{s}} \cdot \mathbf{r} = |\mathbf{\hat{s}}| |\mathbf{r}| \cos(0) = |\mathbf{r}| = r \]

and the electric field may be written in a simplified form:

\[ \mathbf{E}_0 \exp \left[ i \omega_0 \left( \frac{n}{c} r - t \right) \right] \exp \left[ -\kappa \frac{\omega_0}{c} r \right] = \mathbf{E}_0 \exp \left[ i \omega_0 \left( \frac{n}{c} r - t \right) \right] \exp \left[ -\frac{r}{\delta_0} \right] \]

where the parameter \( \delta_0 \) is defined in terms of the wavelength \( \lambda_0 \) measured in vacuum and the imaginary part \( \kappa \) of the refractive index:

\[ \delta_0 = \left( \frac{\kappa \omega_0}{c} \right)^{-1} = \frac{c}{\kappa \omega_0} = \frac{\lambda_0}{2\pi \kappa} \]

The quantity \( \delta_0 \) has dimensions \( \text{m} \) and is the distance over which the field amplitude decreases by a factor of \( e^{-1} \approx 0.368 \); it is called the skin depth of the medium. The amplitude decreases exponentially with increasing \( r \), so that the electric field is attenuated as it propagates. Consider the effect in a conductive medium where \( \kappa \) is large. This means that \( \delta_0 \) is small and the electric field decreases rapidly with distance “into” the conductor. This observation demonstrates that electric current is propagated along the surfaces of conductors and not within the “bulk.”

If the imaginary part \( \kappa \) of the refractive index is zero, there is no attenuation and the electric field is the simple oscillating term:

\[ \mathbf{E}_0 \exp \left[ i \omega_0 \left( \frac{n}{c} r - t \right) \right] = \mathbf{E}_0 \exp \left[ i \omega_0 \left( \frac{r}{v_\phi} - t \right) \right] \]

which confirms that the velocity in the medium is the ratio of the velocity in vacuum and \( n \):

\[ v_\phi = \frac{c}{n} \]

The magnetic field is obtained from the electric field via the cross product with the electric-field...
vector:
\[ \mathbf{B} = \frac{k \times \mathbf{E}}{\omega_0} = \left( \frac{n + i\kappa}{c} \right) (\hat{s} \times \mathbf{E}) \]
which indicates that it also decays with with depth within a conductor.

### 2.0.2 Refractive Indices of Different Materials:

Values of the refractive index for common nonconductive materials include:

<table>
<thead>
<tr>
<th>Medium</th>
<th>Refractive Index ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>( = 1.0 ) (by definition)</td>
</tr>
<tr>
<td>oxygen</td>
<td>( \approx 1.000271 ) at S.T.P. (20°C, 1 atm)</td>
</tr>
<tr>
<td>air</td>
<td>( \approx 1.0002929 ) at S.T.P., 1.01241 at 42 atm</td>
</tr>
<tr>
<td>water</td>
<td>( \approx 1.33 \approx \frac{4}{3} )</td>
</tr>
<tr>
<td>“crown” glass</td>
<td>( \approx 1.5 = \frac{3}{2} )</td>
</tr>
<tr>
<td>“flint” glass</td>
<td>( \approx 1.7 )</td>
</tr>
<tr>
<td>diamond</td>
<td>( \approx 2.417 )</td>
</tr>
<tr>
<td>germanium</td>
<td>( \approx 4.0 ) (though only transparent at wavelengths ( \lambda \gtrsim 2 \mu m ))</td>
</tr>
</tbody>
</table>

The entries for air at different pressures demonstrate that \( n \) increases with increasing pressure, which satisfies the intuition that the refractive index should increase with increasing number density of atoms. This also means that \( n \) decreases with increasing temperature. This variation in \( n \) with pressure and temperature is the source of “atmospheric image blur,” where an image through the atmosphere changes from moment to moment due to variations in temperature and pressure that change the refractive index. The image moves both transversely (around the image plane) and longitudinally (to go “in and out of focus”). These changes happen over time intervals of the order of \( \frac{1}{100} \) sec, which means that longer exposures are blurred due to image motion in exactly the same fashion as a long-exposure image of a moving object taken with a fixed camera is blurred.

Observations of atmospheric turbulence indicate that images may be modeled as created through isolated areas of the atmosphere within which the refractive index is approximately uniform (“isoplanatic patches”) if measured over time intervals \( \approx 10^{-2} \) sec. The apparent diameter of these regions is of the order of 100 mm, which sets the ultimate resolution limit for images of astronomical objects taken through the atmosphere.

The complete picture of the effect of the atmosphere on images also depends on the location of the medium relative to the object and imaging system. Think of the two complementary vertical imaging modes through the atmosphere: from the ground up and from orbits down. In the former, the imaging system is “close to” (indeed, it is immersed in the atmosphere), whereas in the latter the object is close to the atmosphere and distant from the imaging system. The transverse motion of the atmosphere may be characterized as an angular motion, which is relatively much larger in the “ground up” situation. Thus it is easier to get good images of the ground from spacecraft than of astronomical objects (or spacecraft) from the ground, which is fortunate for people working in “adaptive optics,” which tries to compensate for the atmospheric blur in real time.

### 2.0.3 Optical Path Length

We have already seen that the phase velocity of light in a medium is slower than in vacuum, which means that light takes longer to travel through a given thickness of material than through the same
“thickness” of vacuum. For a fixed medium thickness \( d \), we know that:

\[
d = v \cdot t \quad (\text{distance} = \text{velocity} \times \text{time})
\]

\[
= c \cdot t_1 \quad (\text{in vacuum})
\]

\[
= \frac{c}{n} \cdot t_2 \quad (\text{in medium of index } n)
\]

\[
\implies t_1 = \frac{t_2}{n} \implies t_2 > t_1
\]

In the time \( t_2 \) required for light to travel the distance \( d \) in a material of index \( n \), light would travel a longer distance in the same time in vacuum: \( nd = ct_2 \). The distance

\[
nd \geq d
\]

which is the distance light travels in vacuum in the same time as it would traverse the distance \( d \) in the medium with index \( n \) is the optical path length within the medium:

\[
\text{OPL} = nd \text{[mm]}
\]

### 2.1 Lorentz’s Model of Refractive Index

(Feynman Lectures on Physics Vol. I §31, §32 and II §32).

Electric waves interacting with a layer of transparent glass: some of the field is reflected and the original field plus a “correction” term are transmitted.

H.A. Lorentz proposed a simplified explanation of the refractive index that is based on the “solar system” model of atomic structure, where positively charged massive nucleus of protons and neutrons surrounded by a cloud of negatively charged electrons. To ensure stability of the atomic structure, the electrostatic force between electrons and protons must be balanced by a force of motion, which we assume applies only to electrons since they are so much less massive than protons \((m_p \cong 1836 m_e)\).

The forces on the electrons produce an equation of motion that has an oscillatory solution whose temporal frequency depends upon the parameters of the atom or molecule. An electron oscillates at a “natural” frequency \( \omega_0 \) that depends on its mass and the force. An incident electric field can stimulate the electron; it absorbs and re-emits the light, which is called scattering. Under the influence of the incident electric field, the electron acts as a forced harmonic oscillator. The
scattering from the electron “system” varies with the temporal frequency of the incident radiation, and it is this variation in response that is called dispersion.

The configuration of the system is shown in the figure; a point source is located a large distance away to the left, so that the incident electric field is a plane wave. The observation point is a large distance to the right (the figure is “not to scale!”). The observed electric field is the sum of the original electric source field \(E_s\) and a “correction term” \(E_2\) of electric fields generated by moving electrical charges in the plate; this is the “scattered light.” We assume that only electrons can absorb (and therefore emit) light, since the massive protons are much more difficult to move. The index of refraction is due to this correction term \(E_2\), which makes the electric field inside the glass “appear” to be moving at a different (slower) phase velocity.

In reality, a single electron in the glass is affected by the incident field and by the fields generated by all other electrons in the glass, just as the motions of all other charges are influenced by the single observed electron. To simplify the problem, we assume that the influences of the other electrons are small compared to the effect of the incident field, so that the total field at the observation point is little affected by the motions of the other charges. In effect, we are assuming that the index of refraction of the glass is very close to unity. The calculation will produce a field \(E_2\) that travels in the same direction as the incident field and a field \(E_1\) that travels in the opposite direction (the “reflected” light), but \(E_1\) and \(E_2\) are small because \(n \approx 1\).

### 2.1.1 Refractive Index in terms of Phase

Because its source is far away, the incident electric field is a traveling plane wave that may be written in 1-D complex notation:

\[
E_s = E_0 \exp \left[ +i (k_0 z - \omega t) \right] = E_0 \exp \left[ +i \omega \left( \frac{z}{c} - t \right) \right]
\]

We avoid use of the notation “\(\omega_0\)” for the temporal frequency of the variable incident (“driving”) field in this derivation because this frequency will be varied and because the subscripted notation is reserved for the “natural oscillation” frequency of the electron due to the forces from the nucleus.

Assume that \(z = 0\) at the “front” (input) side of the thin plate of glass and \(z = \Delta z\) at the back side. The phase of the electric field at the front of the plate is a function only of the observation time:

\[
\phi[0, t] = k_0 \cdot 0 - \omega t = -\omega t
\]

If there were no glass, then the phase at the location of the back of the plate would be incremented by the additional propagation to \(z = \Delta z\):

\[
\phi[\Delta z, t'; \text{vacuum}] = \omega \left( \frac{\Delta z}{c} - t \right) = -\omega t + \omega \left( \frac{\Delta z}{c} \right)
\]

Insertion of the glass “slows down” the light by an additional factor because the phase velocity in the glass is \(v_\phi < c\). The phase of light at the rear of the glass includes an additional factor due to the slower phase velocity \(v_\phi\):

\[
\phi[\Delta z, t; \text{glass}] = \omega \left( \frac{\Delta z}{v_\phi} - t \right) > \omega \left( \frac{\Delta z}{c} - t \right)
\]

If we substitute the index of refraction \(n\), then the parts of the phase due to the distance \(\Delta z\) differ by its scale factor:

\[
n = \frac{c}{v_\phi}
\]

\[
\Rightarrow \phi[\Delta z, t; \text{glass}] = \omega \left( \frac{n \cdot \Delta z}{c} - t \right) > \omega \left( \frac{\Delta z}{c} - t \right)
\]
The “additional” phase due to the extra time to travel through the glass is:

\[
\phi [\Delta z, t; \text{glass}] - \phi [\Delta z, t'; \text{vacuum}] = \omega \left( \frac{n \cdot \Delta z}{c} - t \right) - \omega \left( \frac{\Delta z}{c} - t \right)
\]

\[
\Delta \phi = \omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right)
\]

The electric field at the back of the glass is the product of the incident field and an additional phase delay:

\[
E [\Delta z, t] = E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \cdot \exp [-i \cdot \Delta \phi]
\]

\[
= E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \cdot \exp \left[ -i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) \right]
\]

This allows the contribution of the glass plate to the electric field to be interpreted as a multiplicative contribution to the optical phase instead of as a multiplicative contribution to the amplitude. The phase interpretation is much easier to analyze because it separates into additive terms. The incremental phase may be expanded into a Taylor series:

\[
\exp [-i \cdot \Delta \phi] = \exp \left[ -i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) \right]
\]

\[
= 1 - i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) + \frac{(-i)^2}{2!} \left( \frac{\omega \cdot (n - 1) \cdot \Delta z}{c} \right)^2 + \cdots
\]

If we assume that the glass is sufficiently thin (so that \( 1 \gg \Delta z \gg 0 \implies \Delta z \gg (\Delta z)^2 \)), then we can ignore all terms of order two or larger without much error and the exponential reduces to the sum of two terms:

\[
\exp [-i \cdot \Delta \phi] \approx 1 - i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right)
\]

which, after substitution, leaves:

\[
E_{\text{after}} \approx E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \cdot \left( 1 - i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) \right)
\]

\[
= \left( E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \right) - \left( i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) \left( E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \right) \right)
\]

The first term is just the source field at the front of the plate:

\[
E_s = E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right]
\]

and the second term is identified as the approximate contribution from the charges within the glass, which is labeled in the Figure as \( E_2 \):

\[
E_2 \approx -i\omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) \left( E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \right)
\]

\[
= \exp \left[ -i\frac{\pi}{2} \right] \left( \omega \left( \frac{(n - 1) \cdot \Delta z}{c} \right) \left( E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \right) \right)
\]

The leading factor of \( -i = \exp \left[ -i\frac{\pi}{2} \right] \) may be interpreted as a phase difference of the electric field due to the charges in the glass; in words, the electric field \( E_2 \) is in quadrature to the original electric field. The vector (phasor) contributions of the two fields are shown in the figure.
2.1 LORENTZ’S MODEL OF REFRACTIVE INDEX

Argand diagram of phasor contributions from incident field $E_0$ and field due to charges in the glass $E_2$, which is oriented approximately perpendicular to $E_0$ and “delays” the phase of the electric field.

If the field $E_2$ thus evaluated is interpreted as due to the oscillating charges in the glass, then we have explained the concept of refractive index, which is quantified in terms of the physical forces in the next section.

2.1.2 Refractive Index in Terms of Forces

The task now is to evaluate the fields in terms of the forces on the electrons in the system. Again, we assume that the incident field has the form of a plane wave that oscillates sinusoidally at angular temporal frequency $\omega$:

$$E_s[z,t] = E_0 \exp [+i(k_0 z - \omega t)]$$

The source field at the “front edge” of the glass may be evaluated by setting $z = 0$:

$$E_s[0,t] = E_0 \exp [-i\omega t]$$

The electrons at this part of the glass “feel” this field and are driven to move in the same direction by the force due to the electric field $E_s$:

$$F = eE_s[0,t] = eE_0 \exp [-i\omega t]$$

(we assume no specific orientation – polarization – of the incident field) These electrons have mass $m$ and act as though bound to the atomic nuclei (protons) by little “springs” that exert restoring forces proportional to the distance of the electron from its equilibrium position:

$$F = -k(x - x_0)$$

Note that the dimensions of $k/m$ are

$$\left[ \frac{k}{m} \right] = \frac{N}{\text{kg s}^2} = \frac{\text{kg m}}{\text{kg s}^2} = \frac{\text{kg m}}{\text{kg s}^2} = \left( \frac{\text{radians}}{s} \right)^2$$

which is the square of an angular frequency: the resonant frequency $\omega_0$, which is the “normal oscillating frequency” of the system composed of electron + spring in classical mechanics:

$$\sqrt{\frac{k}{m}} = \omega_0$$

The equation of motion of an electron under the influence of the driving field is that of an undamped driven harmonic oscillator:

$$m \frac{d^2x}{dt^2} + m\omega_0^2 x = F = eE_0 \exp [-i\omega t]$$
where the last term is the “driving force” due to the electric field. This equation is solved by standard methods of differential equations, by assuming that the position $x$ also oscillates at the same rate. The position $x$ and its time derivatives are:

$$x = x_0 \exp[-i\omega t]$$

$$\frac{dx}{dt} = \dot{x} = -i\omega x_0 \exp[-i\omega t] = -i\omega x$$

$$\frac{d^2x}{dt^2} = \ddot{x} = (-i\omega)^2 x_0 \exp[-i\omega t] = -\omega^2 x$$

These are substituted into the equation of motion simplifies to obtain an equation for the amplitude $x_0$ of the oscillation:

$$m \omega_0^2 x_0 \exp[-i\omega t] + m\omega_0^2 x_0 \exp[-i\omega t] = eE_0 \exp[-i\omega t]$$

$$\implies (-m\omega^2 + m\omega_0^2) x_0 = eE_0$$

$$\implies x_0 = \frac{eE_0}{m(\omega_0^2 - \omega^2)}$$

We assume that the motions of ALL individual electrons in the glass plate due to the incident electric field are identically described by the same simple expression:

$$x(t) = x_0 \exp[-i\omega t] = \frac{eE_0}{m(\omega_0^2 - \omega^2)} \exp[-i\omega t]$$

Note that this does not include the initial positions of the charges, and thus the phase of the position $x(t)$, which are (obviously) different for each.

We now must calculate the field at the observation point (well beyond the plate) due to a “thin plane” of charges that all move with the same equation of motion $x(t)$. We find the field at the observation point by adding the contributions from each of the charges in the glass. The electric field radiated by each electron in the glass is proportional to the acceleration just evaluated:

$$\frac{d^2x}{dt^2} = (-i\omega)^2 x_0 \exp[-i\omega t] = -\omega^2 x$$

The electric field at large distances from the oscillating charge falls off approximately as the reciprocal of the distance and includes the time delay for the field to arrive:

$$E[r, t] \simeq \frac{e}{r} (-\omega^2 x_0) \exp[-i\omega \left(t - \frac{r}{c}\right)]$$

We assume that the observation point is so far away that the field oscillates approximately perpendicular to the “line of sight.”

The total field generated by the charges at the observation point is the vector sum of the contributions from the individual electrons, which may be integrated in polar coordinates because the contributions along all azimuth angles are assumed to be identical. If $\eta$ is the “area number density” of electrons in the glass (number of electrons per unit area, which may be multiplied by $\Delta z$ to find the number density per unit volume), then the electric field is:

$$E_{all} = E_2 = \int_{\rho=0}^{\rho=+\infty} \frac{e}{r} (-\omega^2 x_0) \exp[-i\omega \left(t - \frac{r}{c}\right)] \cdot \eta \cdot 2\pi \rho \cdot d\rho$$

$$= 2\pi \eta \frac{e}{c} (-\omega^2 x_0) \exp[+i\omega \left(t - \frac{r}{c}\right)] \int_{\rho=0}^{\rho=+\infty} \rho \cdot d\rho$$

where the length of the 3-D vector $r$ is the Pythagorean sum of the lengths of the 2-D vector $\rho$ and
The longitudinal distance \( z \):

\[
|E|^2 = r^2 = \rho^2 + z^2
\]

\[
\Rightarrow r \cdot dr = \rho \cdot d\rho
\]

The integral in the expression for the total electric field \( E_{\text{all}} \) becomes:

\[
\int_{\rho=0}^{r=+\infty} \exp \left[ +i\frac{\omega r}{c} \right] \cdot \frac{1}{r} \cdot \rho \cdot d\rho = \int_{r=z}^{r=+\infty} \exp \left[ +i\frac{\omega r}{c} \right] \cdot \frac{1}{r} \cdot r \cdot dr
\]

\[
= \int_{r=z}^{r=+\infty} \exp \left[ +\frac{\omega}{c} \right] \cdot dr
\]

\[
= \frac{c}{i\omega} \left( \exp \left[ +i \cdot \infty \right] - \exp \left[ +i\frac{\omega z}{c} \right] \right)
\]

\[
\approx \frac{c}{i\omega} \left( - \exp \left[ +i\frac{\omega z}{c} \right] \right)
\]

\[
= +\frac{ic}{\omega} \exp \left[ +i\frac{\omega z}{c} \right]
\]

The first term oscillates “infinitely rapidly” so that the integral of its contributions must be negligible. The electric field due to all charges is:

\[
E_2 = 2\pi\eta c \left( -\omega^2 x_0 \right) \exp \left[ -i\omega t \right] \cdot \left( \frac{ic}{\omega} \exp \left[ +i\frac{\omega z}{c} \right] \right)
\]

\[
= (2\pi\eta ce) \cdot (-i\omega) \cdot x_0 \cdot \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right]
\]

which shows that the field measured at the observation point due to all of the oscillating charges is out of phase by \( \frac{\pi}{4} \) radians and delayed due to the finite propagation time.

All that is left to do is to substitute the formula for \( x_0 \) derived above for the driven harmonic oscillator:

\[
E_2 \approx (2\pi\eta ce) \cdot (-i\omega) \cdot \left( \frac{E_0}{m (\omega_0^2 - \omega^2)} \right) \cdot \exp \left[ -i\omega t \right] \exp \left[ +i\omega \left( \frac{z}{c} \right) \right]
\]

\[
= -i \left( \frac{(2\pi\eta ce^2) E_0}{m} \right) \left( \frac{\omega}{\omega_0^2 - \omega^2} \right) \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right]
\]

which has the form of a sinusoidal traveling wave. In words, the electrons in the glass that oscillate due to the incident field emit a wave that travels in the same direction (towards \( z = +\infty \)). The amplitude of the wave is proportional to the area density of atoms \( \eta \) and to the strength of the source field \( E_0 \). This field resembles that of \( E_2 \) that was evaluated in the last section:

\[
E_2 \approx -i\omega \frac{(n-1)}{c} \Delta z \left( E_0 \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right] \right)
\]

\[
= -i\omega \frac{(2\pi\eta ce^2) E_0}{m (\omega_0^2 - \omega^2)} \exp \left[ +i\omega \left( \frac{z}{c} - t \right) \right]
\]

If we equate the two and cancel the common factors, we obtain an expression for the index of
refraction in terms of the parameters of the system and the driving frequency $\omega$:

$$\frac{(n - 1) \cdot \Delta z}{c} = \frac{2\pi\eta e^2}{m(\omega_0^2 - \omega^2)}$$

$$\Rightarrow n = 1 + \frac{(2\pi\eta e^2)}{m(\omega_0^2 - \omega^2)} \cdot \frac{\eta}{\Delta z}$$

We now define $N$ to be the number density of electrons per unit volume in the glass, which is the ratio of the area density $\eta$ and the thickness $\Delta z$:

$$\frac{\eta \text{ [mm}^{-2}]}{\Delta z \text{ [mm]}} = N \text{ [mm}^{-3}]$$

to obtain the final expression for the frequency-dependent refractive index in the simple model of oscillating bound (but undamped) electrons:

$$n = 1 + \frac{2\pi N \omega_0^2 \omega}{m(\omega_0^2 - \omega^2)} \hspace{1cm} \text{(no damping)}$$

The magnitude of the refractive index is approximately unity if evaluated at angular temporal frequencies $\omega$ distant from the natural frequency $\omega_0$: $\omega << \omega_0$ or $\omega >> \omega_0$. If $\omega \approx \omega_0$, $n >> 1$; if $\omega \approx \omega_0$ then $n << 1$ and may be negative. This last behavior is a clue that the picture is incomplete. In fact, the electron response actually is “damped” by other forces in the medium, which is the subject of the next section.
resonance due to the lack of damping force. The index decreases with increasing \( \lambda \) (normal dispersion) except at the discontinuity.

2.1.3 Damping Forces and Complex Refractive Index

The oscillator will absorb energy from the electric field if the electron motion is damped, as by frictional forces. This leads to a complex-valued refractive index:

\[
\hat{n} = n + i\kappa
\]

The equation of motion includes a term proportional to the velocity:

\[
m \frac{d^2x}{dt^2} + \alpha \frac{dx}{dt} + kx = -eE
\]

Again, the incident field \( E_0 \) oscillates sinusoidally at the driving frequency \( \omega \):

\[
E = E_0 \exp[-i\omega t] \implies x = x_0 \exp[-i\omega t]
\]

The derivatives are substituted into the equation of motion to evaluate the amplitude \( x_0 \) of the oscillation:

\[
\frac{dx}{dt} = -i\omega x_0 \exp[-i\omega t]
\]

\[
\frac{d^2x}{dt^2} = -\omega^2 x_0 \exp[-i\omega t]
\]

\[
(m \frac{d^2}{dt^2} + \alpha \frac{d}{dt} + k) x_0 \exp[-i\omega t] = (-m\omega^2 - i\omega\alpha + k) x_0 \exp[-i\omega t]
\]

\[
\implies (-m\omega^2 - i\omega\alpha + k) x_0 \exp[-i\omega t] = -eE_0 \exp[-i\omega t]
\]

\[
\implies x_0 = \frac{eE_0}{(-m\omega^2 - i\omega\alpha + k)} = \frac{eE_0}{m(\omega^2 + i\omega\frac{\alpha}{m} - \frac{k}{m})}
\]

where we have already identified:

\[
\frac{k}{m} \equiv \omega_0^2
\]

We now also rename the term:

\[
\frac{\alpha}{m} \equiv \gamma \ [\text{Hz}]
\]

which is the reciprocal of a time constant, which may be interpreted as the time interval over which the amplitude of the oscillation decays to a specific value. The expression for the amplitude of the oscillation may be written in two equivalent forms:

\[
x_0[\omega] = \frac{\varepsilon m E_0}{(\omega^2 - \omega_0^2) + i\omega\gamma}
\]

\[
= \frac{\varepsilon m E_0}{(\omega^2 - \omega_0^2) + i\omega\gamma} \cdot \left( \frac{(\omega^2 - \omega_0^2) - i\omega\gamma}{(\omega^2 - \omega_0^2) - i\omega\gamma} \right)
\]

\[
= \frac{E_0}{m} \frac{(\omega^2 - \omega_0^2) - i\omega\gamma}{(\omega^2 - \omega_0^2)^2 + \gamma^2\omega^2}
\]

which is clearly complex valued. Note that if \( \omega = \omega_0 \), the amplitude \( x_0 \) no longer is undefined, but rather is imaginary:

\[
x_0[\omega = \omega_0] = -i\frac{eE_0}{m} \frac{1}{\gamma\omega_0}
\]
We can insert the new expression for the amplitude into the formula for polarizability:

\[ P = -Ne_0 \]

\[ = -Ne \left( \frac{E_0}{(\omega^2 - \omega_0^2) + i\omega \gamma} \right) \]

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

First, we should test to see that this reverts to the static value for \( P \) if \( \omega = 0 \) (no incident light):

\[ \lim_{\omega \to 0} \{ P[\omega] \} = \frac{Ne^2 E_0}{\omega_0^2} \]

The refractive index \( n \) may be evaluated exactly as before:

\[ \frac{(n - 1) \cdot \Delta z}{c} = \frac{2\pi \eta e^2}{m (\omega_0^2 - \omega^2 - i\omega \gamma)} \]

\[ \Rightarrow n = 1 + \frac{2\pi Ne^2 e^2}{m (\omega_0^2 - \omega^2 - i\omega \frac{\eta}{m})} \]

Note that \( n \) now is complex valued; it often is denoted \( \hat{n} \equiv n + i\kappa \):

\[ \hat{n} = 1 + \frac{2\pi Ne^2 e^2}{m (\omega_0^2 - \omega^2 - i\omega \gamma)} \]

This expression is rewritten into its real and imaginary parts:

\[ \hat{n} = 1 + \frac{2\pi Ne^2 e^2}{m (\omega_0^2 - \omega^2)} \]

\[ = \frac{(2\pi Ne^2 e^2) E_0}{m ((\omega_0^2 - \omega^2) - i\omega \frac{\eta}{m}) ((\omega_0^2 - \omega^2) + i\omega \gamma)} \]

\[ = 1 + \frac{(2\pi Ne^2 e^2) E_0}{m (\omega_0^2 - \omega^2)^2 + (\omega \gamma)^2} \]

\[ = \left( 1 + \frac{(2\pi Ne^2 e^2) E_0}{m (\omega_0^2 - \omega^2)^2 + (\omega \gamma)^2} \right) + i \left( \frac{(2\pi Ne^2 e^2) E_0}{m (\omega_0^2 - \omega^2)^2 + (\omega \gamma)^2} \right) \]

\[ n = 1 + \frac{(2\pi Ne^2 e^2) E_0}{m (\omega_0^2 - \omega^2)^2 + (\omega \gamma)^2} \]

\[ \kappa = \frac{(2\pi Ne^2 e^2) E_0}{m (\omega_0^2 - \omega^2)^2 + (\omega \gamma)^2} \]
2.1 LORENTZ’S MODEL OF REFRACTIVE INDEX

Attenuation due to \(\text{Im} [n]\)

The complex-valued refractive index may be inserted into the propagation phase:

\[
E[z, t] = E_0 \exp \left[ +i\omega \left( \frac{n_z z}{c} - t \right) \right]
\]

\[
= E_0 \exp \left[ +i\omega \left( \frac{(n + i\kappa) z}{c} - t \right) \right]
\]

\[
= E_0 \exp \left[ +i\omega \left( \frac{n z}{c} - t \right) \right] \exp \left[ +i\omega \left( \frac{\kappa z}{c} \right) \right]
\]

\[
= E_0 \exp \left[ +i\omega \left( \frac{n_z z}{c} - t \right) \right] \exp \left[ - \left( \frac{z}{c/\kappa} \right) \right]
\]

where the last term may be rewritten in the form:

\[
\exp \left[ - \frac{z}{c/\kappa} \right] = \exp \left[ - \frac{z}{\delta} \right]
\]

where \(\delta = \frac{\kappa}{\omega}\) has dimensions of length. This last term is the source of the decaying amplitude of the scattered sinusoidal waves emitted by the charges. The distance \(\delta\) is the decay length over which the amplitude decays to the factor \(e^{-1} \approx 0.367\) of its original maximum.

Refraction index in the limit of no damping

Note that the real part of the refractive index differs from the undamped index, but reverts to the undamped formula if \(\gamma = 0\), which implies that its reciprocal \(\gamma^{-1}\) with units of time goes to infinity:

\[
\lim_{\alpha \to 0} (\text{Re} [n]) = \lim_{\alpha \to 0} (n) = \lim_{\alpha \to 0} \left( 1 + \frac{2\pi N e^2 e^2}{m} \frac{(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + (\omega\gamma)^2} \right)
\]

\[
= 1 + \frac{2\pi N e^2 e^2}{m (\omega_0^2 - \omega^2)} = n_{\text{undamped}}
\]

\[
\lim_{\alpha \to 0} (\text{Im} [n]) = \lim_{\alpha \to 0} (\kappa) = \lim_{\alpha \to 0} \left( \frac{2\pi N e^2 e^2 a}{m} \frac{\omega\gamma}{(\omega_0^2 - \omega^2)^2 + (\omega\gamma)^2} \right) = 0
\]

The real part of the refractive index is graphed as functions of the driving frequency \(\omega\) and driving wavelength \(\lambda\), along with the undamped case (dashed line, for comparison). Both \(\omega_0\) and \(\lambda_0\) were selected to be unity for convenience. The damping “smooths” the dispersion curve so that \(n\) varies more slowly with \(\lambda\) and the real part decreases

with increasing \(\lambda\) (normal dispersion) except in the vicinity of the resonance where the index increases with \(\lambda\) (anomalous dispersion).
Model of the real part of the refractive index in the damped case in vicinity of a resonance, plotted as function of both frequency $\omega$ and wavelength $\lambda$ and compared to the undamped case (dashed line). Again, the constants are arbitrary.

The complex refractive index also is shown as functions of both $\omega$ and $\lambda$. Note the Lorentzian form of the imaginary part, which is the source of the spectral absorption.
The complex refractive index in the vicinity of a resonance located at $\omega_0 = 1$ and $\lambda_0 = 1$, plotted as real part (solid) and imaginary part (dashed). The imaginary part attenuates the transmission.

**Attenuation due to Im[$n$]**

Recall the formula for the electric field traveling down the $z$-axis in a medium with complex refractive index $\tilde{n} = n + i\kappa$

$$E_0 \exp \left[ +i\omega_0 \left( \frac{\tilde{n}}{c} z - t \right) \right] = E_0 \exp \left[ +i\omega_0 \left( \frac{n + i\kappa}{c} z - t \right) \right]$$

$$= E_0 \exp \left[ +i\omega_0 \left( \frac{n \kappa}{c} t - \frac{\kappa}{c} z \right) \right]$$

$$= E_0 \exp \left[ +i\omega_0 \left( \frac{n \kappa}{c} t - \frac{\kappa}{c} z \right) \right] \cdot \exp \left[ -\frac{\kappa}{c} z \right]$$

$$= E_0 \exp \left[ +i\omega_0 \left( \frac{n \kappa}{c} t \right) \right] \cdot \exp \left[ -\frac{\kappa}{c} z \right]$$

The amplitude of the last term decreases exponentially with increasing $z$ at a rate determined by $\frac{\kappa}{\kappa} \equiv \delta$, the attenuation depth. This represents the attenuation (or absorption) due to the imaginary part of $\tilde{n}$.

### 2.1.4 Dispersion Curves

Though the behaviors of different optical materials differ in details, they have many properties in common. We have already stated that the index of refraction $n$ relates the phase velocity of light in vacuum with that in matter:

$$n = \frac{c}{v_\phi} \geq 1.$$

In a transparent dispersive medium over the visible spectrum, the index $n$ decreases with increasing $\lambda$, which ensures that the phase velocity $\frac{\omega}{k}$ (of the average wave) is larger than the group velocity $\frac{d\omega}{dk}$ (of the modulation wave).

![Dispersion curves for various optical glasses in the visible region of the spectrum.](image)
Typical dispersion curve for glass at visible wavelengths, showing the decrease in $n$ with increasing $\lambda$ and the three spectral wavelengths specified by Fraunhofer and used to specify the “refractivity”, “mean dispersion”, and “partial dispersion” of a material.

Dispersion curve of a material from very short to very long wavelengths. The index tends to increase with increasing $\lambda$ as additional resonances are passed, but the index of refraction for visible wavelengths (in bold face) decreases with increasing wavelength.

The dispersion curves for optically transparent materials, such as glass and air, exhibit some very similar features, though the details may be very different. Starting at very short wavelengths ($\lambda \ll 0$), the refractive index $n \approx 1$. In words, the wavelength is so short, the frequency so large, and the energy so large that the photons pass through the material without interacting with the atoms. For nonzero, but still very short, wavelengths (“hard” X rays), the refractive index actually is slightly less than unity, which means that X rays incident on a prism are refracted away from the prism’s base, rather than towards the base in the manner of visible light. This is the reason why X rays can be totally reflected at grazing incidence, which is the focusing mechanism in X-ray telescopes (such as Chandra). As the wavelength of the incident light increases further, though still within the X-ray region, the radiation incident on the material will be heavily absorbed. This is the “K-absorption edge” where the energy of the incident X rays is just sufficient to ionize an electron in the innermost atomic “shell” – the “K shell.” The wavelength of this absorption is $\lambda_K \approx 0.67 \text{ nm}$.
for silicon. Other absorptions occur at yet longer wavelengths (smaller incident photon energies), where electrons in the L and M shells, etc., of the atom are ionized. The spectrum of a material with a large atomic number (and thus several filled electron shells) will exhibit several such resonant absorptions.

Ionization of a K-shell electron by an incoming X ray of sufficient energy. This is the reason for the large absorptions of “hard” X rays by materials. Lower-energy (longer-wavelength) X rays will ionize electrons in the L or M shells, thus producing other absorption “edges.”

As the wavelength of the incident radiation increases further, into the “far ultraviolet” region of the spectrum, the real part of the refractive index decreases to a value much less than unity within a wide band of anomalous dispersion. The fact that \( n < 1 \) in this region may be confusing because it seems that the velocity of light exceeds \( c \), but these waves do not propagate in the material due to the strong absorption (large value of \( \kappa \)). The wavelength of maximum absorption corresponds to the largest of the several “natural oscillation frequencies” of bound electrons in the material.

In the visible region of the spectrum, the dispersion curve exhibits the familiar decrease in \( n \) with \( \lambda \) that was shown above. For example, the index of air is \( n \approx 1.000279 \) at \( \lambda = 486.1 \) nm (Fraunhofer’s “F” line) and \( n \approx 1.000276 \) at \( \lambda = 656.3 \) nm (“C” line). The corresponding values for diamond are \( n_F = 2.4354 \) and \( n_C = 2.4100 \). The closer the nearest ultraviolet absorption to the visible spectrum, the steeper will be the slope \( \frac{dn}{d\lambda} \) in the visible region and thus the larger the visible dispersion (defined below).

The dispersion curve descends yet more steeply somewhere in the near infrared region and then rises due to anomalous dispersion in the vicinity of an infrared absorption band (labeled “\( \lambda_2 \)” on the graph). For quartz (crystalline SiO\(_2\)), the center of this band is located at \( \lambda \approx 8.5 \) \( \mu \)m, but the absorption already is quite strong for wavelengths as short as \( \lambda \approx 4 \) \( \mu \)m. Most optical materials have several such infrared absorption bands and the “base level” of the index of refraction is larger after each such band. This behavior is confirmed by far-infrared measurements of the refractive index of quartz (crystalline SiO\(_2\)), which varies over the interval \( 2.40 \leq n \leq 2.14 \) for \( 51 \mu \)m \( \leq \lambda \leq 63 \) \( \mu \)m. The large values of \( n \) ensure that the focal length of a convex quartz lens is much shorter at far-infrared wavelengths than at visible wavelengths.

As the wavelength is increased still further into the radio region of the spectrum, eventually all of the absorption bands will be passed. The refractive index decreases slowly and approaches a limiting value of \( \sqrt{\frac{c}{\epsilon_0}} \).

**2.1.5 Empirical Expressions for Dispersion**

To a first approximation, the indices of refraction of optical materials in the visible region of the spectrum vary as \( \lambda^{-1} \), which allows us to write an empirical expression for the refractivity of the medium \( n - 1 \):

\[
n[\lambda] - 1 \approx a + \frac{b}{\lambda}
\]

where \( a \) and \( b \) are parameters determined from measurements. Clearly the refractive index \( n \approx a = \sqrt{\frac{c}{\epsilon_0}} \) as \( \lambda \to +\infty \). The observation that the index decreases with increasing \( \lambda \) ensures that \( b > 0 \).
Cauchy’s Equation

Cauchy came up with a better empirical relation for the refractivity by adding some additional parameters \( A, B, C, \ldots \):

\[
n_\lambda - 1 \approx A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \cdots
\]

The parameters are determined by measurements of \( n \) for the corresponding material. Again, \( A = \sqrt{\frac{c}{\varepsilon_0}} \) is the refractive index for long wavelengths \( (\lambda \rightarrow +\infty) \). To determine three empirical constants, it obviously is necessary to measure \( n \) at three wavelengths. Again, the behavior of normal dispersion ensures that \( A \) and \( B \) are both positive.

The dispersion (variation of \( n \) with wavelength \( \lambda \)) may be expressed as the derivative of \( n \) with respect to wavelength \( \lambda \):

\[
\frac{dn}{d\lambda} \approx -2B - 4C + \cdots \approx -2B + \cdots \text{ if } C \lesssim B
\]

which again agrees with the observation that the variation in refractive index becomes “flatter” for longer wavelengths.

Sellmeier’s Equation

These two approximations for \( n_\lambda \) both follow “smooth” curves that decrease with increasing wavelength, and thus do not model anomalous dispersion. Sellmeier proposed an equation for dispersion in 1871 that includes an empirically fixed wavelength where the electrons in the material oscillate at their “natural” or “resonant” frequency \( \omega_0 \):

\[
n_\lambda = \sqrt{1 + \frac{A\lambda^2}{\lambda^2 - \lambda_0^2}} = \sqrt{1 + \frac{A}{1 - \left(\frac{\lambda_0}{\lambda}\right)^2}}
\]

where the two constants \( A > 1 \) and \( \lambda_0 = \frac{2\pi\omega_0}{\omega_0} \) are determined empirically. The denominator \( 1 - \left(\frac{\lambda_0}{\lambda}\right)^2 < 0 \) and \( n_\lambda < 1 \) for wavelengths less than \( \lambda_0 \); the computed \( n_\lambda > 1 \) for \( \lambda > \lambda_0 \). This expression exhibits the behavior of the refractive index in regions of anomalous dispersion, as shown in the figure.
2.1 LORENTZ’S MODEL OF REFRACTIVE INDEX

Index of refraction model by Sellmeier, showing anomalous dispersion in the vicinity of the resonance at $\lambda = \lambda_0$.

Multiple resonances may be accommodated by adding terms that resonate at their specific wavelengths:

$$n[\lambda] = \sqrt{\frac{1 + \sum_{i=1}^{N} \frac{A_i \lambda^2}{\lambda^2 - \lambda_i^2}}{1 + \sum_{i=1}^{N} \frac{A_i}{1 - (\lambda/\lambda_i)^2}}}$$

In the limit $\lambda \to +\infty$, this evaluates to:

$$n[\lambda \to \infty] = \sqrt{1 + \sum_{i=1}^{N} A_i} = \sqrt{\frac{\varepsilon}{\varepsilon_0}}$$

which, in words, shows that the index of refraction at “DC” ($\omega = 0$) includes the sum of the weighting constants for each resonance.

**Refractive Constants for Glasses**

The refractive properties of glass are approximately specified by the refractivity and the measured differences in refractive index at the three Fraunhofer wavelengths F, D, and C:

<table>
<thead>
<tr>
<th>Refractivity</th>
<th>$n_D - 1$</th>
<th>$1.75 \leq n_D \leq 1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Dispersion</td>
<td>$n_F - n_C &gt; 0$</td>
<td>differences between blue and red indices</td>
</tr>
<tr>
<td>Partial Dispersion</td>
<td>$n_D - n_C &gt; 0$</td>
<td>differences between yellow and red indices</td>
</tr>
<tr>
<td>Abbé Number</td>
<td>$\nu \equiv \frac{n_D - 1}{n_F - n_C}$</td>
<td>ratio of refractivity and mean dispersion, $25 \leq \nu \leq 65$</td>
</tr>
</tbody>
</table>

Glasses are specified by six-digit numbers abcd ef, where $n_D = 1.abc$, to three decimal places, and $\nu = de.f$. Note that larger values of the refractivity mean that the refractive index is larger and thus so is the deviation angle in Snell’s law. A larger Abbé number means that the mean dispersion is smaller and thus there will be a smaller difference in the angles of refraction. Such glasses with larger Abbé numbers and smaller indices and less dispersion are crown glasses, while glasses with smaller Abbé numbers are flint glasses, which are “denser”. Examples of glass specifications include Borosilicate crown glass (BSC), which has a specification number of 517645, so its refractive index in the D line is 1.517 and its Abbé number is $\nu = 64.5$. The specification number for a common flint glass is 619364, so $n_D = 1.619$ (relatively large) and $\nu = 36.4$ (smallish). Now consider the refractive indices in the three lines for two different glasses: “crown” (with a smaller $n$) and “flint:”

<table>
<thead>
<tr>
<th>Line</th>
<th>$\lambda$ [nm]</th>
<th>$n$ for Crown</th>
<th>$n$ for Flint</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>656.28</td>
<td>1.51418</td>
<td>1.69427</td>
</tr>
<tr>
<td>D</td>
<td>589.59</td>
<td>1.51666</td>
<td>1.70100</td>
</tr>
<tr>
<td>F</td>
<td>486.13</td>
<td>1.52225</td>
<td>1.71748</td>
</tr>
</tbody>
</table>

The glass specification numbers for the two glasses are evaluated to be:

For the crown glass:

- **Refractivity**: $n_D - 1 = 0.51666 \approx 0.517$
- **Abbé number**: $\nu = \frac{1.51666 - 1}{1.52225 - 1.51418} \approx 64.0$

**Glass number = 517640**
For the flint glass:
refractivity: \( L n_D - 1 = 0.70100 \equiv 0.701 \)
Abbé number: \( \nu = \frac{0.70100 - 1}{1.7148 - 1.69427} \equiv 30.2 \)

Glass number = 701302

### 2.1.6 Negative Refractive Index

In 1968, a Russian physicist named Victor Veselago presented theoretical arguments that the index of refraction need not be positive (V.G. Veselago, “The electrodynamics of substances with simultaneously negative values of \( \varepsilon \) and \( \mu \),” Sov. Phys. Uspekhi 10, 509-514, 1968). We have already demonstrated that the refractive index is related to the electrical permittivity and magnetic permeability:

\[
c = \sqrt{\frac{1}{\mu_0 \varepsilon_0}} \\
\nu = \sqrt{\frac{1}{\mu \varepsilon}} \\
n = \frac{c}{\nu} = \sqrt{\frac{\mu}{\mu_0} \cdot \frac{\varepsilon}{\varepsilon_0}}
\]

The presence of the square root has always been taken to imply that \( n > 0 \) (and, in fact, that \( n \geq 1 \) in real materials). Veselago showed that if both \( \varepsilon \) and \( \mu \) are negative, then the negative sign must be taken:

\[
n = -\sqrt{\frac{\mu}{\mu_0} \cdot \frac{\varepsilon}{\varepsilon_0}} \text{ if } \varepsilon < 0 \text{ and } \mu < 0
\]

In the case \( n < 0 \), the direction of propagation is reversed so that light travels backwards towards the source. This also implies that the laws of geometrical optics are “flipped;” light emerging from a “rare” medium to a “dense” medium is refracted away from the normal, rather than towards it.

Materials whose electromagnetic properties arise from features of the internal structure, rather than from the constituent elements, are metamaterials. An analogy is a material that exhibits apparent color due to interference from reflective coatings or structures within the material. One example of a metamaterial with a negative index was recently announced for which \( n = -0.6 \) at \( \lambda = 780 \text{ nm} \). These metamaterials are said by some to be the key technology for making “cloaking devices” that would make an object invisible by forcing light from behind to propagate “around” it.

Chapter 3

Polarization

Maxwell’s equations demonstrated that light is a transverse wave (as opposed to longitudinal waves, e.g., sound). Both the \( \mathbf{E} \) and \( \mathbf{B} \) vectors are perpendicular to the direction of propagation of the radiation. Even before Maxwell, Thomas Young inferred the transverse character of light in 1817 when he passed light through a calcite crystal (calcium carbonate, CaCO\textsubscript{3}). Two beams emerged from the crystal, which Young brilliantly deduced were orthogonally polarized, i.e., the \( \mathbf{E} \) vectors of the two beams oscillate in orthogonal directions.

The polarization of electromagnetic radiation is defined by the plane of vibration of the electric vector \( \mathbf{E} \) because the effect of the magnetic field on a free charge (electron) is much smaller. This is seen from the factor of \( c^{-1} \) applied to the magnetic field Lorentz force law:

\[
\mathbf{F} \propto q_0 \left( \mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B} \right)
\]

\( q_0 = \) charge [coulombs]

\( \mathbf{F} = \) force on the charge [measured in newtons: \( 1 \text{N} = 1 \frac{\text{kg} \cdot \text{m}}{\text{s}^2} \)]

\( \mathbf{v} = \) velocity of the charge \( q_0 \) [\( \frac{\text{m}}{\text{s}} \)]

\( c = \) velocity of light \( [3 \cdot 10^8 \text{m/s}] \)

3.1 Plane Polarization = Linear Polarization

In the most familiar type of polarization, the \( \mathbf{E} \)-vector oscillates in the same plane at all points on the wave. This is called planar or linear polarization. Any state of linear polarization can be expressed as a linear combination (sum) of two orthogonal states (basis states), e.g., the \( x \)- and \( y \)-components of the \( \mathbf{E} \)-vector for a wave traveling toward \( z = \pm \infty \):

\[
\mathbf{E} = \mathbf{E}[\mathbf{r},t] = [\hat{x}E_x + \hat{y}E_y] \cos[k_0z - \omega_0t + \phi]
\]

\( \hat{x}, \hat{y} = \) unit vectors along \( x \) and \( y \)

\( E_x, E_y = \) amplitudes of the \( x \)- and \( y \)-components of \( \mathbf{E} \)

For a wave of amplitude \( E_0 \) polarized at an angle \( \theta \) relative to the \( x \)-axis:
CHAPTER 3 POLARIZATION

\[ E_x = E_0 \cos[\theta] \]
\[ E_y = E_0 \sin[\theta] \]
\[ E_0 = \sqrt{E_x^2 + E_y^2} \]
\[ \theta = \tan^{-1}\left(\frac{E_y}{E_x}\right) \]

Linearly polarized radiation oscillates in the same plane at all times and at all points in space. Especially note that \( E_x \) and \( E_y \) are in phase for linearly polarized light, i.e., both components have zero-crossings and have extrema at the same points in time and space.

### 3.2 Circular Polarization

If the \( \mathbf{E} \)-vector describes a helical motion in space, the projection of the \( \mathbf{E} \)-vector onto a plane normal to the propagation direction \( \mathbf{k} \) exhibits circular motion over time, hence the name for this polarization:

Circular polarization occurs when the electric fields along orthogonal axes have the same amplitude by their phases differ by \( \pm \frac{\pi}{2} \) radians.
3.2 CIRCULAR POLARIZATION

To an observer sitting at a fixed point in space \((z = z_0)\), the motion of the \(\mathbf{E}\)-vector is the sum of two orthogonal linearly polarized states with one component delayed in phase by \(90^\circ = \frac{\pi}{2}\) radians. The math is identical to that used to describe oscillator motion as the projection of rotary motion:

\[
motion = \hat{x} \cos [k_0 z_0 - \omega_0 t] + \hat{y} \cos \left[k_0 z_0 - \omega_0 t + \frac{\pi}{2}\right] = \hat{x} \cos [\omega_0 t] \pm \hat{y} \sin [\omega_0 t]
\]

We can consider the constant \(k_0 z_0\) to be an initial phase value \(0\):

\[
motion = \hat{x} \cos [\omega_0 t - \phi_0] \pm \hat{y} \sin [\omega_0 t - \phi_0]
\]

where the upper sign applies to right-handed circular polarization (angular momentum convention).

3.2.1 Nomenclature for Circular Polarization

Like linearly polarized light, circularly polarized light has two orthogonal states, i.e., clockwise and counterclockwise rotation of the \(\mathbf{E}\)-vector. These are termed right-handed (RHCP) and left-handed (LHCP). There are two conventions for the nomenclature:

1. Angular Momentum Convention (my preference because easier to visualize): Point the thumb of the \(\{\text{right, left}\}\) hand in the direction of propagation. If the fingers point in the direction of rotation of the \(\mathbf{E}\)-vector, then the light is \(\{\text{RHCP, LHCP}\}\).

2. Optics (also called screw-shape or screwy) Convention: The path traveled by the \(\mathbf{E}\)-vector of RHCP light is the same path described by a right-hand screw. Of course, the natural laws defined by Murphy ensure that the two conventions are opposite: RHCP light by the angular momentum convention is LHCP by the screw convention.

3.2.2 Elliptical Polarization, Reflections

If the amplitudes of the \(x\)- and \(y\)-components of the \(\mathbf{E}\)-vector are not equal, or if the phase difference is not \(\pm \frac{\pi}{2} = \pm 90^\circ\), then the projection of the path of the \(\mathbf{E}\)-vector is an ellipse and the polarization (obviously) is elliptical. Note that elliptical polarization may be either right- or left-handed, as defined above.

3.2.3 Change of Handedness on Reflection

By conservation of angular momentum, the direction of rotation of the \(\mathbf{E}\)-vector does not change on reflection. Since the direction of propagation reverses, the handedness of the circular or elliptical polarization changes:

![Change in “handedness” of a circularly (or elliptically) polarized wave upon reflection by a mirror.](image-url)
3.2.4 Natural Light

The superposition of emissions from a large number of thermal source elements (as in a light bulb) has a random orientation of polarizations. The state of polarization of the resulting light changes direction randomly over very short time intervals ($\approx 10^{-14}$ s). The radiation is *unpolarized*, even though the electric field oscillates in one plane if measured over this short time period. Natural light is neither totally polarized nor totally unpolarized, but rather *partially polarized*.

3.3 Description of Polarization States

3.3.1 Jones Vector

The components of the electric field in the two orthogonal directions may be used to represent a vector with complex components. This is called a Jones vector, which is useful only for completely polarized light.

Assume that light propagates down the $z$-axis (or rotate the coordinate system to ensure this situation). The electric field is:

$$\mathbf{E} = \text{Re} \{ \mathbf{E}_0 \exp [+i (k_0 z - \omega_0 t)] \} = [\text{Re} \{ E_x \exp [+i (k_0 z - \omega_0 t)] \}, \text{Re} \{ E_y \exp [+i (k_0 z - \omega_0 t - \delta)] \}]$$

$$= \text{Re} \left\{ [E_x, E_y e^{-i\delta}] \exp [+i (k_0 z - \omega_0 t)] \right\}$$

The Jones vector that describes this state has components equal to the amplitudes in the orthogonal directions, plus any phase delay $\delta$ in one of the components (generally taken to be the $y$-component):

$$\mathbf{\mathcal{E}} = \begin{bmatrix} E_x \\ E_y e^{-i\delta} \end{bmatrix}$$

The total irradiance is the sum of the irradiances in orthogonal polarizations:

$$I_{\text{total}} = I_x + I_y = \langle E_x E_x^* \rangle + \langle E_y E_y^* \rangle = \langle \mathbf{E} \cdot \mathbf{E}^* \rangle = \langle \mathbf{\mathcal{E}} \cdot \mathbf{\mathcal{E}}^\dagger \rangle$$

where the “dagger” is used to indicate the Hermitian conjugate vector, which is the complex conjugate of the transpose:

$$\mathbf{\mathcal{E}}^\dagger \equiv (\mathbf{\mathcal{E}}^*)^T = (\mathbf{\mathcal{E}}^T)^*$$

In words, the irradiance is the time average of the scalar product of the Jones vector with its complex conjugate.

**Examples of Jones Vectors:**

1. Plane-polarized light along $x$-axis

$$\mathbf{\mathcal{E}} = \begin{bmatrix} E_0 \\ 0 \end{bmatrix}$$

2. Plane-polarized light along $y$-axis:

$$\mathbf{\mathcal{E}} = \begin{bmatrix} 0 \\ E_0 \end{bmatrix}$$
3. Plane-polarized light at angle $\theta$ to $x$-axis:

$$\mathbf{E} = \begin{bmatrix} E_0 \cos \theta \\ E_0 \sin \theta \end{bmatrix}$$

4. RHCP

$$\mathbf{E} = \hat{x} E_0 \cos [k_0 z - \omega_0 t] + \hat{y} E_0 \sin [k_0 z - \omega_0 t]$$

$$= \hat{x} E_0 \cos [k_0 z - \omega_0 t] + \hat{y} E_0 \cos \left[ k_0 z - \omega_0 t - \frac{\pi}{2} \right] \implies \delta = +\frac{\pi}{2}$$

($\delta = +\frac{\pi}{2}$ because of negative sign in phase of Jones vector)

$$= \text{Re} \left\{ \begin{bmatrix} E_0 \\ E_0 \exp \left[ +\frac{i\pi}{2} \right] \end{bmatrix} \exp \left[ +i (k_0 z - \omega_0 t) \right] \right\}$$

$$\implies \mathbf{E} = E_0 \begin{bmatrix} 1 \\ \exp \left[ +\frac{i\pi}{2} \right] \end{bmatrix} = E_0 \begin{bmatrix} 1 \\ +i \end{bmatrix} \text{ for RHCP}$$

so the Jones vector for LHCP obviously is:

$$\mathbf{E} = E_0 \begin{bmatrix} 1 \\ \exp \left[ -\frac{i\pi}{2} \right] \end{bmatrix} = E_0 \begin{bmatrix} 1 \\ -i \end{bmatrix} \text{ for LHCP}$$

The Jones vectors of two orthogonally polarized states satisfies the condition:

$$\mathbf{E}_1 \cdot \mathbf{E}_2^* = 0$$

Other representations of the state of polarization are available (e.g., Stokes' parameters, coherency matrix, Mueller matrix, Poincare sphere). They are more complicated, and hence more useful, i.e., they can describe partially polarized states. For more information, see (for example), *Polarized Light* by Shurcliff.

### 3.3.2 Actions of Optical Elements on Jones Vectors

We can now imagine the action of an optic on a state of polarization expressed by the corresponding Jones vector. The optic is expressed as a $2 \times 2$ matrix $\mathbf{M}$, where again we recall that Jones vectors may only represent completely polarized light. The form of the operation is:

$$\mathbf{M} \mathbf{E} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} E_x \\ E_y e^{+i\delta} \end{bmatrix}$$

We can infer the form of the matrix for some different cases. If the incident light is linearly polarized along $x$ and the optic passes only light that is linearly polarized along $x$, then all of the light should pass to the output:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} E_x \\ 0 \end{bmatrix} = \begin{bmatrix} E_x \\ 0 \end{bmatrix}$$

$$\implies a = 1, c = 0 \implies \mathbf{M}_{g=0} = \begin{bmatrix} 1 & b \\ 0 & d \end{bmatrix}$$
If light that is linearly polarized along $y$ is incident on the same polarizer, the output should be zero:

$$\begin{bmatrix} 1 & b \\ 0 & d \end{bmatrix} \begin{bmatrix} 0 \\ E_y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\implies b = 0, d = 0 \implies M_{\theta=0} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

If the linear polarizer is oriented along the $y$ direction, it is easy to show that:

$$M_{\theta=\frac{\pi}{2}} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Note that $\det M_{\theta=0} = \det M_{\theta=\frac{\pi}{2}} = 0$

If the $x$-matrix is applied to light that is linearly polarized at $\theta = \pm \frac{\pi}{4}$, the output is:

$$M_{\theta=0} E_x = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_0 \sqrt{2} \\ I_0 \sqrt{2} \end{bmatrix} = \begin{bmatrix} I_0 \sqrt{2} \\ 0 \end{bmatrix}$$

A linear polarizer oriented at $\theta$ measured relative to the $x$-axis may be derived in similar fashion:

$$M_{\theta} E = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}$$

$$M_{\theta} E = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \cos \theta \pm \frac{\pi}{4} \\ \sin \theta \pm \frac{\pi}{4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$M_{\theta} E_0 \begin{bmatrix} \cos \theta \pm \frac{\pi}{4} \\ \sin \theta \pm \frac{\pi}{4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

These two relations may be solved simultaneously:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \cos \theta \pm \sin \theta \\ \sin \theta \pm \cos \theta \end{bmatrix} = \begin{bmatrix} \cos \theta & 0 \\ \sin \theta & 0 \end{bmatrix}$$
3.3 DESCRIPTION OF POLARIZATION STATES

which means that:

\[
\mathbf{M}_\theta = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \cos \theta & 0 \\ \sin \theta & 0 \end{bmatrix} \cdot \begin{bmatrix} \cos \theta & \mp \sin \theta \\ \sin \theta & \pm \cos \theta \end{bmatrix}^{-1}
\]

\[
= \begin{bmatrix} \cos \theta & 0 \\ \sin \theta & 0 \end{bmatrix} \cdot \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}
\]

\[
= \begin{bmatrix} \cos^2 \theta & \sin \theta \cos \theta \\ \sin \theta \cos \theta & \sin^2 \theta \end{bmatrix}
\]

Note that \(\det \mathbf{M}_\theta = 0\), as was true for \(\mathbf{M}_{\theta=0}\) and \(\mathbf{M}_{\theta=\pi/2}\).

A polarization rotator will convert light that is linearly polarized at angle \(\theta\) and rotate it to \(\theta + \beta\), so that:

\[
M_\beta = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \cos \beta \\ \sin \beta \end{bmatrix} = \begin{bmatrix} \cos (\theta + \beta) \\ \sin (\theta + \beta) \end{bmatrix}
\]

\[
\Rightarrow \begin{bmatrix} a \cos \beta + b \sin \beta \\ c \cos \beta + d \sin \beta \end{bmatrix} = \begin{bmatrix} \cos \theta \cos \beta - \sin \theta \sin \beta \\ \sin \theta \cos \beta + \cos \theta \sin \beta \end{bmatrix}
\]

\[
\Rightarrow a = \cos \beta, \quad b = -\sin \beta, \quad c = +\sin \beta, \quad d = \cos \beta
\]

\[
M_\beta = \begin{bmatrix} \cos \beta & -\sin \beta \\ +\sin \beta & \cos \beta \end{bmatrix}
\]

\[
\det [M_\beta] = 1
\]

We can also retard the phase of one component more or less than the other:

\[
\mathbf{M}_{\Delta \phi} = \begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} E_x e^{+i\phi_x} \\ E_y e^{+i\phi_y} \end{bmatrix}
\]

\[
\Rightarrow \mathbf{M}_{\Delta \phi} = \begin{bmatrix} e^{+i\phi_x} & 0 \\ 0 & e^{+i\phi_y} \end{bmatrix}
\]

so that \(\det [\mathbf{M}_{\Delta \phi}] = e^{+i\phi_x} \cdot e^{+i\phi_y} = \exp [+i(\phi_x + \phi)]\).

A quarter-wave retarder changes the phase of one axis by \(\pm \frac{\pi}{2}\) radians:

\[
\mathbf{M}_{\Delta \phi} = \begin{bmatrix} 1 & 0 \\ 0 & e^{\pm i\frac{\pi}{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \pm i \end{bmatrix}
\]

A quarter-wave plate acting on light that is linearly polarized at \(\theta = 45^\circ\) yields:

\[
\begin{bmatrix} 1 & 0 \\ 0 & \pm i \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \pm i \end{bmatrix}
\]

which is circularly polarized.
3.3.3 Coherency Matrix $\mathbf{J}$

For light with amplitudes $E_x$ and $E_y$ in the $x$ and $y$ directions, respectively, the coherency matrix $\mathbf{J}$ is the $2 \times 2$ array formed from the outer product of time averages of the Jones vector:

$$\mathbf{J} = \left( \begin{array}{c} E_x \\ E_y \end{array} \right) \left( \begin{array}{c} E_x^* \\ E_y^* \end{array} \right)^*$$

Note that the coherency matrix is identical to the complex conjugate of its transpose: it is a Hermitian matrix:

$$\mathbf{J} = (\mathbf{J}^T)^* = \mathbf{J}^\dagger$$

The irradiance $I$ is the sum of the diagonal terms, which is called the trace of the matrix:

$$I = \langle |E_x|^2 \rangle + \langle |E_y|^2 \rangle = \langle E_x E_x^* \rangle + \langle E_y E_y^* \rangle = tr[\mathbf{J}]$$

Coherency Matrix of Completely Linearly Polarized Light

The coherency matrix of completely polarized light has the form:

$$\mathbf{J} = \begin{bmatrix} \langle E_x^2 \rangle & \langle E_x E_y^* \rangle \\ \langle E_x^* E_y \rangle & \langle E_y^2 \rangle \end{bmatrix}$$

Note that

$$\det \mathbf{J} = \langle E_x^2 \rangle \langle E_y^2 \rangle - \langle E_x^* E_y \rangle \langle E_x E_y^* \rangle = 0$$

for linearly polarized light.

We can rotate the coordinate system of the matrix $\mathbf{J}$ to find the equivalent diagonal form (call it $\mathbf{\Omega}$), where the diagonal elements are the eigenvalues. Since $\mathbf{J}$ is only $2 \times 2$, the eigenvalues are easily derived using the well-known “brute force” method. We set the determinant of the “secular equation” to zero and solve the quadratic equation to find two values for $\lambda$:

$$\det [\mathbf{J} - \lambda \mathbf{I}] = 0$$

$$= \det \begin{bmatrix} J_{xx} - \lambda & J_{xy} \\ J_{yx} & J_{yy} - \lambda \end{bmatrix}$$

$$= (J_{xx} - \lambda)(J_{yy} - \lambda) - J_{xy}J_{yx}$$

$$= \lambda^2 - \lambda(J_{xx} + J_{yy}) + (J_{xx}J_{yy} - J_{xy}J_{yx})$$

$$= \lambda^2 - \lambda (tr[\mathbf{J}]) + \det \mathbf{J} = 0$$

where $tr[\mathbf{J}]$ is the sum of the diagonal elements of $\mathbf{J}$. The eigenvalues are the solutions of this
3.3 DESCRIPTION OF POLARIZATION STATES

\[
\lambda_+ = \frac{\text{tr}[\mathbf{J}]}{2} \left( 1 + \sqrt{1 - \frac{4 \cdot \text{det}[\mathbf{J}]}{\text{tr}[\mathbf{J}]}} \right) \\
\lambda_- = \frac{\text{tr}[\mathbf{J}]}{2} \left( 1 - \sqrt{1 - \frac{4 \cdot \text{det}[\mathbf{J}]}{\text{tr}[\mathbf{J}]}} \right)
\]

so that \( \lambda_+ \geq \lambda_- \) and the diagonal form is:

\[
\mathbf{\Omega} = \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix}
\]

For completely linearly polarized light, \( \lambda_+ = I_{\text{total}} \) and \( \lambda_- = 0 \), so the diagonal form of the coherency matrix of linearly polarized light is:

\[
\mathbf{\Omega} = \begin{bmatrix} I_{\text{total}} & 0 \\ 0 & 0 \end{bmatrix}
\]

**Coherency Matrix of Unpolarized Light**

The two electric field amplitudes of unpolarized light are uncorrelated, so that \( \langle E_x E_y e^{+i\delta} \rangle = \langle E_x E_y e^{-i\delta} \rangle = 0 \) and the coherency matrix reduces to:

\[
\mathbf{J} = \frac{1}{2} I_{\text{total}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\]

which is diagonal, so the eigenvalues of the coherency matrix for unpolarized light are equal:

\[
\lambda_+ = \lambda_- = \frac{1}{2} I_{\text{total}}
\]

and \( \text{det}\mathbf{J} = I_{\text{total}} \).

**Examples:**

- LP at angle \( \theta \)

\[
\mathbf{J} = \langle \mathbf{E} \mathbf{E}^\dagger \rangle = \left\langle \begin{bmatrix} E_0 \cos[\theta] \\ E_0 \sin[\theta] \end{bmatrix} \begin{bmatrix} E_0^* \cos[\theta] & E_0^* \sin[\theta] \end{bmatrix} \right\rangle = \langle |E_0|^2 \rangle \begin{bmatrix} \cos^2[\theta] & \cos[\theta] \sin[\theta] \\ \cos[\theta] \sin[\theta] & \sin^2[\theta] \end{bmatrix}
\]

\[
\mathbf{J}_{LP, \theta} = I_0 \begin{bmatrix} \cos^2[\theta] & \cos[\theta] \sin[\theta] \\ \cos[\theta] \sin[\theta] & \sin^2[\theta] \end{bmatrix}
\]

\[
\text{det}\mathbf{J} = \cos^2[\theta] \cdot \sin^2[\theta] - (\cos[\theta] \sin[\theta])^2 = 0
\]
• Horizontal LP Light (along $x$)

$$\mathbf{J}_{LP} x = \left\langle \mathbf{E} \mathbf{E}^\dagger \right\rangle = I_0 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\det \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = 0$$

• Vertical LP (along $y$)

$$\mathbf{J}_{LP} y = I_0 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\det \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

• Diagonal LP at $\theta = +\frac{\pi}{4}$

$$\mathbf{J}_{LP} +\frac{\pi}{4} = I_0 \begin{bmatrix} \cos^2 [\theta] & \cos [\theta] \sin [\theta] \\ \cos [\theta] \sin [\theta] & \sin^2 [\theta] \end{bmatrix} \bigg|_{\theta = +\frac{\pi}{4}}$$

$$= I_0 \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = I_0 \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\det \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = 0$$

Diagonal LP at $\theta = -\frac{\pi}{4}$

$$\mathbf{J}_{LP} -\frac{\pi}{4} = \left\langle \mathbf{E} \mathbf{E}^\dagger \right\rangle$$

$$= I_0 \begin{bmatrix} \cos^2 [\theta] & \cos [\theta] \sin [\theta] \\ \cos [\theta] \sin [\theta] & \sin^2 [\theta] \end{bmatrix} \bigg|_{\theta = -\frac{\pi}{4}}$$

$$= I_0 \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = I_0 \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\det \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = 0$$
3.3 DESCRIPTION OF POLARIZATION STATES

- RHCP

\[
\mathbf{J}_{\text{RHCP}} = \frac{I_0}{2} \begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}
\]

\[
\det \begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix} = 0
\]

- LHCP

\[
\mathbf{J}_{\text{LHCP}} = \frac{I_0}{2} \begin{bmatrix} 1 & +i \\ -i & 1 \end{bmatrix}
\]

Coherency Matrix of Partially Polarized Light

In this case we decompose the matrix into a weighted sum of polarized and unpolarized pieces. This is easy to see in the diagonal forms:

\[
\mathbf{J} = \begin{bmatrix} \langle E_x^2 \rangle & \langle E_x E_y e^{+i\delta} \rangle \\ \langle E_x E_y e^{-i\delta} \rangle & \langle E_y^2 \rangle \end{bmatrix}
\]

\[
\mathbf{\Omega} = \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix} = \begin{bmatrix} \lambda_+ - \lambda_- & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \lambda_- & 0 \\ 0 & \lambda_+ \end{bmatrix}
\]

\[
= (\lambda_+ - \lambda_-) \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \lambda_- \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\]

The degree of polarization \( \mathcal{P} \) is the ratio of the polarized irradiance to the total irradiance as obtained from the eigenvalues:

\[
\mathcal{P} = \frac{I_{\text{polarized}}}{I_{\text{total}}} = \frac{\lambda_+ - \lambda_-}{\lambda_+ + \lambda_-}
\]

\[
= + \sqrt{1 - \frac{4 \cdot \det [\mathbf{J}]}{(\text{tr} [\mathbf{J}])^2}}
\]

If either \( \lambda_+ = 0 \) or \( \lambda_- = 0 \), then \( \mathcal{P} = 1 \) and the light is completely polarized; if \( \lambda_+ = \lambda_- \), then \( \mathcal{P} = 0 \) and the light is completely unpolarized. If \( \lambda_+ \neq \lambda_- \), the light is partially polarized.

Action of Optic on Coherency Matrix

The action of the optical element on a coherency matrix \( \mathbf{J} \) is:

\[
\mathbf{J} \mathbf{L} \mathbf{L}^\dagger = \mathbf{J}'
\]
where the output irradiance is again the invariant trace of the matrix, $\text{tr} \left[ \mathbf{J} \right] = \text{tr} \left[ \mathbf{J}' \right]$. 

Example: Rotation of Plane of Polarization:

\[ \mathbf{L} = \mathbf{R} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \]

\[ \mathbf{E}' = \mathbf{L} \mathbf{E} = \begin{bmatrix} E_x \cos \theta + E_y \sin \theta \\ -E_x \cos \theta + E_y \sin \theta \end{bmatrix} \]

\[ I_{\text{total}} = \sqrt{E_x^2 + E_y^2} \]

\[ I'_{\text{total}} = \left< \left( E_x \cos \theta + E_y \sin \theta \right)^2 + \left( -E_x \cos \theta + E_y \sin \theta \right)^2 \right> \]

\[ = \left( E_x \cos \theta + E_y \sin \theta \right)^2 + \left( -E_x \cos \theta + E_y \sin \theta \right)^2 \]
3.4 Generation of Polarized Light

3.4.1 Selective Emission:

If all emitting elements of a source (e.g., electrons in a bulb filament) vibrate in the same direction, the radiated light will be polarized in that direction. This is difficult to achieve at optical frequencies ($\Delta t \lesssim 10^{-14} \text{s} \implies \nu \gtrsim 10^{14} \text{Hz}$), but is easy at radio or microwave frequencies ($\nu \lesssim 10^{8} \text{Hz}$) by proper design of the antenna that radiates the energy. For example, a radio-frequency oscillator attached to a simple antenna forces the free electrons in the antenna to oscillate along the long (vertical) dimension of the antenna. The emitted radiation is therefore mostly oscillating in the vertical direction; it is vertically polarized.

\[\text{Radio Frequency (RF) Oscillator (} \nu \sim 10^8 \text{Hz)}\]

Emission of electromagnetic radiation (“light”) by a “dipole” radiator is polarized in the direction of motion of the emitting electrons (vertical, in this case).

Rather than generating polarized light at the source, we can obtain light of a selected polarization from natural light by removing unwanted states of polarization. This is the mechanism used in the next section.

3.4.2 Selective Transmission or Absorption

A man-made device for selecting a state of polarization by selective absorption is Polaroid. This operates like the microwave-polarizing skein of wires. The wires are parallel to the $y$-axis in the figure. Radiation incident on the wires drives the free electrons in the wires in the direction of polarization of the radiation. The electrons driven in the $y$-direction along the surface of the wire and strike other such electrons, thus dissipating the energy in thermal collisions. What energy that is reradiated by such electrons is mostly directed back toward the source (reflected). The $x$-component of the polarization is not so affected, since the electrons in the wire are constrained against movement in that direction. The $x$-component of the radiation therefore passes nearly unaffected.

Common Polaroid sheet acts as a skein of wires for optical radiation. It is made from clear polyvinyl acetate which has been stretched in one direction to produce long chains of hydrocarbon molecules. The sheet is then immersed in iodine to supply lots of free electrons.

\[\text{Unpolarized Light} \rightarrow \text{Polarized Light} \rightarrow \text{Transmitted Polarization}\]

Polarization by “skein of wires” – the radiation polarized parallel to the direction of the wires in the skein is absorbed, so the radiation polarized \textbf{perpendicular to the wires} is transmitted. In other words, the “picket fence” model of polarization is not appropriate in this case.
3.4.3 Generating Polarized Light by Reflection – Brewster’s Angle

We have already observed that \( r_{TM} = 0 \) if the incident angle is Brewster’s angle \( \theta_B = \tan^{-1} \left( \frac{n_2}{n_1} \right) \).

One polarization of obliquely incident light “sees” the bound electrons in the dense medium “differently” and therefore is reflected “differently.” The reflected wave is polarized to some extent; the degree of polarization depends on the angle of incidence and the index of refraction \( n_2 \). The polarization mechanism is simply pictured as a forced electron oscillator. The bound electrons in the dielectric material are driven by the incident oscillating electric field of the radiation \( \mathbf{E} \exp \left[ +i \left( k_0 z_0 + \omega_0 t \right) \right] \), and hence vibrate at frequency \( \nu_0 = \frac{\omega_0}{2\pi} \). Due to its acceleration, the vibrating electron reradiates radiation at the same frequency \( \nu_0 \) to produce the reflected wave. The state of polarization of the reflected radiation is a function of the polarization state of the incident wave, the angle of incidence, and the indices of refraction on either side of the interface. If the reflected wave and the refracted wave are orthogonal (i.e., \( \theta_0 + \theta_t = 90^\circ \Rightarrow \theta_t = \frac{\pi}{2} - \theta_0 \)), then the reflected wave is completely plane polarized parallel to the surface (and thus polarized perpendicular to the plane of incidence). This angle appeared in the discussion of the reflectance coefficients in the previous section. In this case, the electrons driven in the plane of the incidence will not emit radiation at the angle required by the law of reflection. This angle of complete polarization is called Brewster’s Angle \( \theta_B \), which we mentioned earlier during the discussion of the Fresnel equations.

Brewster’s angle: the incident beam at \( \theta_0 = \theta_B \) is unpolarized. The reflectance coefficient for light polarized in the plane (TM waves) is 0, and the sum of the incident and refracted angle is \( 90^\circ = \frac{\pi}{2} \).

Thus \( \theta_B + \theta_t = \frac{\pi}{2} \Rightarrow \theta_t = \frac{\pi}{2} - \theta_B \).

3.4.4 Polarization by Scattering

Light impinging on an air molecule drives the electrons of the molecule in the direction of vibration of the electric field vector. This motion causes light to be reradiated in a dipole pattern; i.e., no light is emitted along the direction of electron vibration. If we look at scattered light (e.g., blue sky) at \( 90^\circ \) from the source, the light is completely linearly polarized. Note that if the light is multiply scattered, as in fog, each scattering disturbs the state of polarization and the overall linear state is perturbed into unpolarized radiation.
Scattering of sunlight by atmospheric molecules. The light scattered into the eye at an angle of 90° is completely linearly polarized perpendicular to the line from the sun to the point in the sky.

3.5 Birefringence – Double Refraction

H§8.4

Many natural crystals and man-made materials interact with the two orthogonal polarizations differently. This is often due to an anisotropy (nonuniformity) in the crystalline structure; such materials are called dichroic or birefringent. Many crystals (e.g., calcite) divide a nonpolarized light wave into two components with orthogonal polarizations. The two indices of refraction are sometimes denoted $n_f$ and $n_s$ for fast and slow axes, where $n_f < n_s$. They are also denoted $n_o$ and $n_e$ for ordinary and extraordinary axes. By dividing the incoming natural light into two beams in such a crystal, we can select one of the two polarizations.

3.5.1 Examples:

Refractive indices along the fast and slow axes at $\lambda = 589.3 \text{ nm}$

<table>
<thead>
<tr>
<th>Material</th>
<th>$n_s$</th>
<th>$n_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcite (CaCO$_3$)</td>
<td>1.6584</td>
<td>1.4864</td>
</tr>
<tr>
<td>Crystalline Quartz (SiO$_2$)</td>
<td>1.5534</td>
<td>1.5443</td>
</tr>
<tr>
<td>Ice (crystalline H$_2$O)</td>
<td>1.313</td>
<td>1.309</td>
</tr>
<tr>
<td>Rutile (TiO$_2$)</td>
<td>2.903</td>
<td>2.616</td>
</tr>
<tr>
<td>Sodium Nitrate (SiNO$_3$)</td>
<td>1.5854</td>
<td>1.3369</td>
</tr>
</tbody>
</table>

The wavelength of light in a medium is $\lambda' = \frac{\lambda}{n}$, so light along the two polarization directions have different wavelengths:

$$\lambda'_s = \frac{\lambda}{n_s} < \lambda'_f = \frac{\lambda}{n_f}$$

3.5.2 Phase Delays in Birefringent Materials; Wave Plates

Consider light incident on a birefringent material of thickness $d$. The electric field as a function of distance $z$ and time $t$ is:

$$\mathbf{E}[z,t] = (\hat{x} E_x + \hat{y} E_y) \exp \left[ i (\omega_0 t - k_0 z) \right].$$
At the input face of the material \( z = 0 \) and the output face \( z = d \), the fields are:

\[
\mathbf{E}[z = 0, t] = (\mathbf{\hat{x}}E_x + \mathbf{\hat{y}}E_y) \exp[-i\omega_0 t]
\]

\[
\mathbf{E}[z = d, t] = (\mathbf{\hat{x}}E_x + \mathbf{\hat{y}}E_y) \exp[+i(k_0 d - \omega_0 t)]
\]

If \( n_x = n_s > n_y = n_f \), then \( \lambda_f > \lambda_s \) and:

\[
k_s = k_x = \frac{2\pi n_s}{\lambda} > k_f = \frac{2\pi n_f}{\lambda}
\]

The field at the output face \( (z = d) \) is therefore:

\[
\mathbf{E}[d, t] = \left(\mathbf{\hat{x}}E_x + \mathbf{\hat{y}}E_y \exp\left[\frac{2\pi i}{\lambda} d(n_f - n_s)\right]\right) \exp\left[+i\frac{2\pi d n_s}{\lambda}\right]
\]

By defining a constant phase term \( \delta = \frac{2\pi}{\lambda} d(n_f - n_s) \), the electric field at the output face of the birefringent material can be expressed as:

\[
\mathbf{E}[d, t] = (\mathbf{\hat{x}}E_x + \mathbf{\hat{y}}E_y e^{+i\delta}) \exp\left[+i\frac{2\pi d n_s}{\lambda}\right]
\]

On emergence from the material, the \( y \)-component of the polarization has a different phase than the \( x \)-component; the phase difference is \( \delta \).

**Quarter-Wave Plate:**

If \( E_x = E_y \), (i.e., the incident wave is linearly polarized @ 45° to the \( x \)-axis), then the emerging wave is circularly polarized. This is the principle of the circular polarizer.

**Half-Wave Plate:**

If \( \delta = \pi \implies d = \frac{\lambda}{2(n_s - n_f)} \), and the relative phase delay is 180°. Such a device is a half-wave plate. If the incident light is linearly polarized along the orientation midway between the fast and slow axes, the plane of polarization of the exiting linearly polarized light is rotated by 90°.

**Circular Polarizer:**

A circular polarizer is a sandwich of a linear polarizer and a \( \frac{\lambda}{4} \) plate, where the polarizing axis is oriented midway between the fast and slow axes of the quarter-wave plate. The LP ensures that equal amplitudes exist along both axes of the quarter-wave plate, which delays one of the components to create circularly polarized light. Light incident from the back side of a circular polarizer is not circularly polarized on exit; rather it is linearly polarized. A circular polarizer can be recognized and properly oriented by placing it on a reflecting object (e.g., a dime). If the image of the coin is
dark, the polarizer has the linear polarizer on top. This is because the handedness of the light is changed on reflection; the light emerging from the plate is now linearly polarized perpendicular to the axis of the LP and no light escapes.

![Diagram of light polarization](image)

*A circular polarizer is a sandwich of a linear polarizer and a quarter-wave plate.*

### 3.5.3 Critical Angle – Total Internal Reflection

We mentioned the critical angle during the discussion of the Fresnel equations. From Snell, we have the relation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

If $n_1 > n_2$ then a specific angle $\theta_C$ exists that satisfies the condition where its sine is equal to the ratio of the two indices:

$$\frac{n_1}{n_2} \sin \theta_C = 1 \implies \sin \theta_C = \frac{n_2}{n_1} < 1 \implies \theta_2 = \frac{\pi}{2}$$

The outgoing ray is refracted parallel to the interface (the “surface”):

$$\text{Critical Angle: } \theta_C = \sin^{-1}\left[\frac{n_2}{n_1}\right]$$

[Reflection at a “dense-to-rare” interface for light incident at the critical angle $\theta_c$.]

For crown glass with $n_d = 1.52$, $\theta_C = \sin^{-1}\left[\frac{1}{1.52}\right] \approx 0.718$ radians $\approx 41^\circ$. For a common flint glass with $n_d = 1.70$, then $\theta_C \approx 0.629$ radians $\approx 36^\circ$. If the incident angle $\theta_1 > \theta_C$ and $n_1 > n_2$ (e.g., the first medium is glass and the second is air), then no real-valued solution for Snell’s law exists, and there is no refracted light. This is the well-known phenomenon of total internal reflection where all incident light is reflected at the interface.
This may be analyzed rigorously by applying Maxwell’s equations to show that the refracted angle $\theta_2$ is complex valued instead of real valued, so that the electromagnetic field is attenuated exponentially as it crosses the interface. In other words, the electric field decays so rapidly across the interface that no energy can flow across the boundary, and hence no light escapes. However, we can “frustrate” the total internal reflection by placing another medium (such as another piece of glass) within a few light wavelengths of the interface. If close enough to the boundary, then some electric field can get into the second glass and a refracted wave “escapes”.

Schematic of “frustrated total internal reflection”: some energy can “jump” across a small gap between two pieces of glass even though the incident angle exceeds the critical angle. As the width $\tau$ of the gap increases, then the quantity of energy coupled across the gap decreases very quickly.
Chapter 4

Electromagnetic Waves at an Interface

In optics, and particularly in imaging, we are interested in evaluating electromagnetic fields on both sides of an interface between two media, e.g., air and glass. We use these properties to design particular shapes of the interfaces to make the light “do what we want,” so that we can redirect the fields to converge at images.

4.1 Fresnel Equations

A beam of light (implicitly a plane wave) in vacuum or in an isotropic medium propagates in the particular fixed direction specified by its Poynting vector until it encounters the interface with a different medium. The light causes the charges (electrons, atoms, or molecules) in the medium to oscillate and thus emit additional light waves that can travel in any direction (over the sphere of $4\pi$ steradians of solid angle). The oscillating particles vibrate at the frequency of the incident light and re-emit energy as light of that frequency (this is the mechanism of light “scattering”). If the emitted light is “out of phase” with the incident light (phase difference $\Delta \Phi \cong \pm \pi$ radians), then the two waves interfere destructively and the original beam is attenuated. If the attenuation is nearly complete, the incident light is said to be “absorbed.” Scattered light may interfere constructively with the incident light in certain directions, forming beams that have been reflected and/or transmitted. The constructive interference of the transmitted beam occurs at the angle that satisfies Snell’s law; while that after reflection occurs for $\theta_{\text{reflected}} = \theta_{\text{incident}}$. The mathematics are based on Maxwell’s equations for the three waves and the continuity conditions that must be satisfied at the boundary. The equations for these three electromagnetic waves are not difficult to derive, though the process is somewhat tedious. The equations determine the properties of light on either side of the interface and lead to the phenomena of:

1. equal angles of incidence and reflection;
2. Snell’s Law that relates the incident and refracted wave;
3. relative “strengths” and phases of the three light waves, and;
4. orientations of the electric fields of the three waves (the states of polarization of the three waves).

4.1.1 Definitions of Vectors

For simplicity, we consider only plane waves, so that all beams (incident, reflected, and transmitted) are specified by single wavevectors $\mathbf{k}_n$ that are valid at all points in a medium and that point in the
direction of propagation. The length of the wavevector in a medium with index \( n \) is:

\[
|\vec{k}_n| = \frac{2\pi}{\lambda_n} = \frac{2\pi}{\left( \frac{\lambda_0}{n} \right)} = \frac{2\pi n}{\lambda_0}
\]

where \( \lambda_0 \) is the wavelength in vacuum and \( \lambda_n = \frac{\lambda_0}{n} \) is the wavelength in the medium. The interface between the media is assumed to be the \( x - y \) plane located at \( z = 0 \). The incident wavevector \( \vec{k}_0 \), the reflected vector \( \vec{k}_r \), the transmitted (refracted) vector \( \vec{k}_t \) and the unit vector \( \hat{n} \) normal to the interface are shown:

All angles \( \theta_0 \), \( \theta_r \), and \( \theta_t \) are measured from the normal, so that \( \theta_0 \) and \( \theta_t \) are positive and \( \theta_r < 0 \) as drawn.

The incident and reflected beams are in the same medium (where \( n = n_1 \)) and so have the same wavelength and their \( \vec{k} \) vectors have the same magnitude:

\[
\lambda_1 = \frac{2\pi n_1}{|\vec{k}_0|} = \frac{2\pi n_1}{|\vec{k}_r|}
\]

\[
|\vec{k}_0| = |\vec{k}_r| = \frac{\omega_0}{v_1} = \frac{2\pi n_1}{\lambda_0}
\]

The wavelength of the transmitted (refracted) beam is different because of the different index of refraction:

\[
\lambda_2 = \frac{2\pi n_2}{|\vec{k}_t|}
\]

As drawn, the normal to the surface is specified by the unit vector perpendicular to the interface; in this case, it points in the direction of the positive \( z \)-axis:

\[
\hat{n} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}
\]

(n.b., we could have defined \( \hat{n} \) in the opposite direction, which would have changed the signs of the angles but would have had no effect on the physics).

The incident electric field is a sinusoidal oscillation that may be written in complex notation:

\[
\vec{E}_{\text{incident}} = \vec{E}_0 \exp \left[ +i (\vec{k}_0 \cdot \vec{r} - \omega_0 t) \right]
\]
where \( \mathbf{r} = [x, y, z] \) is the position vector of the location where the phase \( k_0 \cdot \mathbf{r} - \omega_0 t \) is measured; note that the phases measured at all positions in a plane perpendicular to the incident wavevector \( k_0 \) are identical because this is a plane wave.

The reflected and transmitted waves have the general forms:

\[
\begin{align*}
E_{\text{reflected}} &= E_r \exp \left[ +i \left( k_r \cdot \mathbf{r} - \omega_r t + \phi_r \right) \right] \\
E_{\text{transmitted}} &= E_t \exp \left[ +i \left( k_t \cdot \mathbf{r} - \omega_t t + \phi_t \right) \right]
\end{align*}
\]

where we have yet to demonstrate that \( \omega_r = \omega_t = \omega_0 \) and that \( |k_r| = |k_0| \). The constants \( \phi_r \) and \( \phi_t \) are the (perhaps different) initial phases of the reflected and transmitted waves, i.e., the phases measured at \( \mathbf{r} = \mathbf{0} \) and \( t = 0 \).

4.1.2 Snell’s Law for Reflection and Refraction of Waves

One boundary condition that must be satisfied is that the phases of all three waves must match at the interface (specified by \( z = 0 \)) at all times \( t \):

\[
(k_0 \cdot \mathbf{r} - \omega_0 t)_{|z=0} = (k_r \cdot \mathbf{r} - \omega_r t + \phi_r)_{|z=0} = (k_t \cdot \mathbf{r} - \omega_t t + \phi_t)_{|z=0}
\]

This equivalence at all times immediately implies that the temporal frequencies \( \omega \) of the three waves must be identical \( (\omega = \omega_0) \) because otherwise the phases would differ at different times. In words, this demonstrates that the temporal frequency is invariant with medium, which is equivalent to saying that the “color” of the light does not change if the light propagates into a different medium. By cancelling the temporal parts of the phases, we see that the spatial vectors must satisfy the conditions:

\[
(k_0 \cdot \mathbf{r})_{|z=0} = (k_r \cdot \mathbf{r} + \phi_r)_{|z=0} = (k_t \cdot \mathbf{r} + \phi_t)_{|z=0}
\]

Since the scalar products of the three wavevectors with the same position vector \( \mathbf{r} \) must be equal, then the three vectors \( k_0, k_r, \) and \( k_t \) must all lie in the same plane (call it the \( x-z \) plane, as shown in the drawing). The number of waves per unit length at any instant of time must be equal at the boundary for all three waves, as shown, which means that the \( x \)-components of the three wavevectors must be equal:

\[
(k_0)_x = (k_r)_x = (k_t)_x
\]

The \( x \)-components of the three wavevectors (for the incident, reflected, and transmitted refracted waves) must be equal at the interface to ensure that each produces the same number of waves per unit length along the interface, so that the three wavefronts “match” despite the difference in wavelengths in the two media.
From the definitions of the wavevectors, we can also see that:

\[
\begin{align*}
(k_0)_x &= |k_0| \cos \left( \frac{\pi}{2} - \theta_0 \right) = |k_0| \sin \theta_0 \\
(k_r)_x &= |k_r| \cos \left( \frac{\pi}{2} - \theta_r \right) = |k_r| \sin [-\theta_r]
\end{align*}
\]

where the factor of \(-1\) applied to the reflected angle arises from the fact that this angle is clockwise from the normal, and hence negative. The equality of the lengths of the incident and reflected wavevectors immediately demonstrates that:

\[
\begin{align*}
(k_0)_x &= (k_r)_x = |k_0| \sin \theta_0 = |k_r| \sin [-\theta_r] \\
\implies |k_0| \sin \theta_0 &= |k_r| \sin [-\theta_r] \\
\implies \sin \theta_0 &= \sin [-\theta_r] \\
\implies \theta_r &= -\theta_0
\end{align*}
\]

In words, the angle of reflection is equal to the negative of the angle of incidence. The sign of the angle often is ignored, leading to the statement that the angles of incidence and reflection are equal.

Now make the same observation for the transmitted wave; its angle is measured counter-clockwise from the normal and hence is positive:

\[
\begin{align*}
(k_0)_x &= |k_0| \sin \theta_0 = \frac{2\pi n_1}{\lambda_0} \sin \theta_0 \\
(k_t)_x &= |k_t| \cos \left( \frac{\pi}{2} - \theta_t \right) = |k_t| \sin \theta_t = \frac{2\pi n_2}{\lambda_0} \sin \theta_t
\end{align*}
\]

We equate these to derive the relationship of the angles of the incident and transmitted wavevectors:

\[
\frac{2\pi n_1}{\lambda_0} \sin \theta_0 = \frac{2\pi n_2}{\lambda_0} \sin \theta_t
\]

\[
\implies n_1 \sin \theta_0 = n_2 \sin \theta_t
\]

We recognize this to be (of course) Snell’s law for refraction.

The reflection law may be cast into the form of Snell’s refraction law by assuming that the index of refraction is negative for the reflected beam:

\[
n_1 \sin \theta_0 = -n_1 \sin \theta_r
\]

\[
\implies \sin \theta_r = -\sin \theta_0
\]

\[
\implies \theta_r = -\theta_0
\]

Note that these laws were derived without having to consider the vector nature of the electric and magnetic fields, but rather just the spatial frequencies of the waves at the boundaries. The next task is not quite this simple.....

### 4.1.3 Boundary Conditions for Electric and Magnetic Fields

We’ve determined the angles of the reflected and transmitted (refracted) plane waves in the form of Snell’s law(s). We also need to evaluate the “quantity” of light reflected and refracted due to the boundary. Since the geometries of the fields will depend on the directions of the electric field vectors, we will have to consider this aspect in the derivations. In short, this discussion will depend on the “orientation” of the electric field relative to the interface, which is what is called the “polarization” of the electric field (n.b., this is different from the “polarizability” of the medium). We will again have to match appropriate boundary conditions at the boundary, but these conditions apply to the vector components of the electric and magnetic fields on each side of the boundary. We use the same notation as before for amplitudes of the electric fields of the incident, reflected, and transmitted fields.
(refracted) waves. Faraday’s and Ampère’s laws (the Maxwell equations involving curl) for plane waves can be recast into forms that are more useful for the current task:

\[
\nabla \times \mathbf{E} \propto -\frac{\partial \mathbf{B}}{\partial t}
\]

\[
\nabla \times \mathbf{B} \propto +\frac{\partial \mathbf{E}}{\partial t}
\]

We need the constants of proportionality in this derivation. Recall that they depend on the system of units. In the MKS system, the laws have the form:

\[
\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}
\]

\[
\nabla \times \mathbf{B} = +\epsilon \mu \frac{\partial \mathbf{E}}{\partial t}
\]

where \(\epsilon\) and \(\mu\) are the permittivity and permeability of the medium, respectively and the phase velocity of light in the medium is:

\[v_0 = \sqrt{\frac{1}{\epsilon \mu}}\]

The incident field is assumed to be a plane wave of the form already used:

\[
\mathbf{E}_{\text{incident}}[x, y, z, t] = \mathbf{E}_0 \exp \left[+i (k_0 \cdot \mathbf{r} - \omega_0 t)\right]
\]

\[
= \mathbf{E}_0 \exp \left[+i \left(k_0 x + k_0 y + k_0 z - \omega_0 t\right)\right]
\]

\[
= \mathbf{E}_0 \exp \left[+i (k_{0x} x + k_{0y} y + k_{0z} z - \omega_0 t)\right]
\]

\[
= (\hat{x} E_{0x} + \hat{y} E_{0y} + \hat{z} E_{0z}) \exp \left[+i (k_{0x} x + k_{0y} y + k_{0z} z - \omega_0 t)\right]
\]

where the usual convention for naming the scalar components of the vector along the three axes has been used, and we know that \(\mathbf{E}_0 \perp k_0\). In our coordinate system, the incident wave vector lies in the \(x-z\) plane (the plane defined by \(k_0\) and \(\hat{n}\)), so that \(k_{0y} = 0\):

\[
\mathbf{E}_{\text{incident}}[x, y, z, t] = \mathbf{E}_0 \exp \left[+i (k_0 \cdot \mathbf{r} - \omega_0 t)\right]
\]

\[
= (\hat{x} E_{0x} + \hat{y} E_{0y} + \hat{z} E_{0z}) \exp \left[+i (k_{0x} x + k_{0y} y + k_{0z} z - \omega_0 t)\right]
\]

The boundary conditions that must be satisfied by the electric fields and by the magnetic fields at the boundary are perhaps not obvious. Consider the figure on the left:
components in this geometry) must cancel:

\[ \epsilon_1 \mathbf{E}_1 \cdot \hat{n} - \epsilon_2 \mathbf{E}_2 \cdot \hat{n} = 0 \]

\[ \implies \epsilon_1 E_{1z} = \epsilon_2 E_{2z} \]

\[ \mathbf{B}_1 \cdot \hat{n} - \mathbf{B}_2 \cdot \hat{n} = 0 \]

\[ \implies B_{1z} = B_{2z} \]

The flux of the electric field in a medium is the so-called “displacement” field \( \mathbf{D} = \epsilon \mathbf{E} \) and the flux of the magnetic field is the field \( \mathbf{B} \). Thus Gauss’ law determines that the normal components of \( \mathbf{D} \) and of \( \mathbf{B} \) are continuous across the boundary of the medium.

The figure on the right is a rectangular path (a “loop”) that also straddles the boundary. The unit vector \( \hat{t} \perp \hat{n} \) points along the interface surface. If the “height” of the loop \( dh \to 0 \), then the circulations of the electric and magnetic fields must cancel:

\[ \mathbf{E}_1 \cdot \hat{t} - \mathbf{E}_2 \cdot \hat{t} = 0 \]

\[ \implies E_{1x} = E_{2x} \]

\[ \frac{\mathbf{B}_1}{\mu_1} \cdot \hat{t} - \frac{\mathbf{B}_2}{\mu_2} \cdot \hat{t} = 0 \]

\[ \implies \frac{B_{1x}}{\mu_1} = \frac{B_{2x}}{\mu_2} \]

We now want to solve Maxwell’s equations for an incident plane wave, which will depend on the incident angle \( \theta_0 \) and on the vector direction of the electric field. It is convenient to evaluate these conditions in two cases of linearly polarized waves: (1) where the polarization is perpendicular to the plane of incidence defined by \( \hat{n} \) and \( \mathbf{k}_0 \) and (2) the polarization is parallel to the plane of incidence defined by \( \hat{n} \) and \( \mathbf{k}_0 \). The traditional jargon for the first state is “s polarization,” though Born and Wolf call it the “perpendicular” or “\( \perp \)” polarization, while other authors (including Jackson) call it the transverse electric (TE) polarization because the electric field vector is “perpendicular” to the interface. The second case is called “p” or “\( \parallel \)” or the transverse magnetic (TM) polarization. We shall follow Jackson’s naming convention of TE and TM. The two cases are depicted below:

The electric field perpendicular to the plane of incidence; this is the TRANSVERSE ELECTRIC field (TE, also called the “s” or “\( \perp \)” polarization).
4.1 FRESNEL EQUATIONS

The electric field is parallel to the plane of incidence; this is the TRANSVERSE MAGNETIC field (TM, also called the "p" or "||" polarization).

4.1.4 Transverse Electric (TE) Waves = "s" = “⊥“ Polarization

In the TE case in our geometry, the incident electric field $E_{\text{incident}}$ is oriented along the y direction and the wavevector $k_0$ has components in the x and z directions:

$$E_{\text{incident}}[x, y, z, t] = (\hat{x} \cdot 0 + \hat{y} \cdot E_0 + \hat{z} \cdot 0) \exp[i (k_0 x + k_0 z - \omega_0 t)]$$

$$= \hat{y} E_0 \exp[i (k_0 x + k_0 z - \omega_0 t)]$$

The magnetic field is derived from the relation:

$$B = \frac{n}{c} k_0 \times E$$

$$B_{\text{incident}}[x, y, z, t] = \left[ -\cos[\theta] \cdot n_1 \frac{|E_0|}{c} \right] \hat{x} + 0 \cdot \hat{y} + \left[ +\sin[\theta] \cdot n_1 \frac{|E_0|}{c} \right] \hat{z} \cdot \exp[i (k_0 x + k_0 z - \omega_0 t)]$$

The reflected fields are:

$$E_{\text{reflected}}[x, y, z, t] = \hat{y} \cdot |E_0| \exp[i (k_{rx} x + k_{rz} z - \omega_0 t)]$$

$$B_{\text{reflected}}[x, y, z, t] = \left[ +\cos[-\theta] \cdot n_1 \frac{|E_0|}{c} \right] \hat{x} + \left[ -\sin[-\theta] \cdot n_1 \frac{|E_0|}{c} \right] \hat{z} \cdot \exp[i (k_0 x + k_0 z - \omega_0 t)]$$

and the transmitted (refracted) fields are:

$$E_{\text{transmitted}}[x, y, z, t] = \hat{y} \cdot |E_t| \exp[i (k_{tx} x + k_{tz} z - \omega_0 t)]$$

$$B_{\text{transmitted}}[x, y, z, t] = \left[ -\cos[\theta_t] \cdot n_2 \frac{|E_t|}{c} \right] \hat{x} + \left[ \sin[\theta_t] \cdot n_2 \frac{|E_t|}{c} \right] \hat{z} \cdot \exp[i (k_0 x + k_0 z - \omega_0 t)]$$
The only components of the electric field at the interface are transverse, so the only boundary conditions to be satisfied are the tangential electric field:

\[ E_0 + E_r = E_t \implies 1 + \frac{E_r}{E_0} = \frac{E_t}{E_0} \]

This is typically expressed in terms of the reflection and transmission coefficients for the amplitude of the waves (not the power of the waves; these are the reflectance \( R \) and transmittance \( T \) of the interface, which will be considered very soon):

\[
\begin{align*}
r_{TE} &= \frac{E_r}{E_0} \\
t_{TE} &= \frac{E_t}{E_0}
\end{align*}
\]

where the subscripts denote the transverse electric polarization. The boundary condition for the normal magnetic field yields the expression:

\[
\frac{n_1}{c} \sin [\theta_0] (E_0 + E_r) = \frac{n_2}{c} \sin [\theta_t] E_t
\]

while that for the tangential magnetic field:

\[
\frac{n_1}{\mu_1 c} \cos [\theta_0] (E_0 - E_r) = \frac{n_2}{\mu_2 c} \cos [\theta_t] E_t
\]

These may be solved simultaneously for \( r \) and \( t \) to yield expressions in terms of the indices, permeabilities, and angles:

**Reflectance Coefficient for TE Waves**

\[
r_{TE} = \frac{E_r}{E_0} = \frac{n_1}{\mu_1} \cos [\theta_0] - \frac{n_2}{\mu_2} \cos [\theta_t]
\]

\[
r_{TE} = \frac{n_1 \cos [\theta_0] - n_2 \cos [\theta_t]}{n_1 \cos [\theta_0] + n_2 \cos [\theta_t]} \quad \text{if } \mu_1 = \mu_2 \quad \text{(usual case)}
\]

This equation for the amplitude reflectance (amplitude of the reflected wave) of the transverse electric polarization is one of the Fresnel Equations that calculates the effect of an interface on incident electromagnetic waves. The corresponding Fresnel equation for the amplitude transmittance is:

**Transmission Coefficient for TE Waves**

\[
t_{TE} = \frac{E_t}{E_0} = \frac{+2 \frac{n_1}{\mu_1} \cos [\theta_0]}{n_1 \cos [\theta_0] + n_2 \cos [\theta_t]}
\]

\[
t_{TE} = \frac{+2n_1 \cos [\theta_0]}{n_1 \cos [\theta_0] + n_2 \cos [\theta_t]} \quad \text{if } \mu_1 = \mu_2
\]

Again, these are the amplitude coefficients; the reflectance and transmittance of light at the surface relate the energies or powers. These measure the ratios of the reflected or transmitted power to the incident power. The power is proportional to the product of the magnitude of the Poynting vector and the area of the beam. The areas of the beams before and after reflection are identical, which means that the reflectance is just the ratio of the magnitudes of the Poynting vectors. This reduces to the square of the amplitude reflection coefficient:

\[
R = r^2
\]
which reduces to this expression for the TE case:

$$R_{TE} = \left( \frac{n_1 \cos \theta_0 - n_2 \cos \theta_t}{n_1 \cos \theta_0 + n_2 \cos \theta_t} \right)^2$$

The transmittance $T$ is a bit more complicated to compute, because the refraction at the interface changes the “width” of the beam (and thus its the cross-sectional area in one direction). The example in the figure shows a case with $n_1 > n_2$, where the width of the transmitted (“refracted”) beam along the x-axis is larger, and thus the area of the transmitted beam is larger, in the medium with the larger index:

Demonstration that the cross-sectional area of the beam is changed by refraction at the interface between two media with different refractive indices. The area is larger in the medium with the larger index. This difference must be accounted for in the calculation of the power transmission $T$ across the interface.

The magnitude of the Poynting vector is proportional to the product of the index of refraction and the squared magnitude of the electric field:

$$|\mathbf{s}_1| \propto n_1 |E_0|^2$$
$$|\mathbf{s}_2| \propto n_2 |E_t|^2$$

The ratio of the transmitted to incident power is:

$$T = \frac{|\mathbf{s}_2| \cdot A_2}{|\mathbf{s}_1| \cdot A_1} = \frac{n_2 |E_t|^2 \cdot A_2}{n_1 |E_0|^2 \cdot A_1} = \frac{n_2}{n_1} \cdot t^2 \cdot \frac{A_2}{A_1}$$

The area of the transmitted beam changes in proportion to the dimension along the x-axis in this case, which allows us to see that:

$$\frac{A_2}{A_1} = \frac{w_2}{w_1} = \frac{\sin \left[ \frac{\pi}{2} - \theta_t \right]}{\sin \left[ \frac{\pi}{2} - \theta_0 \right]} = \frac{\cos \theta_t}{\cos \theta_0}$$

which is larger than unity if $|\theta_t| < |\theta_0|$, as is the case if $n_2 > n_1$. This leads to the final expression
for the transmission at the interface:

\[
T = \left( \frac{n_2 \cos \theta_1}{n_1 \cos \theta_0} \right) \cdot t^2
\]

Snell’s law is used to derive the relationship between the incident and transmitted angles:

\[
n_1 \sin \theta_0 = n_2 \sin \theta_1
\]

\[
\implies \sin \theta_t = \frac{n_1}{n_2} \sin \theta_0
\]

\[
\implies \cos \theta_t = \sqrt{1 - \sin^2 \theta_t} = \sqrt{1 - \left( \frac{n_1}{n_2} \sin \theta_0 \right)^2} = \frac{1}{n_2} \sqrt{n_2^2 - n_1^2 \sin^2 \theta_0}
\]

Thus we can write down the transmittance \(T\) in terms of the refractive indices and the incident angle:

\[
T = \left( \frac{n_2 \cos \theta_1}{n_1 \cos \theta_0} \right) \cdot t^2 = \left( \frac{\sqrt{n_2^2 - n_1^2 \sin^2 \theta_0}}{n_1 \cos \theta_0} \right) \cdot t^2
\]

where the appropriate expression for \(t\) for the polarization is inserted. For the TE case, the transmission is:

\[
TE = \left( \frac{\sqrt{n_2^2 - n_1^2 \sin^2 \theta_0}}{n_1 \cos \theta_0} \right) \cdot \left( \frac{1}{n_1 \cos \theta_0 + n_2 \cos \theta_t} \right)^2
\]

These will be plotted for some specific cases after we evaluate the coefficients for TM waves.

### 4.1.5 Transverse Magnetic (TM) Waves = “p” = “||” polarization

In the TM case in our geometry, the electric field is in the x-z plane and the wavevector has components in the x and z directions:

\[
E_{\text{incident}} [x, y, z, t] = (\hat{x} \cdot |E_0| \cos \theta_0 + \hat{z} \cdot |E_0| \sin \theta_0) \exp [+i \left( k_{ox}x + k_{oz}z - \omega_0 t \right)]
\]

\[
E_{\text{incident}} [x, y, z, t] = (\hat{x} \cdot |E_0| \cos \theta_0 - \hat{z} \cdot |E_0| \sin \theta_0) \exp [+i \left( k_{ox}x + k_{oz}z - \omega_0 t \right)]
\]

The magnetic field is in the y-direction:

\[
B_{\text{incident}} [x, y, z, t] = \left( \frac{n_1 |E_0|}{c} \hat{y} \right) \exp [+i \left( k_{ox}x + k_{oz}z - \omega_0 t \right)]
\]

The reflected fields are:

\[
E_{\text{reflected}} [x, y, z, t] = (\hat{x} \cdot -|E_0| \cos \theta_0 - \hat{z} \cdot |E_0| \sin \theta_0) \exp [+i \left( k_{ox}x + k_{oz}z - \omega_0 t \right)]
\]

\[
E_{\text{reflected}} [x, y, z, t] = \left( \frac{n_1 |E_0|}{c} \hat{y} \right) \exp [+i \left( k_{ox}x + k_{oz}z - \omega_0 t \right)]
\]
4.1 FRESNEL EQUATIONS

and the transmitted (refracted) fields are:

\[
\mathbf{E}_{\text{transmitted}}[x, y, z, t] = (\hat{\mathbf{x}} \cdot \mathbf{E}_0) \cos [\phi t] - \hat{\mathbf{y}} \cdot \mathbf{E}_0 \sin [\phi t]) \exp \left[ +i \left( k_0 x + k_0 z - \omega_0 t \right) \right]
\]

\[
\mathbf{B}_{\text{transmitted}}[x, y, z, t] = \left( \frac{n_2 \mathbf{E}_0}{c} \hat{\mathbf{y}} \right) \exp \left[ +i \left( k_0 x + k_0 z - \omega_0 t \right) \right]
\]

In the case, the boundary condition on the normal component of \( \mathbf{B} \) is trivial, but the other components are:

\[
\mu_1 \sin [\phi] (E_0 + E_r) = \mu_2 \sin [\theta] E_t
\]

\[
\cos [\phi] (E_0 - E_r) = \cos [\theta] E_t
\]

\[
\frac{n_1}{\mu_1 c} (E_0 + E_r) = \frac{n_2}{\mu_2 c} E_t
\]

These are solved for the reflection and transmission coefficients:

**Transverse Magnetic Waves**

\[
r_{TM} = \frac{\pm \frac{n_2}{\mu_2} \cos [\phi] - \frac{n_1}{\mu_1} \cos [\theta]}{\pm \frac{n_2}{\mu_2} \cos [\phi] + \frac{n_1}{\mu_1} \cos [\theta]}
\]

which simplifies if the permeabilities \( \mu_n \) are equal (as they usually are):

\[
r_{TM} = \frac{\pm n_2 \cos [\phi] - n_1 \cos [\theta]}{\pm n_2 \cos [\phi] + n_1 \cos [\theta]} \quad \text{if} \quad \mu_1 = \mu_2
\]

The corresponding reflectance is:

\[
R_{TM} = \left( \frac{\pm n_2 \cos [\phi] - n_1 \cos [\theta]}{\pm n_2 \cos [\phi] + n_1 \cos [\theta]} \right)^2
\]

The amplitude transmission coefficient evaluates to:

\[
t_{TM} = \frac{2 \frac{n_1}{\mu_1} \cos [\phi]}{\pm \frac{n_2}{\mu_2} \cos [\phi] + \frac{n_1}{\mu_1} \cos [\theta]}
\]

again, if the permeabilities are equal, this simplifies to:

\[
t_{TM} = \frac{2 n_1 \cos [\phi]}{\pm n_2 \cos [\phi] + n_1 \cos [\theta]} \quad \text{if} \quad \mu_1 = \mu_2
\]

The corresponding transmittance function is:

\[
T_{TM} = \left( \frac{n_2^2 - n_1^2 \sin^2 [\theta]}{n_1 \cos [\theta]} \right) \cdot \left( \frac{2 n_1 \cos [\theta]}{\pm n_2 \cos [\phi] + n_1 \cos [\theta]} \right)^2
\]

4.1.6 Comparison of Coefficients for TE and TM Waves

The reflectance coefficients for the two cases of TE and TM waves are:
CHAPTER 4 ELECTROMAGNETIC WAVES AT AN INTERFACE

\( r_{TE} = \frac{n_1 \cos \theta_0 - n_2 \cos \theta_t}{n_1 \cos \theta_0 + n_2 \cos \theta_t} \)

\( r_{TM} = \frac{+n_2 \cos \theta_0 - n_1 \cos \theta_t}{+n_2 \cos \theta_0 + n_1 \cos \theta_t} \)

where the angles are also determined by Snell’s law:

\( n_1 \sin \theta_0 = n_2 \sin \theta_t \)

\[ \Rightarrow \cos \theta_t = \sqrt{1 - \left( \frac{n_1}{n_2} \sin \theta_0 \right)^2} \]

Note that angles and the indices for the TE case are “in” the same media, i.e., the index \( n_1 \) multiplies the cosine of \( \theta_0 \), which is in the same medium, and the same holds for \( n_2 \) and \( \theta_t \). However, the opposite situation applies to the TM case: \( n_1 \) is applied to \( \cos \theta_t \) and \( n_2 \) to \( \cos \theta_0 \). These same observations also apply to the corresponding transmission coefficients:

\( t_{TE} = \frac{+2n_1 \cos \theta_0}{n_1 \cos \theta_0 + n_2 \cos \theta_t} = \frac{+2 \cos \theta_0}{\cos \theta_0 + \frac{n_2}{n_1} \cos \theta_t} \)

\( t_{TM} = \frac{+2n_1 \cos \theta_0}{+n_2 \cos \theta_0 + n_1 \cos \theta_t} = \frac{+2 \cos \theta_0}{\frac{n_2}{n_1} \cos \theta_0 + \cos \theta_t} \)

**Normal Incidence** \((\theta_0 = 0)\)

In the case of normal incidence, where \( \theta_0 = \theta_r = \theta_t = 0 \), then the TE and TM equations evaluate to:

\( r_{TE}|_{\theta_0=0} = \frac{n_1 - n_2}{n_1 + n_2} \)

\( r_{TM}|_{\theta_0=0} = \frac{+n_2 - n_1}{+n_2 + n_1} = -\left( r_{TE}|_{\theta_0=0} \right) \)

\( t_{TE}|_{\theta_0=0} = \frac{+2n_1}{n_1 + n_2} \)

\( t_{TM}|_{\theta_0=0} = \frac{+2n_1}{n_1 + n_2} = t_{TE}|_{\theta_0=0} \)

Note that the reflectance coefficients are not identical, even though the two polarizations are indistinguishable at normal incidence. Since the areas of the incident and transmitted waves are identical, there is no area factor in the amplitude transmittance. The resulting formulas for the observable reflectance and transmittance are identical:

**normal incidence** \((\theta_0 = 0)\)

\( R_{TE}(\theta_0 = 0) = R_{TM}(\theta_0 = 0) \equiv R = \left( \frac{n_1 - n_2}{n_1 + n_2} \right)^2 \)

\( T = \frac{4n_1 n_2}{(n_1 + n_2)^2} \)

**“Rare-to-Dense” Reflection at Normal Incidence** If the input medium has a smaller refractive index \( n \) (a rarer medium) than the second (denser) medium (e.g., “air to glass” with \( n_1 = 1.0 < n_2 = 1.5 \), then the Fresnel coefficients are:
4.1 FRESNEL EQUATIONS

\[
\begin{align*}
r_{TE} &= \frac{1.0 - 1.5}{1.0 + 1.5} = -0.2 = 0.2e^{+i\pi} \\
r_{TM} &= \frac{1.5 - 1.0}{1.5 + 1.0} = +0.2 \\
t_{TE} &= t_{TM} = \frac{2 \cdot 1.0}{1.0 + 1.5} = +0.8 \\
\end{align*}
\]

for "rare-to-dense" reflection with \(n_1 = 1.0\) and \(n_2 = 1.5\)

Note the value of \(r_{TE}\); in words, the phase of the reflected light is changed by \(\pi\) radians = 180° if reflected at a “rare-to-dense” interface such as the usual air-to-glass case. The reflectivity \(R\) and the transmittance \(T\) are:

\[
\begin{align*}
R_{TE} &= R_{TM} = \left(\frac{1 - 1.5}{1 + 1.5}\right)^2 = 0.04 \\
T_{TE} &= T_{TM} = \frac{4 \cdot 1.1.5}{(1 + 1.5)^2} = 0.96 \\
\end{align*}
\]

which means that:

\[R + T = 1\]

“Dense-to-Rare” Reflection at Normal Incidence If the input medium is “denser” \((n_1 = 1.5 > n_2 = 1.0)\), then the amplitude reflection coefficients at normal incidence are:

\[
\begin{align*}
r_{TE} &= \frac{1.5 - 1.0}{1.5 + 1.0} = +0.2 \\
r_{TM} &= \frac{1.0 - 1.5}{1.0 + 1.5} = -0.2 = 0.2e^{+i\pi} \\
\end{align*}
\]

The reflectivity is the same as in the “rare-to-dense” reflection:

\[R = (\pm0.2)^2 = 0.04\]

The amplitude transmission coefficients at normal incidence in a “dense-to-rare” reflection are identical:

\[
t_{TE} = t_{TM} = \frac{2 \cdot 1.5}{1.0 + 1.5} = +1.2 > 1.0
\]

The fact that these values are larger than unity may seem strange, at least until we recall that the transmittance requires the additional geometrical factor:

\[
T = \left(\frac{n_2 \cos \theta_t}{n_1 \cos \theta_0}\right) \cdot t^2 \\
= \left(\frac{1}{1.5}\right) \cdot (+1.2)^2 = 0.96
\]

Again, we see that energy is conserved:

\[R + T = 1\]

4.1.7 Angular Dependence of Amplitude Coefficients at “Rare-to-Dense” Interface

The graphs of the Fresnel coefficients using the equations just derived for the cases of the “rare-to-dense” interface \((n_1 = 1 < n_2 = 1.5)\) are plotted vs. incident angle measured in degrees from 0°
normal incidence) to 90° (grazing incidence). Note that the amplitude reflectance coefficient for the TM state (“||” or p polarization) goes to zero at a specific angle $\theta_B \approx 60^\circ$ (Brewster’s Angle), which means that $R_{TM} = 0$ and $T_{TM} = 1$ there. This phenomenon was discovered in 1815 by David Brewster.

4.1.8 Reflectance and Transmittance at “Rare-to-Dense” Interface

The reflectance and transmittance the two polarizations with $n_1 = 1.0$ and $n_2 = 1.5$ as functions of the incident angle $\theta_0$ show that the reflectance of the TM wave at Brewster’s angle is zero.
4.1 FRESNEL EQUATIONS

Reflectance and transmittance for $n_1 = 1.0$ and $n_2 = 1.5$ for TE and TM waves. Note that $R_{TM} = 0$ and $T_{TM} = 1$ at “Brewster’s angle.”

Brewster’s Angle for Complete Polarization of Reflected Wave

In a “rare-to-dense” reflection at Brewster’s Angle $\theta_0 = \theta_B$, we see that $r_{TM} = 0$ and therefore the reflected amplitude of this wave is zero. In other words, the only light reflected at this angle must have the TE polarization only. This is Brewster’s angle where the electrons driven in the plane of the incidence will not emit radiation at the angle required by the law of reflection; this is sometimes called the angle of complete polarization. The incident angle $\theta_B$ may be evaluated by setting $r_{TM} = 0$:

\[
\begin{align*}
\frac{+n_2 \cos [\theta_B] - n_1 \cos [\theta_t]}{+n_2 \cos [\theta_B] + n_1 \cos [\theta_t]} &= 0 \\
\Rightarrow +n_2 \cos [\theta_B] &= n_1 \cos [\theta_t] \\
\Rightarrow \cos^2 [\theta_t] &= \left(\frac{n_2}{n_1}\right)^2 \cos^2 [\theta_B]
\end{align*}
\]

We also have an expression for $\sin^2 [\theta_t]$ from Snell’s law:

\[n_1 \sin [\theta_B] = n_2 \sin [\theta_t] \Rightarrow \sin^2 [\theta_t] = \left(\frac{n_1}{n_2}\right)^2 \sin^2 [\theta_B]\]

Square and add to the expression for $\cos^2 [\theta_t]$:

\[\cos^2 [\theta_t] + \sin^2 [\theta_t] = 1 = \left(\frac{n_2}{n_1}\right)^2 \cos^2 [\theta_B] + \left(\frac{n_1}{n_2}\right)^2 \sin^2 [\theta_B]\]
But we also know that $\cos^2[\theta_B] + \sin^2[\theta_B] = 1$, so the two expressions may be equated:

$$\cos^2[\theta_B] + \sin^2[\theta_B] = \left(\frac{n_2}{n_1}\right)^2 \cos^2[\theta_B] + \left(\frac{n_1}{n_2}\right)^2 \sin^2[\theta_B]$$

$$\Rightarrow \frac{n_2^2 - n_1^2}{n_1^2} \cos^2[\theta_B] + \frac{n_1^2 - n_2^2}{n_2^2} \sin^2[\theta_B] = 0$$

$$\Rightarrow \frac{\sin^2[\theta_B]}{\cos^2[\theta_B]} = \tan^2[\theta_B] = \frac{n_2^2}{n_1^2}$$

So the expression that is satisfied by Brewster’s angle at a “rare-to-dense” reflection is:

$$\theta_B = \tan^{-1}\left(\frac{n_2}{n_1}\right)$$

If $n_1 = 1$ (air) and $n_2 = 1.5$ (glass), then $\theta_B \cong 56.3^\circ$. For incident angles larger than about 56°, the reflected light is plane polarized parallel to the plane of incidence. We can also find $\theta_t$ by applying Snell’s law:

$$n_1 \sin[\theta_B] = n_2 \sin[\theta_t]$$

$$1 \cdot \sin\left[\tan^{-1}\left(\frac{1.5}{1}\right)\right] = 1.5 \cdot \sin[\theta_t]$$

$$\theta_t = \sin^{-1}\left[\frac{1}{1.5} \sin\left[\tan^{-1}\left(\frac{1.5}{1}\right)\right]\right] \approx 33.7^\circ$$

Note that

$$\theta_B + \theta_t = 56.3^\circ + 33.7^\circ = 90^\circ = \frac{\pi}{2} \text{ radians}$$

If the dense medium is water ($n_2 = 1.33$), then $\theta_B \cong 52.4^\circ$ and $\theta_t = 37.6^\circ$; again, $\theta_B + \theta_t = \frac{\pi}{2}$ radians. This means that light incident at Brewster’s angle $\theta_B$ is refracted at $\theta_t$ such that:

$$\theta_t + \theta_B = \frac{\pi}{2} \Rightarrow \sin[\theta_t] = \sin\left[\frac{\pi}{2} - \theta_B\right] = \cos[\theta_B]$$

Polarization of reflected light at Brewster’s angle. The incident beam at $\theta_0 = \theta_B$ is unpolarized. The reflectance coefficient for light polarized in the plane (TM waves) is 0, and the sum of the incident and refracted angle is $90^\circ = \frac{\pi}{2}$. Thus $\theta_B + \theta_t = \frac{\pi}{2} \Rightarrow \theta_t = \frac{\pi}{2} - \theta_B$.

The reflection at Brewster’s angle provides a handy means to determine the polarization axis of a linear polarizer – just look through a linear polarizer at light reflected at a shallow angle relative
to the surface (e.g., a waxed floor). Note that the transmitted (refracted) light contains both polarizations, though not in equal amounts.

### 4.1.9 Reflection and Transmittance at “Dense-to-Rare” Interface, Critical Angle

At a “dense-to-rare” interface \((n_1 > n_2)\), the reflectance of the TM wave \((s\) polarization) is:

\[
r = \frac{-n_2 \cos[\theta_0] + n_1 \cos[\theta_t]}{n_2 \cos[\theta_0] + n_1 \cos[\theta_t]}
\]

The numerator evaluates to zero for a particular incident angle \(\theta_C\) that satisfies:

\[
n_2 \cos[\theta_C] = n_1 \cos[\theta_t]
\]

\[
\frac{n_1}{n_2} = \frac{\cos[\theta_C]}{\cos[\theta_t]}
\]

This corresponds to the situation where Snell’s law requires that:

\[
\sin[\theta_t] = \frac{n_1}{n_2} \sin[\theta_C] \implies \sin[\theta_C] = \frac{n_2}{n_1}
\]

If the incident angle exceeds this critical angle, then the amplitude reflectance coefficients \(r_{TE}\) and \(r_{TM}\) are both unity, and thus so are the reflectances \(R_{TE}\) and \(R_{TM}\). This means that light incident at \(\theta_0 \geq \theta_C\) is totally reflected at a “dense-to-rare” interface. This phenomenon is the source of total internal reflectance (TIR); the reflection is “internal” because it within the dense medium, e.g., glass. TIR is why optical fibers are useful in communications. If \(n_1 = 1.5\) and \(n_2 = 1.0\), then

\[
\sin[\theta_C] = \frac{2}{3} \implies \theta_0 \approx 0.73 \text{ radians} \approx 41.8^\circ \approx \theta_C
\]

Compare this to Brewster’s angle in the same case satisfies:

\[
\theta_B = \tan^{-1}\left[\frac{n_2}{n_1}\right] = \tan^{-1}\left[\frac{2}{3}\right] \implies \theta_B \approx 0.59 \text{ radians} \approx 33.7^\circ < \theta_C
\]

\[\text{Amplitude reflectance coefficients for TE and TM waves at a “dense-to-rare” interface with}\]
$n_1 = 1.5$ (glass) and $n_2 = 1.0$ (air). Both polarizations rise to $r = +1.0$ at the “critical angle” $\theta_c$, for which $\theta_i = 90^\circ = \frac{\pi}{2}$. Also noted is Brewster’s angle, where $r_{TM} = 0$. The coefficients for $\theta_0 > \theta_c$ may be interpreted as being complex-valued.

4.1.10 Practical Applications for Fresnel’s Equations

The 4% reflectivity at normal incidence for one surface of glass is the reason why windows resemble mirrors at night for persons in brightly lit rooms. The windows at the ends of the gas tube in a He:Ne laser are often oriented at Brewster’s angle to eliminate reflective losses; the emitted laser light is linearly polarized. The basic principle of optical fiber transmission is total internal reflection to propagate the beam. Hollow optical fibers use high-incidence-angle near-unity reflections.
Chapter 5

Optical Interference and Coherence

We now return to the more accurate model of light as a wave with the goal of understanding the performance limitations of imaging systems. We first consider “interference,” the result of adding waves from a small number of sources which may “cancel out” or reinforce, depending on the amplitudes and relative phase angles. This study leads to the concept of “coherence,” which is a measure of the statistical behavior of the optical phase over time, space, or both. Finally, we consider the subject of diffraction, which is really just an extension of the concept of interference to large numbers of sources. We shall see that the effects of diffraction lead to the ultimate limit on the performance of an optical imaging system.

The basic concept of interference is the difference in the phase of light waves that are superposed at one point; the two waves may have been generated from two locations on the same wavefront or they may have been derived from the same point on the wavefront and traveled two different paths to recombination. In either event, the significant aspect is the optical phase difference between the recombined waves. The importance of the optical phase difference is illustrated by a simple addition of two oscillations and then will be generalized to the sum of traveling waves.

5.1 Addition of Oscillations

The simple mathematical sum of two sinusoids with the same amplitude and different frequencies and phases may be evaluated via the identity for the sum of two cosines:

\[
\cos [\varphi_1] + \cos [\varphi_2] = 2 \cos \frac{\varphi_1 + \varphi_2}{2} \cos \frac{\varphi_1 - \varphi_2}{2}
\]

\[
\equiv 2 \cos [\varphi_{\text{avg}}] \cdot \cos [\varphi_{\text{mod}}]
\]

This relationship may be derived easily using complex analysis. This result may then be generalized to the sum of two sinusoidal oscillations with the same amplitude \(A\) but different angular temporal frequencies \(\omega_1\) and \(\omega_2\):

\[
y_1 [t] = A \cos [\omega_1 t + \phi_1]
\]

\[
y_2 [t] = A \cos [\omega_2 t + \phi_2]
\]

\[
y_1 [t] + y_2 [t] = 2A \cos \left[ \frac{\omega_1 + \omega_2}{2} t + \frac{\phi_1 + \phi_2}{2} \right] \cos \left[ \frac{\omega_1 - \omega_2}{2} t - \frac{\phi_1 - \phi_2}{2} \right]
\]

\[
\equiv 2A \cos [\omega_{\text{avg}} t + \phi_{\text{avg}}] \cdot \cos [\omega_{\text{mod}} t + \phi_{\text{mod}}]
\]

where \(\omega_{\text{avg}} \equiv \frac{\omega_1 + \omega_2}{2}\), \(\omega_{\text{mod}} \equiv \frac{\omega_1 - \omega_2}{2}\), \(\phi_{\text{avg}} \equiv \frac{\phi_1 + \phi_2}{2}\), and \(\phi_{\text{mod}} \equiv \frac{\phi_1 - \phi_2}{2}\). In words, the sum of two oscillations with different frequencies is identical to the product of two oscillations: one at the slower varying modulation \(\omega_{\text{mod}}\) and the other is the more rapidly oscillating average frequency.
The latter is called the carrier frequency in some applications. A perhaps familiar example of the modulation results from the excitation of two mistuned piano strings, where a low-frequency oscillation (the beat) is heard. As one string is tuned to the other, the frequency of the beat decreases until reaching zero when the string frequencies match. Acoustic beats may be thought of as interference of the summed oscillations in time.

**Sidebar: Addition of Different-Amplitude Oscillations**

As an aside, the corresponding result for waves with different amplitudes may be evaluated by decomposing the larger term into two pieces, of which one has the same amplitude:

\[
A_1 \cos [\omega_1 t + \phi_1] + A_2 \cos [\omega_2 t + \phi_2] \\
= (A_1 - A_2) \cos [\omega_1 t + \phi_1] + A_2 (\cos [\omega_1 t + \phi_1] + \cos [\omega_2 t + \phi_2]) \\
= (A_1 - A_2) \cos [\omega_1 t + \phi_1] + 2A_2 \cos [\omega_{\text{avg}} t + \phi_{\text{avg}}] \cdot \cos [\omega_{\text{mod}} t + \phi_{\text{mod}}]
\]

**Sidebar: Multiplication of Oscillations**

We have just seen that the sum of two sinusoidal oscillations with the same amplitude may be written as the product of oscillations with the average and modulation frequencies. In passing, we note the (also useful) complementary result that the product of two oscillations with the sum and difference frequencies:

\[
\cos [\omega_1 t + \phi_1] \cdot \cos [\omega_2 t + \phi_2] \\
= \frac{1}{2} \cos [(\omega_1 + \omega_2) t + (\phi_1 + \phi_2)] + \frac{1}{2} \cos [(\omega_1 - \omega_2) t + (\phi_1 - \phi_2)]
\]

In words, the product of two oscillations is equivalent to the sum of two oscillations: one at the sum frequency (with the sum of the initial phases) and one at the difference frequency. This is the principle of signal mixing or modulation, as in AM radio where the transmitted radio-frequency signal (RF) is multiplied by an intermediate frequency (IF) generated within the receiver to produce the sum of high-frequency and low-frequency signals. The sum-frequency signal is stripped away by a lowpass filter, leaving the desired audio signal.

“Heterodyne” detection by multiplying two oscillations. The difference frequency is passed by the lowpass filter.

### 5.1.1 Addition of 1-D Traveling Waves

The expression for the summation of two oscillations may be extended to 1-D traveling waves with the same amplitude $A$:

\[
f_1 [z, t] = A \cos [(k_1 z - \omega_1 t) + \phi_1] \\
f_2 [z, t] = A \cos [(k_2 z - \omega_2 t) + \phi_2] \\
f_1 [z, t] + f_2 [z, t] = 2A \cdot \cos [(k_{\text{mod}} z - \omega_{\text{mod}} t) + \phi_{\text{mod}}] \cdot \cos [(k_{\text{avg}} z - \omega_{\text{avg}} t) + \phi_{\text{avg}}]
\]
5.1 ADDITION OF OSCILLATIONS

\[ k_{\text{mod}} = \frac{k_1 - k_2}{2} \]
\[ v_{\text{mod}} = \frac{\omega_{\text{mod}}}{k_{\text{mod}}} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\Delta \omega}{\Delta k} \]
\[ \phi_{\text{mod}} = \frac{\phi_1 - \phi_2}{2} \]

\[ k_{\text{avg}} = \frac{k_1 + k_2}{2} \]
\[ v_{\text{avg}} = \frac{\omega_{\text{avg}}}{k_{\text{avg}}} = \frac{\omega_1 + \omega_2}{k_1 + k_2} \]
\[ \phi_{\text{avg}} = \frac{\phi_1 + \phi_2}{2} \]

In words, the superposition of two traveling waves with different temporal frequencies (and thus different wavelengths) generates the product of two component traveling waves, one oscillating more slowly in both time and space, i.e., a “traveling modulation.” Note that both the average and modulation waves move along the \(z\)-axis. In this case, \(k_1, k_2, \omega_1,\) and \(\omega_2\) are all positive, and so \(k_{\text{avg}}\) and \(\omega_{\text{avg}}\) must be also. However, the modulation wavenumber and frequency may be negative. In fact, the algebraic sign of \(k_{\text{mod}}\) may be negative even if \(\omega_{\text{mod}}\) is positive (or vice versa). In either of these cases, the modulation wave moves in the opposite direction to the average wave.

Example: 1-D Waves Traveling in Opposite Directions \((k_2 = -k_1)\)

Note that if the two 1-D waves traveling in the same direction along the \(z\)-axis have the same frequency \(\omega\), they must have the same wavelength \(\lambda_0\) and the same wavenumber \(k = \frac{2\pi}{\lambda_0}\). The modulation terms \(k_{\text{mod}}\) and \(\omega_{\text{mod}}\) must be zero, and the summation wave exhibits no modulation. Recall also such waves traveling in opposite directions generate a waveform that “moves” but does not “travel:”

\[
\begin{align*}
  f_1 [z, t] &= A \cos [(k_1 z - \omega_1 t) + \phi_1] \\
  f_2 [z, t] &= A \cos [(-k_1 z - \omega_1 t) + \phi_2] = A \cos [(k_1 z + \omega_1 t) - \phi_2] \\
  f_1 [z, t] + f_2 [z, t] &= A \cos [(k_1 z - \omega_1 t) + \phi_1] + A \cos [(k_1 z + \omega_1 t) - \phi_2] \\
  &= 2A \cos \left[ \left( \frac{k_1 + k_1}{2} \right) z - \left( \frac{\omega_1 - \omega_1}{2} \right) t + \left( \frac{\phi_1 - \phi_2}{2} \right) \right] \\
  &\quad \times \cos \left[ \left( \frac{k_1 - k_1}{2} \right) z - \left( \frac{\omega_1 + \omega_1}{2} \right) t + \left( \frac{\phi_1 + \phi_2}{2} \right) \right] \\
  &= 2A \cos \left[ k_1 z + \frac{\Delta \phi}{2} \right] \cdot \cos \left[ -\omega_1 t + \frac{\phi_1 + \phi_2}{2} \right] \\
  &= 2A \cos [k_1 z + \phi_{\text{mod}}] \cdot \cos [\omega_1 t + \phi_{\text{avg}}] \\
  &= 2A \cos [k_1 z + \phi_{\text{mod}}] \cdot \cos [\omega_1 t - \phi_{\text{avg}}] 
\end{align*}
\]

where the symmetry of \(\cos[\theta]\) was used in the last step. The result is the product of two terms: one each that oscillates in space and time. The spatial variation is “fixed,” but the amplitude oscillates in time at \(\omega_1\) radians per second. This is commonly called a “standing wave.”
The behavior of a standing wave over time: the spatial variation does not move but its amplitude does oscillate at an angular frequency of $\omega_0$ radians/second. Note that the amplitude at a value of $z$ is negative half of the time.

Though not apparent at this point, the goal of optical interference is to produce a standing wave that may be recorded and measured.

**Traveling Waves with “Similar” Frequencies $\omega_1 \approx \omega_2$**

If $\omega_1 \approx \omega_2$ (which means that $k_1 = |k_1| \approx k_2$), then the propagation velocities of the two waves are approximately equal

\[
\begin{align*}
\omega_{\text{avg}} &= \frac{\omega_1 + \omega_2}{2} \approx \omega_1 \\
k_{\text{avg}} &= k_1 \approx k_2 \\
\omega_{\text{mod}} &= \frac{\omega_1 - \omega_2}{2} \approx 0 \\
k_{\text{mod}} &= \frac{k_1 - k_2}{2} \approx 0
\end{align*}
\]

The propagation velocity of the “average wave” is:

\[
v_{\text{avg}} = \frac{\omega_{\text{avg}}}{k_{\text{avg}}} \approx \frac{\omega_1}{k_1} = v_1 \approx v_2
\]

The velocity of the “modulation wave” is:

\[
v_{\text{mod}} = \frac{\omega_{\text{mod}}}{k_{\text{mod}}} = \frac{\Delta \omega}{\Delta k} = \frac{\Delta \omega}{\Delta k} \rightarrow \frac{d\omega}{dk} \text{ in the limit } k_2 \rightarrow k_1
\]

Another common name for $v_{\text{mod}}$ is the **group velocity**, which is a measure of the velocity of propagation of information carried by the modulation.

The “game” of interference is to combine two (or more) sinusoidal traveling waves that differ in some attribute: different propagation directions, different phases, different path lengths, from different locations on a source, etc. The two waves may have been extracted from different points on the same source wavefront (“division of wavefront” interference) or obtained by dividing the light amplitude from the same source wavefront with some type of beamsplitter (“division of amplitude” interference). The next chapter generalizes the concept of interference to include a “large” number of waves (usually an infinite number). The resulting concept of “diffraction” is very important in
5.2 Division-of-Wavefront Interference

Figure 5.1: Two classes of interference used to measure statistical variation in the phase: (a) light from two locations on the same wavefront is added in division-of-wavefront interference, which measures the spatial coherence; (b) light from two different wavefronts is combined using a beamsplitter in division-of-amplitude interference, which measures the temporal coherence.

imaging because it is the fundamental limit to the capability of a system to distinguish different objects.

5.2 Division-of-Wavefront Interference

References: Hecht, Optics §8

The first example of optical interference involves summing two wavefronts that were generated by one source and traveled different paths before being recombined. This was the first type of interference observed, having been noted by Thomas Young in the early 1800s. The mathematical derivation is a simple extension of the addition of 1-D traveling waves to three dimensions.

5.2.1 Addition of 3-D Traveling Waves with same $\omega$

Traveling waves also may be defined over two or three spatial dimensions; the waves have the form $f[x, y, t]$ and $f[x, y, z, t]$, respectively. The direction of propagation of such a wave in a multidimensional space is determined by a vector analogous to $k_0$; a 3-D wavevector $k_0$ has components $(k_0)_x, (k_0)_y, (k_0)_z$. The vector may be written:

$$k_0 = (k_0)_x \hat{x} + (k_0)_y \hat{y} + (k_0)_z \hat{z} = \begin{bmatrix} (k_0)_x \\ (k_0)_y \\ (k_0)_z \end{bmatrix}$$

The corresponding wave travels in the direction of the wavevector $k$ and has wavelength $\lambda_0 = \frac{2\pi}{k}$. In other words, the length of $k$ is the magnitude of the wavevector:

$$|k_0| = \sqrt{k_0 \cdot k_0} = \sqrt{(k_0)_x^2 + (k_0)_y^2 + (k_0)_z^2} = \frac{2\pi}{\lambda_0}$$

The angular temporal frequency $\omega_0$ is determined from the magnitude of the wavevector through the dispersion relation:

$$\omega_0 = v_0 \cdot |k_0| \implies \nu_0 = \frac{v_0}{\lambda_0}$$
Example: \((k_1)_y = (k_2)_y = 0\)

For illustration, consider a simple 2-D analogue of the 1-D traveling plane wave. The wave travels in the direction of a 2-D wavevector \(\mathbf{k}\) that lies the \(x - z\) plane, so that \(k_y = 0\):

\[
\mathbf{k}_0 = \begin{bmatrix}
(k_0)_x \\
0 \\
(k_0)_z
\end{bmatrix}
\]

The points of constant phase with phase angle \(\phi = C\) radians is the set of points in the 2-D space \(\mathbf{r} = [x, y, z] = (r, \theta)\) such that the scalar product \(\mathbf{k}_0 \cdot \mathbf{r} = C\):

\[
\mathbf{k}_0 \cdot \mathbf{r} = \mathbf{r} \cdot \mathbf{k}_0 = |\mathbf{r}| |\mathbf{k}_0| \cos[\theta] = (k_0)_x x + 0 \cdot y + (k_0)_z z = C\ \text{radians, for point of constant phase}
\]

Therefore, the equation of a 2-D wave traveling in the direction of \(\mathbf{k}_0\) with linear wavefronts is:

\[
f [x, y, t] = A \cos \left( (k_0)_x x + (k_0)_z z - \omega_0 t \right)
\]

In three dimensions, the set of points with the same phase lie on a planar surface so that the equation of the traveling wave is:

\[
f [x, y, z, t] = f [\mathbf{r}, t] = A \cos \left( (k_0)_x x + (k_0)_y y + (k_0)_z z - \omega_0 t \right)
\]

This plane wave could have been created by a point source at a large distance to the left and below the \(z\)-axis.
5.2 DIVISION-OF-WAVEFRONT INTERFERENCE

Now, we will apply the equation derived when adding oscillations with different temporal frequencies. In general, the form of the sum of two traveling waves is:

\[ f_1[x, y, z, t] + f_2[x, y, z, t] = A \cos(k_1 \cdot \mathbf{r} - \omega_1 t) + A \cos(k_2 \cdot \mathbf{r} - \omega_2 t) \]

\[ = 2A \cos(k_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t) \cdot \cos(k_{\text{mod}} \cdot \mathbf{r} - \omega_{\text{mod}} t) \]

where the average and modulation wavevectors are:

\[ k_{\text{avg}} = \frac{k_1 + k_2}{2} = \left( \frac{(k_1)_x + (k_2)_x}{2} \right) \hat{x} + \left( \frac{(k_1)_y + (k_2)_y}{2} \right) \hat{y} + \left( \frac{(k_1)_z + (k_2)_z}{2} \right) \hat{z} \]

\[ k_{\text{mod}} = \frac{k_1 - k_2}{2} = \left( \frac{(k_1)_x - (k_2)_x}{2} \right) \hat{x} + \left( \frac{(k_1)_y - (k_2)_y}{2} \right) \hat{y} + \left( \frac{(k_1)_z - (k_2)_z}{2} \right) \hat{z} \]

and the average and modulation angular temporal frequencies are:

\[ \omega_{\text{avg}} = \frac{\omega_0 + \omega_0}{2} = \omega_0 \]

\[ \omega_{\text{mod}} = \frac{\omega_0 - \omega_0}{2} = 0 \]

The average and modulation wavevectors \( k_{\text{avg}} \) and \( k_{\text{mod}} \) generally point in different directions and thus the corresponding waves move in different directions at velocities determined from:

\[ v_{\text{avg}} = \frac{\omega_{\text{avg}}}{|k_{\text{avg}}|} \]

\[ v_{\text{mod}} = \frac{\omega_{\text{mod}}}{|k_{\text{mod}}|} \]

Because the phase of the multidimensional traveling wave is a function of two parameters (its wavevector \( k_n \) and angular temporal frequency \( \omega_n \)), the phases of two traveling waves usually differ even if the temporal frequencies are equal. Since the temporal frequencies are equal, so must be the wavelengths:

\[ \lambda_1 = \lambda_2 = \lambda \rightarrow |k_1| = |k_2| \equiv |k| \equiv k \]

In this case, the component waves travel in different directions so the components of the wavevectors differ:

\[ k_1 = [(k_1)_x, (k_1)_y, (k_1)_z] \neq k_2 = [(k_2)_x, (k_2)_y, (k_2)_z] \]

The summation of two traveling waves with identical magnitudes and temporal frequencies is:

\[ f_1[x, y, z, t] + f_2[x, y, z, t] = A \cos(k_1 \cdot \mathbf{r} - \omega_0 t) + A \cos(k_2 \cdot \mathbf{r} - \omega_0 t) \]

\[ = 2A \cos(k_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t) \cdot \cos(k_{\text{mod}} \cdot \mathbf{r}) \]

because \( \omega_{\text{mod}} = 0 \). The superposition of two 2-D wavefronts with the same temporal frequency but traveling in different directions results in two multiplicative components: a traveling wave in the direction of \( k_{\text{avg}} \), and a stationary wave in space along the direction of \( k_{\text{mod}} \).
These two waves could have been generated at point sources located above and below the \( z \)-axis at a large distance to the left. This is the classic “Young’s double slit” experiment, where light from a single source is split into two waves (spherical waves in this case) that propagate large distances to the observation plane.

How two “tilted” plane waves are generated in the Young double-aperture experiment. The two apertures in the opaque screen on the left divide the incoming wave into two expanding spherical waves. After propagating a long distance, the spherical waves approximate plane waves that are tilted relative to the axis by \( \theta = \frac{d}{2L} \).
The “tilts” of the two waves are evaluated from the two (equal) distances:

\[ \tan [\theta] = \frac{(d/2)}{L} = \frac{d}{2L} \]

If \( L >> d \), then

\[ \tan [\theta] \cong \sin [\theta] \cong \theta \cong \frac{d}{2L} \]

The superposition of the two electric fields is:

\[ f[x,y,z,t] = f_1[x,y,z,t] + f_2[x,y,z,t] \]

\[ = 2A \cos \left[ \frac{k_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t}{2} \right] \cdot \cos [k_{\text{mod}} \cdot \mathbf{r}] \]

\[ = 2A \cos \left[ 2\pi \frac{z}{\lambda_0} \cos [\theta] - \omega_0 t \right] \cdot \cos \left[ 2\pi \frac{x}{\lambda_0} \sin [\theta] \right] \]

The first term (with the time dependence) is a traveling wave in the direction specified by \( \mathbf{k} = [0, 0, k_z] = \frac{2\pi}{\lambda_0} \cos [\theta] \cdot \hat{z} \), while the second term (with no dependence on time) is a spatial wave along the \( x \) direction. The amplitude variation of this modulation wave is:

\[ 2A \cos \left[ 2\pi \frac{x}{\lambda_0} \sin [\theta] \right] = 2A \cos \left[ 2\pi \left( \frac{x}{\lambda_0} \sin [\theta] \right) \right] \]

which is a sinusoidal oscillation with period \( \frac{\lambda_0}{\sin [\theta]} \).

The irradiance (measurable intensity) of the superposition is:

\[ |f[x,y,z,t]|^2 = 4A^2 \cos^2 \left[ 2\pi \frac{z}{\lambda_0} \cos [\theta] - \omega_0 t \right] \cos^2 \left[ 2\pi \frac{x}{\lambda_0} \sin [\theta] \right] \]

Both cosine terms can be rewritten using:

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

\[ \cos^2 \left[ 2\pi \frac{z}{\lambda_0} \cos [\theta] - \omega_0 t \right] = \frac{1}{2} \left( 1 + \cos \left[ 2\pi \frac{z}{\lambda_0} \frac{2\cos [\theta]}{2\cos [\theta]} - 2\omega_0 t \right] \right) \]

\[ \cos^2 \left[ 2\pi \frac{x}{\lambda_0} \sin [\theta] \right] = \frac{1}{2} \left( 1 + \cos \left[ 2\pi \frac{x}{\lambda_0} \frac{2\sin [\theta]}{2\sin [\theta]} \right] \right) \]

As before, the argument of the first term varies rapidly because \( \omega_0 \cong 10^{14} \text{ radians/sec} \). The result of the measurement is the average value of \( \frac{1}{r} \):

\[ \left\langle |f[x,y,z,t]|^2 \right\rangle = 4A^2 \cdot \frac{1}{2} \cdot \frac{1}{2} \left( 1 + \cos \left[ 2\pi \frac{x}{\lambda_0} \frac{2\sin [\theta]}{2\sin [\theta]} \right] \right) \]

\[ = A^2 \left( 1 + \cos \left[ 2\pi \frac{x}{\lambda_0} \frac{\sin [\theta]}{2\sin [\theta]} \right] \right) \]
This derivation may also be applied to find the irradiance of the individual component traveling waves:

\[ I_1 = \left| f_1 [x, y, z, t] \right|^2 = |A \cos \left[ k_1 \cdot r - \omega_0 t \right] |^2 = A^2 \cos^2 \left[ k_1 \cdot r - \omega_0 t \right] = A^2 \cdot \frac{1}{2} \]

\[ I_2 = \left| f_2 [x, y, z, t] \right|^2 = A^2 \cdot \frac{1}{2} = I_1 \equiv I_0 \]

So the irradiance of the sum of the two waves can be rewritten in terms of the irradiance of a single wave:

\[ \left| f [x, y, z, t] \right|^2 = 2I_0 \left( 1 + \cos \left[ \frac{2\pi x}{\lambda_0 \sin [\theta]} \right] \right) \]

This is a “biased sinusoid” that is positive or zero everywhere. In other words, the modulation of the irradiance is sinusoidal between 0 and 2I_0 \cdot (2) = 4I_0 with period \( X_0 = \frac{\lambda_0}{2 \sin [\theta]} \), so that the average irradiance is 2I_0. The period varies directly with \( \lambda \) and inversely with \( \sin [\theta] \); for small \( \theta \), the period of the sinusoid is large, for \( \theta = 0 \) (so that the both waves travel “straight down” the optical axis), there is no modulation of the irradiance. The alternating bright and dark regions of this time-stationary sinusoidal intensity pattern are called interference fringes. The shape, separation, and orientation of the interference fringes are determined by the form of and direction traveled by the incident wavefronts, and the pattern provides information about the incident waves. In fact, if one of the two waves is known, then the parameters of the other wave can be inferred from the interference pattern; this is the principle behind holography. The argument of the cosine function is the optical phase difference of the two waves, which is a function of the wavelength \( \lambda_0 \) and the angle \( \theta \) of each beam from the axis of symmetry. The phase difference does not vary with time, which means that the two waves are temporally coherent. At locations where the optical phase difference is an even multiple of \( \pi \), the cosine evaluates to unity and a maximum of the interference pattern results; this is constructive interference. If the optical phase difference is an odd multiple of \( \pi \), the cosine evaluates to -1 and the irradiance is zero; this is destructive interference.

\[ \text{Interference of two “tilted plane waves” with the same wavelength. The two component traveling waves are shown as “snapshots” at one instant of time on the left (white = 1, black = -1); the sum of the two is shown in the center (white = 2, black = -2), and the squared magnitude on the right (white = 4, black = 0). The modulation in the vertical direction is constant, while that in the horizontal direction is a traveling wave and “averages” out to a constant value of } \frac{1}{2}. \]
The amplitude and irradiance observed at one instant of time when the irradiance at the origin ("on axis") is a maximum is shown:

Interference patterns observed along the $x$-axis at one value of $z$: (a) bipolar amplitude fringes, with period equal to $\frac{\lambda}{\sin[\theta]}$; nonnegative irradiance (intensity) fringes, with period equal to $\frac{\lambda}{2\sin[\theta]}$. This pattern is averaged over time and scales by a factor of $\frac{1}{2}$.

Again, the traveling wave in the images of the amplitude and intensity of the superposed images moves in the $z$-direction (to the right), thus blurring out the oscillations in the $z$-direction. The oscillations in the $x$-direction are preserved as the interference pattern, which is plotted as a function of $x$ below. Note that the spatial frequency of the intensity fringes is twice as large as that of the amplitude fringes.

Irradiance patterns observed at the output plane at several instants of time, showing that the spatial variation of the irradiance is preserved but the averaging reduces the maximum value by half.

### 5.2.2 Superposition of Two Plane Waves with $\omega_1 \neq \omega_2$

For further illustration, consider the case the two waves travel in the same directions but with different temporal frequencies $\omega_1 \neq \omega_2$, which means that $\mathbf{k}_1 = \mathbf{k}_2$ and $|\mathbf{k}_1| \neq |\mathbf{k}_2|$. The average and modulation wavevectors are found as before, but the fact that $\omega_{\text{mod}} \neq 0$ means that modulation wave is no longer stationary along the $x$-axis.
\[ \lambda_1 = 8 \text{ units} \]
\[ \omega_1 = \frac{1}{8} \text{ radians per second} \]
\[ \theta_1 = -\frac{\pi}{6} \text{ radians} = -30^\circ \]

\[ \lambda_2 = 12 \text{ units} \]
\[ \omega_2 = \frac{1}{12} \text{ radians per second} \]
\[ \theta_2 = +\frac{\pi}{6} \text{ radians} = +30^\circ \]

- average and modulation frequencies:
\[ \omega_1 = \frac{3}{2} \omega_2 \]
\[ \omega_{\text{avg}} = \frac{\frac{1}{8} + \frac{1}{12}}{2} = \frac{5}{48} \text{ radians per second} \]
\[ \omega_{\text{mod}} = \frac{\frac{1}{8} - \frac{1}{12}}{2} = \frac{1}{48} \text{ radians per second} \]

- Length of Wavevectors:

  \(- \lambda_1 = 8 \text{ units} \implies \)
  \[ |k_1| = \frac{2\pi}{8} = \frac{\pi}{4} \cong 0.785 \text{ radians per second} \]
  \[ v_1 = \frac{\omega_1}{|k_1|} = \frac{\frac{1}{8} \text{ radians per second}}{\frac{\pi}{4} \text{ unit length}} = \frac{1}{\frac{2\pi}{2 \text{ second}}} \]

  \(- \lambda_1 = 12 \text{ units} \implies \)
  \[ |k_2| = \frac{2\pi}{8} = \frac{\pi}{4} \cong 0.785 \text{ radians per second} \]
  \[ v_2 = \frac{\omega_2}{|k_2|} = \frac{\frac{1}{8} \text{ radians per second}}{\frac{\pi}{4} \text{ unit length}} = \frac{1}{\frac{2\pi}{2 \text{ second}}} \]

  \(- v_1 = v_2 \implies \) nondispersive waves

- Derive \( k_1 \)

  \[ (k_1)_x = \frac{2\pi}{\lambda_1} \sin |\theta_1| = \frac{2\pi}{8} \left( \frac{1}{2} \right) = -\frac{\pi}{8} \cong -0.393 \text{ radians per unit length} \]

  \[ (k_1)_z = \frac{2\pi}{\lambda_1} \cos |\theta_1| = \frac{2\pi}{8} \frac{\sqrt{3}}{2} \cong 0.680 \text{ radians per unit length} \]

- Check: \[ |k_1| = \sqrt{\left( -\frac{\pi}{8} \right)^2 + \left( \frac{\sqrt{3} \pi}{8} \right)^2} = \frac{\pi}{4} \] (agrees with result from wavelength)
5.2 DIVISION-OF-WAVEFRONT INTERFERENCE

- Components of $k_1$:

$$k_1 = \begin{bmatrix}
-\frac{\pi}{8} \\
0 \\
+\frac{\sqrt{3}\pi}{8}
\end{bmatrix} \cong \begin{bmatrix}
-0.393 \\
0 \\
+0.680
\end{bmatrix}$$

- Derive $k_2$

$$(k_2)_x = \frac{2\pi}{\lambda_2} \sin \left(-\theta_1\right) = \frac{\pi}{6} \cdot \left(+\frac{1}{2}\right) = +\frac{\pi}{12} \cong +0.262 \text{ radians/unit length}$$

$$(k_2)_z = \frac{2\pi}{\lambda_2} \cos \left[-\theta_1\right] = \frac{2\pi}{12} \cdot \frac{\sqrt{3}}{2} = \frac{\sqrt{3}\pi}{12} \cong 0.453 \text{ radians/unit length}$$

- Check: $|k_2| = \sqrt{\left(\frac{\pi}{12}\right)^2 + \left(\frac{\sqrt{3}\pi}{12}\right)^2} = \frac{\pi}{6}$ (also agrees)

- Components of $k_2$:

$$k_2 = \begin{bmatrix}
+\frac{\pi}{12} \\
0 \\
+\frac{\sqrt{3}\pi}{12}
\end{bmatrix} \cong \begin{bmatrix}
+0.262 \\
0 \\
+0.453
\end{bmatrix}$$

- Average and modulation wavevectors:

$$k_{\text{avg}} = \frac{k_1 + k_2}{2} = \frac{1}{2} \begin{bmatrix}
-\frac{\pi}{8} \\
0 \\
+\frac{\sqrt{3}\pi}{8}
\end{bmatrix} + \begin{bmatrix}
+\frac{\pi}{12} \\
0 \\
+\frac{\sqrt{3}\pi}{12}
\end{bmatrix} = \begin{bmatrix}
-\frac{\pi}{48} \\
0 \\
\frac{5\sqrt{3}\pi}{48}
\end{bmatrix}$$

$$|k_{\text{avg}}| = \frac{\sqrt{19\pi}}{24} \text{ radians/unit length} \cong 0.571 \text{ radians/unit length}$$

$$\theta_{\text{avg}} = \tan^{-1} \left[\frac{-\frac{\sqrt{3}\pi}{48}}{\frac{5\sqrt{3}\pi}{48}}\right] \cong -0.115 \text{ radians} \cong -6.6^\circ$$

$$k_{\text{mod}} = \frac{\sqrt{7\pi}}{24} \text{ radians/unit length} \cong 0.346 \text{ radians/unit length}$$

$$\theta_{\text{mod}} = \tan^{-1} \left[\frac{-\frac{\sqrt{3}\pi}{48}}{\frac{\sqrt{3}\pi}{48}}\right] \cong -1.237 \text{ radians} \cong -71^\circ$$

Note that both the average and modulation waves travel; they are headed in different directions with different frequencies and different velocities. The temporal frequencies are $\nu_{\text{avg}} = \frac{\pi}{48} \text{ Hz}$ and $\nu_{\text{mod}} = \frac{\sqrt{3}\pi}{48} \text{ Hz}$. In this case, the time-averaged irradiance of the modulation wave is also “averaged out” to create a uniform irradiance pattern and no stationary fringe pattern is visible from interference of beams with two different wavelengths. Put another way, the phase difference encapsulated in the phase of the modulation oscillation is a function of time and space; this variation means that the light is not coherent.
Chapter 5 Optical Interference and Coherence

Sum of two sinusoidal traveling waves where the periods are related by \( \lambda_2 = \frac{3}{2} \lambda_1 \). The two waves travel in the directions ±30°, respectively. The resulting amplitude sum and power are depicted as “snapshots” at one instant of time. Since the modulation wave now travels too, both waves are averaged to constant values and no fringes are visible.

Intensity patterns observed at the output plane at several instants of time. The velocity of the modulation wave makes this pattern “migrate” towards \(-x\), and thus the time-averaged pattern is a constant; no interference is seen.

The same principles just discussed may be used to determine the form of interference fringes from wavefronts with other shapes. Some examples will be considered in the following sections.

5.3 Fringe Visibility – Temporal Coherence

The visibility of a sinusoidal fringe pattern is a quality that corresponds quite closely to the concept of modulation of sinusoids. Given a nonnegative sinusoidal irradiance (intensity) distribution with maximum \( I_{\text{max}} \) and minimum \( I_{\min} \) (so that \( I_{\min} \geq 0 \)), the visibility of the sinusoidal fringe pattern is:

\[
V = \frac{I_{\text{max}} - I_{\min}}{I_{\text{max}} + I_{\min}}
\]
Note that if $I_{\text{min}} = 0$, then $V = 1$ regardless of the value of $I_{\text{max}}$. The visibility of the fringe pattern is determined by the relative irradiances of the individual wavefronts and by the coherence of the light source, which determines the distance over which the phase of the light from the source is well defined.

To introduce the concept of coherence, consider first the Young's two-aperture experiment where the source is composed of equal-amplitude emission at two distinct wavelengths $\lambda_1$ and $\lambda_2$ incident on the observation screen at $\pm \theta$. Possible pairs of wavelengths could be those of the sodium doublet ($\lambda = 589.0 \text{ nm}$ and $589.6 \text{ nm}$), or the pair of lines emitted by a "greenie" He:Ne laser ($\lambda_1 = 543 \text{ nm}$ (green), $\lambda_2 = 594 \text{ nm}$ (yellow)). In air or vacuum, the corresponding angular frequencies obviously are $\omega_1 = c |k_1| = \frac{2\pi c}{\lambda_1}$ and $\omega_2 = \frac{2\pi c}{\lambda_2}$.

To find the irradiance pattern created by the interference of the four beams, we must compute the superposition of the amplitude of the electromagnetic field, find its squared-magnitude, and then determine the average over time. The sum of the four component terms is straightforward to compute by recognizing that it is the sum of the amplitude patterns from the pairs of waves with the same wavelength. We have already shown that the sum of the two terms with $\lambda = \lambda_1$ is:

$$ f_1 [x, z; \lambda_1] + f_2 [x, z; \lambda_1] = 2A \cos \left[ \frac{2\pi z}{\lambda_1} \cos \theta - \omega_1 t \right] \cos \left[ \frac{2\pi x}{\lambda_1} \sin \theta \right] $$

a traveling wave in the $+z$-direction and a stationary sinusoid along the $x$-direction. Now add a second pair of plane waves with same "tilts" $\theta$ but the light has different wavelength $\lambda_2 \neq \lambda_1$. The interference pattern from this pair of waves alone is easy to evaluate:

$$ f_3 [x, z; \lambda_2] + f_4 [x, z; \lambda_2] = 2A \cos \left[ \frac{2\pi z}{\lambda_2} \cos \theta - \omega_2 t \right] \cos \left[ \frac{2\pi x}{\lambda_2} \sin \theta \right] $$

so the period of the interference pattern differs only by the wavelength. If both wavelengths are used at the same time, the amplitudes of the four waves may still be added in pairs:

$$ \sum_{n=1}^{4} f_n [x, z; \lambda] = (f_1 [x, z; \lambda_1] + f_2 [x, z; \lambda_1]) + (f_3 [x, z; \lambda_2] + f_4 [x, z; \lambda_2]) $$

$$ = 2A \cos \left[ \frac{2\pi x}{(k_{\text{avg}})_1} \cos \theta \right] \cos \left[ (k_{\text{avg}})_1 z - \omega_1 t \right] + 2A \cos \left[ \frac{2\pi x}{(k_{\text{avg}})_2} \cos \theta \right] \cos \left[ (k_{\text{avg}})_2 z - \omega_2 t \right] $$

where:

$$ \left| (k_{\text{avg}})_1 \right| \equiv (k_{\text{avg}})_1 = \frac{2\pi}{\lambda_1 \cos \theta} \quad \left| (k_{\text{avg}})_2 \right| \equiv (k_{\text{avg}})_2 = \frac{2\pi}{\lambda_2 \cos \theta} $$

The velocities of the average waves are:

$$ (v_{\text{avg}})_1 = \frac{\omega_1}{(k_{\text{avg}})_1} = \frac{2\pi \nu_1}{\lambda_1 \cos \theta} = \frac{\nu_1 \lambda_1}{2} \cos \theta = \frac{c}{2} \cos \theta $$

$$ v_2 = \frac{\omega_2}{(k_{\text{avg}})_2} = \frac{2\pi \nu_2}{\lambda_2 \cos \theta} = \frac{\nu_2 \lambda_2}{2} \cos \theta = \frac{c}{2} \cos \theta = v_1 $$

which means that both “average” waves travel down the $z$-axis with the same velocity, as they should.
because there is no dispersion. The field at the observation plane is now the sum of two terms:

\[ \sum_{n=1}^{4} f_n [x, z; \lambda] = 2A \cos \left[ 2\pi \frac{x}{\lambda_1 \sin \theta} \right] \cos \left[ (k_{avg})_1 (z - v_1 t) \right] + 2A \cos \left[ 2\pi \frac{x}{\lambda_2 \sin \theta} \right] \cos \left[ (k_{avg})_2 (z - v_1 t) \right] \]

The irradiance is the time average of the squared magnitude of the summation, which has three terms:

\[ \left( \sum_{n=1}^{4} f_n [x, z; \lambda] \right)^2 = |2A|^2 \cos \left[ 2\pi \frac{x}{\lambda_1 \sin \theta} \right] \cos \left[ (k_{avg})_1 z - \omega_1 t \right] + \cos \left[ 2\pi \frac{x}{\lambda_2 \sin \theta} \right] \cos \left[ (k_{avg})_2 z - \omega_2 t \right]^2 \]

\[ = 4A^2 \left( \cos^2 \left[ 2\pi \frac{x}{\lambda_1 \sin \theta} \right] \right) \left( \cos^2 \left[ (k_{avg})_1 z - \omega_1 t \right] + \cos^2 \left[ 2\pi \frac{x}{\lambda_2 \sin \theta} \right] \cos^2 \left[ (k_{avg})_2 z - \omega_2 t \right] \right) + 4A^2 \left( 2 \cos \left[ 2\pi \frac{x}{\lambda_1 \sin \theta} \right] \cos \left[ 2\pi \frac{x}{\lambda_2 \sin \theta} \right] \cos \left[ (k_{avg})_1 z - \omega_1 t \right] \cos \left[ (k_{avg})_2 z - \omega_2 t \right] \right) \]

The time averages of each of the first two terms (squares of the traveling wave cosines) are identically \( \frac{1}{2} \), but the time average of the product of the two cosine traveling waves may be written as the summation of two bipolar cosines:

\[ \cos \left[ (k_{avg})_1 z - \omega_1 t \right] \cdot \cos \left[ (k_{avg})_2 z - \omega_2 t \right] = \frac{1}{2} \cos \left[ (k_{avg})_1 + (k_{avg})_2 \right] z - (\omega_1 + \omega_2) t + \frac{1}{2} \cos \left[ (k_{avg})_1 - (k_{avg})_2 \right] z - (\omega_1 - \omega_2) t \]

The time average evaluates to 0 if the time period \( T >> \nu^{-1}_{mod} \), so the irradiance of the summation is:

\[ \mathcal{I} [x, z] = 2A^2 \left( \cos^2 \left[ 2\pi \frac{x}{\lambda_1 \sin \theta} \right] \right) + \cos^2 \left[ 2\pi \frac{x}{\lambda_2 \sin \theta} \right] \left( \frac{1}{2} + 0 \right) \]

\[ = 4A^2 \left( \cos^2 \left[ 2\pi \frac{x}{\lambda_1 \sin \theta} \right] \right) \left( \frac{1}{2} + \cos^2 \left[ 2\pi \frac{x}{\lambda_2 \sin \theta} \right] \right) \]

This is a very simple and significant result; it shows that the irradiance from a two-aperture system from two wavelengths is the sum of irradiances from the individual wavelengths.
The sum of four tilted plane waves can be calculated by summing the pair due to one wavelength and that due to the other.

This expression as the sum of two squared cosines may be rewritten yet again by using the identity:

$$\cos^2 [\theta] = \frac{1}{2} + \frac{1}{2} \cos [2\theta]$$

to obtain:

$$I [x, z] = 2A^2 \left( \frac{1}{2} + \frac{1}{2} \cos \left[ 2\pi \frac{x}{\lambda_1 (2 \sin \theta)} \right] + \frac{1}{2} + \frac{1}{2} \cos \left[ 2\pi \frac{x}{\lambda_2 (2 \sin \theta)} \right] \right)$$

$$= 2A^2 \left( \frac{1}{2} + \frac{1}{2} \cos \left[ 2\pi \frac{x}{\lambda_1 (2 \sin \theta)} \right] + \cos \left[ 2\pi \frac{x}{\lambda_2 (2 \sin \theta)} \right] \right)$$

The sum of the two stationary cosine waves also may be recast yet one more time as the product of cosines with the average and modulation frequencies:

$$\cos [2\pi \xi_1 x] + \cos [2\pi \xi_2 x] = 2 \cos \left[ 2\pi \left( \frac{\xi_1 + \xi_2}{2} \right) x \right] \cdot \cos \left[ 2\pi \left( \frac{\xi_1 - \xi_2}{2} \right) x \right]$$

where:

$$\xi_1 = \frac{2 \sin [\theta]}{\lambda_1}$$
$$\xi_2 = \frac{2 \sin [\theta]}{\lambda_2}$$

$$\Rightarrow \frac{\xi_1 \pm \xi_2}{2} = \sin [\theta] \left( \frac{1}{\lambda_1} \pm \frac{1}{\lambda_2} \right) = \sin [\theta] \left( \frac{\lambda_2 \pm \lambda_1}{\lambda_1 \lambda_2} \right)$$
Substitute these expressions into the summation:

\[
\cos \left[ 2\pi \frac{x}{\lambda_1} \right] + \cos \left[ 2\pi \frac{x}{\lambda_2} \right] = \cos \left[ 2\pi \frac{x}{c} \right] + \cos \left[ 2\pi \frac{x}{c} \right] \]

\[
= \cos \left[ 2\pi \left( \frac{2\nu_1 \sin[\theta]}{c} \right) x \right] + \cos \left[ 2\pi \left( \frac{2\nu_2 \sin[\theta]}{c} \right) x \right] \]

\[
= 2 \cos \left[ 2\pi \left( \frac{2\nu_1 + \nu_2 \sin[\theta]}{c} \right) x \right] \cdot \cos \left[ 2\pi \left( \frac{2\nu_2 - \nu_2 \sin[\theta]}{c} \right) x \right] \]

\[
= 2 \cos \left[ 2\pi \left( \frac{x}{\nu_{avg} \sin[\theta]} \right) \right] \cdot \cos \left[ 2\pi \left( \frac{x}{\nu_{mod} \sin[\theta]} \right) \right] \]

where \( \nu_{avg} \equiv \frac{\nu_1 + \nu_2}{2} \) and \( \nu_{mod} \equiv \frac{\nu_1 - \nu_2}{2} \). These may be expressed in terms of the average and modulation wavelengths:

\[
\nu_{avg} = \frac{\nu_1 + \nu_2}{2} = \frac{\lambda_1 + \lambda_2}{2} = \frac{c}{2} \left( \frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right) \]

\[
= \frac{c}{2} \left( \frac{\lambda_1 + \lambda_2}{\lambda_1 \lambda_2} \right) = \frac{c}{\lambda_1 \lambda_2} \lambda_{avg} \]

\[
\nu_{mod} = \frac{\nu_1 - \nu_2}{2} = \frac{\lambda_1 - \lambda_2}{2} = \frac{c}{2} \left( \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right) \]

\[
= \frac{c}{2} \left( \frac{\lambda_2 - \lambda_1}{\lambda_1 \lambda_2} \right) = \frac{c}{\lambda_1 \lambda_2} \lambda_{mod} \]

So the expression for the irradiance becomes:

\[
I [x, z] = 2A^2 \left( 1 + \frac{1}{2} \left( 2 \cos \left[ 2\pi \left( \frac{x}{\nu_{avg} \sin[\theta]} \right) \right] \cdot \cos \left[ 2\pi \left( \frac{x}{\nu_{mod} \sin[\theta]} \right) \right] \right) \right) \]

\[
I [x, z] = 2A^2 \left( 1 + \cos \left[ 2\pi \left( \frac{x}{\nu_{avg} \sin[\theta]} \right) \right] \cdot \cos \left[ 2\pi \left( \frac{x}{\nu_{mod} \sin[\theta]} \right) \right] \right) \]

\[
\equiv 2A^2 \left( 1 + \cos \left[ \frac{2\pi x}{D_{avg}} \right] \cdot \cos \left[ \frac{2\pi x}{D_{mod}} \right] \right) \]

where the periods are:

\[
D_{avg} \equiv \frac{c}{2\nu_{avg} \sin[\theta]} = \frac{\lambda_1 \lambda_2}{2 \sin[\theta] \cdot \lambda_{avg}} \propto (\lambda_{avg})^{-1} \]

\[
D_{mod} \equiv \frac{c}{2\nu_{mod} \sin[\theta]} = \frac{\lambda_1 \lambda_2}{2 \sin[\theta] \cdot \lambda_{mod}} \propto (\Delta \lambda)^{-1} \]

If the two wavelengths are close together, so that \( \lambda_1 \approx \lambda_2 \approx \lambda_{avg} >> \lambda_{mod} \), then the expressions for
5.3 FRINGE VISIBILITY – TEMPORAL COHERENCE

the periods of the two component oscillations may be simplified:

\[ D_{\text{mod}} = D_{\text{avg}} \cdot \frac{\lambda_{\text{avg}}}{\Delta \lambda} \gg D_{\text{avg}} \]

In words, the period of the modulation due to \( \lambda_{\text{mod}} \) is much longer than that due to \( \lambda_{\text{avg}} \) if the emitted wavelengths are approximately equal. Where the amplitude with the longer period \( D_{\text{mod}} \) decays to zero, the short-period fringes are no longer visible. In fact, the sinusoidal fringes due to \( \lambda_{\text{avg}} \) are visible over a range of \( x \) equal to range of \( x \) between the zeros of \( D_{\text{mod}} \), i.e., half the period of \( D_{\text{mod}} \). The pattern resulting from the example considered above is shown. Note that the amplitude of maxima of the irradiance fringes decreases away from the center of the observation screen, where the optical path lengths are equal for all four beams, and thus where they will add constructively.

\[
\text{Interference patterns from two wavelengths with same input amplitude: (a) the two amplitude patterns differ in period in proportion to the wavelength; (b) the sum of the two amplitude patterns at one instant of time; (c) the squared magnitude at the same instant, showing that the amplitude of the fringe varies with } x.\]

In this case where \( \Delta \lambda \ll \lambda_{\text{avg}} \), the fringes are visible over a large interval of \( x \). We speak of such a light source as temporally coherent; the phase difference of light emitted at \( \lambda_1 \) and at \( \lambda_2 \) changes slowly with time, and thus with the position along the \( x \)-axis. Therefore, fringes are visible over a large range of \( x \). On the other hand, if \( \Delta \lambda \) is of the same order as \( \lambda_{\text{avg}} \), the wavelengths are widely separated. The phase difference of light emitted at the two extreme wavelengths changes rapidly with time, and thus with position along the \( x \)-axis. The fringes are visible only if the phase difference remains approximately constant (for \( x = 0 \)) over the averaged time interval. Such a source is said to be temporally incoherent. It is difficult (though not impossible) to see fringes generated by a temporally incoherent source.

5.3.1 Coherence of Light

The coherence of an electromagnetic field is a description of the correlation properties (i.e., the statistics) of the optical phase, which determines the interference properties of the field, i.e., how radiation from different points on the object or that has travelled different optical paths will interact during the imaging process. If the radiation is coherent, then it will interfere constructively and destructively when it produces the image. Incoherent radiation cannot interfere, and thus the image has significantly different properties. Imaging systems that act in coherent light are linear in the amplitude of the field, while systems acting in incoherent light are linear in irradiance (time average of the squared magnitude of the amplitude).

The measure of the coherence of light is based on the statistics of the phase difference measured at two points that differ in location and/or time:

\[ \Delta \Phi [R_1, t_1; R_2, t_2] \equiv \Phi [R_1, t_1] - \Phi [R_2, t_2] \]
which may be written in the equivalent form:
\[
\Delta \Phi [\mathbf{R}_1, t_1; \Delta \mathbf{R}, \Delta t] = \Phi [\mathbf{R}_1, t_1] - \Phi [\mathbf{R}_1 + \Delta \mathbf{R}, t_1 + \Delta t]
\]
where
\[
\mathbf{R}_2 = \mathbf{R}_1 + \Delta \mathbf{R}
\]
\[
t_2 = t_1 + \Delta t
\]

A field that is spatially (or laterally) coherent exhibits a fixed phase difference measured at any pair of points in space separated by the same vector \(\Delta \mathbf{R}\) and at the same time (\(\Delta t = 0\)). A radiation field is temporally (or longitudinally) coherent if phase differences evaluated at any point in space at two different times separated by \(\Delta t\) (i.e., such that \(\Delta \mathbf{R} = 0\)) are invariant. In other words, two beams are combined after one has been delayed in time to determine whether the phase difference is preserved; this leads to the concept of coherence length. Spatial coherence is tested by a division-of-wavefront interferometer that can “shear” a part of the wavefront to one side to superpose it on another part. Light from two different points on a wavefront can interfere if the source is “small.” Temporal coherence is tested using a division-of-amplitude interferometer that can delay one part of the wavefront relative to another; it is a measure of the frequency bandwidth of a point source; a source with a narrow bandwidth is temporally coherent over a longer distance.

The phase difference of a field of light that is both spatially and temporally coherent is identical for all pairs of points separated by the same increments \([\Delta \mathbf{R}, \Delta t]\), regardless of the initial location \([\mathbf{R}_1, t_1]\), which means that the light comes from a point source of purely monochromatic light. Pure coherence is an idealized concept, because neither do ideal point sources nor truly monochromatic waves exist. A true monochromatic wave must have infinite length.

If the two superposed plane waves have the same wavelength, then \(\omega_1 = \omega_2 = \omega_{\text{avg}}\) and \(\omega_{\text{mod}} = 0\), which means that the modulation wave remains fixed in space. The traveling wave is averaged out in the irradiance, leaving only the fixed modulation wave.

\[
I [x, y, z, t] = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |E [x, y, z, t]|^2 \, dt = 4 \cdot \cos^2 \left( k_{\text{mod}} \cdot \mathbf{r} \right)
\]

This stationary pattern of irradiance demonstrates the interference between the two plane waves. If the two waves have different wavelengths, then the modulation pattern also moves and “smears out” the interference pattern. Maxima occur at values of the position vector \(\mathbf{r}\) such that \(k_{\text{mod}} \cdot \mathbf{r} = n \pi\), where \(n\) is any integer. The “strength” of each maximum is 4 units, which is twice as large as the irradiance of the two waves added independently. Minima with “strength” 0 occur halfway between the maxima. This phenomenon of “amplification” and “cancellation” in coherent light does not occur for incoherent light, and is the reason why the transfer functions differ for the two cases.

It is straightforward to superpose wavefronts from other sources, such as cylinders or spheres, to generate modulation waves that have different “shapes.”

Strictly speaking, no radiation field is perfectly coherent or incoherent, but the psf and OTF of optical imaging systems are easily computed in these two limiting cases and we will assume that one of them is applicable.

### 5.3.2 Coherence Time and Coherence Length

If two wavelengths emitted by the source are separated by \(\Delta \lambda\), the corresponding frequency difference often is called the bandwidth of the source:

\[
\Delta \nu \equiv |\nu_1 - \nu_2| = \frac{|c}{\lambda_1} - \frac{c}{\lambda_2} = c \cdot \frac{\lambda_2 - \lambda_1}{\lambda_1 \cdot \lambda_2} = c \cdot \frac{\Delta \lambda}{\lambda_1 \cdot \lambda_2}
\]

Now consider that the spectrum of the source includes a third wavelength \(\lambda_3\) with the same “strength” located midway between the extrema \(\lambda_1\) and \(\lambda_2\) (i.e., \(\lambda_3 = \lambda_{\text{avg}}\)), the factors \(\Delta \lambda\) and
5.3 FRINGE VISIBILITY – TEMPORAL COHERENCE

\( \lambda_{\text{avg}} \) are unchanged, but the irradiance pattern must be different in some way. This pattern is more difficult to calculate directly as before, but the result may be modeled easily by recognizing that the wavefronts generated by \( \lambda_3 \) through the two apertures combine in amplitude to create a third pattern of sinusoidal fringes with a spatial period between those due to \( \lambda_1 \) and \( \lambda_2 \). The three such patterns may be summed and squared to model the irradiance fringes. Consider first the individual fringe patterns due to the extrema \( \lambda_1 \) and \( \lambda_2 \) as shown in (a):

(a) Irradiance pattern resulting from two wavelengths with equal “powers”, showing the long-period fringes due to \( \Delta \lambda \) and the short-period fringes due to \( \lambda_{\text{avg}} \); (b) Irradiance pattern after adding a third wavelength at \( \lambda_{\text{avg}} \) with the same “power” – the distance between peaks of the fringe pattern has increased; (c) pattern obtained from light over the full bandwidth between maximum and minimum – amplitudes of neighboring maxima are smaller.

The irradiance pattern generated from the superposition of these fringe patterns exhibits the same “short-period” fringes due to \( \lambda_{\text{avg}} \) and the same long-period fringes due to \( \Delta \lambda \), but the amplitudes of the adjacent maxima of the long period fringes varies; the distance between the “maxima of the maxima” is longer. Now generalize to the case where the spectrum contains equal amplitudes of all wavelengths between \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \); the distance between the maxima gets larger yet so that the most visible fringes are present only in the region centered about \( x = 0 \).

Because the region where the short-period fringes created by the three-line source are visible is similar in size to that from the two-line source, but (infinitely) much smaller than the region of interference from the single-line source, we say that light from the two wavelengths \( \lambda_{\text{max}} \) and \( \lambda_{\text{min}} \) and light from \( \lambda_{\text{max}} \), \( \lambda_{\text{min}} \), and \( \lambda_{\text{avg}} \) are equally coherent, the light from one wavelength is completely coherent. In short, coherence is quantified from the temporal bandwidth, which may be evaluated from the wavelength passband:

\[
\Delta \nu = \nu_{\text{max}} - \nu_{\text{min}} = \frac{c}{\lambda_{\text{min}}} - \frac{c}{\lambda_{\text{max}}}
\]

\[
= c \left( \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}} \cdot \lambda_{\text{min}}} \right) = c \cdot \frac{\Delta \lambda}{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}
\]

where \( \Delta \lambda \equiv \lambda_{\text{max}} - \lambda_{\text{min}} \). Note that the dimensions of \( \Delta \nu \) are \((\text{sec})^{-1}\). The time delay over which the phase difference of light emitted from one source point is deterministic (and thus over which fringes may be generated) is the reciprocal \((\Delta \nu)^{-1}\):

\[
\Delta \tau = \frac{1}{\Delta \nu} = \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{c \cdot \Delta \lambda},
\]

which is called the coherence time. Obviously, if \( \Delta \lambda \) is large, then so is \( \Delta \nu \) and the corresponding coherence time is small. The coherence length \( \Delta \ell \) is the distance traveled by an electromagnetic
wave during the coherence time:

\[ c \cdot \Delta \tau \equiv \Delta \ell = \frac{c}{\Delta \nu} = \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{\Delta \lambda} \]

This is a measure of the length of the electromagnetic wave “packet” over which the phase difference is predictable. Recall that for interference of waves from a source with wavelength range \( \Delta \lambda \), the range of coordinate \( x \) over which irradiance fringes are visible is half of the modulation period:

\[ \frac{D_{\text{mod}}}{2} = \frac{1}{2 \cdot \sin[\theta]} \left( \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right)^{-1} \frac{1}{\lambda_{\text{max}} \cdot \lambda_{\text{min}}} \frac{1}{\Delta \lambda} \]

\[ = \frac{1}{2 \cdot \sin[\theta]} \left( \frac{1}{c \cdot \Delta \nu} \right) = \frac{1}{2 \cdot \sin[\theta]} \cdot \Delta \ell \]

In words, the domain of \( x \) over which fringes are visible is proportional to the coherence length; if the bandwidth is “wide,” then the coherence length \( \Delta \ell \) is “short” and the fringes are visible over only a small region.

Lasers supply the most coherent light available because, to a very good approximation, many lasers emit a single wavelength \( \lambda_0 \), so that \( \Delta \lambda = 0 \implies \Delta \ell = \infty \implies D_{\text{mod}} = \infty \) and fringes are visible at all \( x \). A coherent source should be employed when an optical interference pattern is to be used to measure a parameter of the system, such as the optical image quality (as was used to test the Hubble space telescope). Thus the range of \( x \) over which fringes are visible is determined by the coherence length (and thus the bandwidth) of the source. Therefore, the visibility of the interference fringes may be used as a measure the source coherence.

**“White-Light” Fringes**

The range of visible wavelengths of light is generally taken to be:

\[ \Delta \lambda = \lambda_{\text{red}} - \lambda_{\text{blue}} \approx 700 \text{ nm} - 400 \text{ nm} = 300 \text{ nm} \]

The corresponding frequency bandwidth is:

\[ \Delta \nu = c \left( \frac{1}{\lambda_{\text{blue}}} - \frac{1}{\lambda_{\text{red}}} \right) \approx 3 \times 10^8 \text{ m/s} \cdot \left( \frac{1}{400 \text{ nm}} - \frac{1}{700 \text{ nm}} \right) \approx 3.21 \times 10^{14} \text{ Hz} \]

The width of the modulation region is:

\[ D_{\text{mod}} = \frac{\lambda_1 \lambda_2}{2 \sin[\theta] \cdot \lambda_{\text{mod}}} \]

\[ = \frac{700 \text{ nm} \cdot 400 \text{ nm}}{\sin[\theta] \cdot 300 \text{ nm}} \]

\[ \approx \frac{1 \text{ \mu m}}{\theta} \]

If the angle \( 2\theta \) between the interfering beams is small, then \( D_{\text{mod}} \) may be large. For example, we can expect that \( D_{\text{mod}} \) would have to be of the order of millimeters to be visible, say \( D_{\text{mod}} = 5 \text{ mm} \), which implies that the angle would have to be:

\[ 2\theta \approx \frac{1}{2500} \text{ radians} \approx 80 \text{ arc seconds} \approx 1.3 \text{ arc minutes} \]

**5.3.3 Polarization and Fringe Visibility**

Up to this point, we have ignored the effect on the orientation of the electric field vectors on the sum of the fields. In fact, this is an essential consideration; two orthogonal electric field vectors cannot add to generate a time-invariant modulation in the irradiance. Consider the sum of two electric field
5.4 VAN CITTERT - ZERNIKE THEOREM

The size of the region over which interference fringes are visible is related to the “degree” of coherence of the sources through the van Cittert-Zernike theorem, which says that the (normalized) Fourier transform of the irradiance distribution across the source is the degree of coherence, also called the spatial coherence function.

\[ \mathcal{F}_2 \{ I [x,y] \} = \langle E [r_1] \cdot E^* [r_2] \rangle \equiv \Gamma [r_1;r_2] \]

5.5 Division-of-Amplitude Interferometers

We have seen several times now that optical interference results when two (or more) waves are superposed in such a way to produce a time-stationary spatial modulation of the superposed electric field, which may be observed by eye or a photosensitive detector. Interferometers use this result to measure different parameters of the light (e.g., wavelength \( \lambda \), bandwidth \( \Delta \lambda \) or coherence length \( \Delta \ell \), angle and sphericity of the wavefront, etc.), or of the system (path length, traveling distance, index of refraction, etc.). We have seen that the interference pattern in a division-of-wavefront interferometer is generated from the \( \mathbf{k}_{\text{mod}} \) portion of the sum of the wavefronts. Division-of-amplitude interferometers use a partially reflecting mirror – a beamsplitter – to divide the wavefront into two beams that then travel down different paths and are recombined by the original or another beamsplitter. The optical interference is generated by the phase difference between the recombined wavefronts.

The beamsplitter reflects part of the wave and transmits the rest. The amplitude reflection
Coefficient for normal incidence is:

\[ r = \frac{n_1 - n_2}{n_1 + n_2} \text{ at normal incidence} \]

which shows that the amplitude is multiplied by a negative number if \( n_1 < n_2 \), meaning that the phase is changed by \( \pi \) radians if reflected at a rare-to-dense interface (second surface has larger \( n \)). The reflection at a dense-to-rare interface exhibits no phase shift. Irradiance depends on the relative phases of the light due to the reflections, which depend on type of reflection. A reflective dielectric coating has an electrically “resistive” surface and a “rare-to-dense” (“external”) reflection induces a phase change \( \Delta \phi = \pi \) radians.

5.5.1 Fizeau Interferometer

The Fizeau interferometer uses a single beamsplitter and may be used to measure the difference in shape between a test optical surface and a reference surface. In the drawing, the physical length difference between the path reflected from of the bottom of the test optic and from the top of the reference surface is \( d \).

Part of the incident beam is reflected from the glass-air interface of the test object. This dense-to-rare reflection has no phase shift. The reflection from the glass reference surface is rare-to-dense, and the phase of the light is changed by \( \pi \) radians. The two waves are recombined when they emerge from the top of the test surface and the irradiance exhibits fringes from the optical path difference. Because the beams traverse the same path in each direction, the optical path difference is doubled.
so an increment in the physical path of \( \frac{\lambda}{2} \) changes the optical path by \( \lambda \) and one fringe cycle is seen. The fringe that appears where the two surfaces are in contact corresponds to a phase difference of \( \pi \) radians and thus is dark.

If the test optic is spherical, then the physical path difference \( d \) may be expressed in terms of the radius of curvature \( R \) and the radial distance \( r \):

\[
R^2 = (R - d)^2 + r^2 \implies r^2 = 2Rd - d^2 \cong 2Rd \text{ for } d \ll R
\]

\[
d \cong \frac{r^2}{2R} \text{ for } d \ll R \quad \text{(Sag Formula)}
\]

\( \text{Sag} \propto \text{square of radial distance} \)

If the interstice between the optical elements in filled with air, and if \( m \) fringes are counted between two points at radial distances \( r_1 \) and \( r_2 \), then the corresponding thickness change is:

\[
m \cdot \frac{\lambda}{2} = \text{OPD} = nd \to d = m \cdot \frac{\lambda}{2} \text{ in air}
\]

This is the same paraboloidal approximation for a sphere that is used in Fresnel diffraction (quadratic dependence on radial distance). If the interstice between optical elements filled with air and \( m \) fringes counted between two points at radial distances \( r_1 \) and \( r_2 \), the corresponding thickness change is:

\[
m \cdot \frac{\lambda}{2} = \text{OPD} = nd \to d = m \cdot \frac{\lambda}{2} \text{ in air}
\]

The test of a spherical surface against an optical “flat” produces “Newton’s Rings”

(a) \hspace{2cm} (b)

*Interference patterns observed using Fizeau interferometer of spherical lens compared to planar test surface: (a) large radius of curvature \( R \); (b) small \( R \).*
5.5.2 Michelson Interferometer

A Michelson interferometer uses mirrors to change the optical path difference of the two beams. The two wavefronts are directed down different paths before recombining to create interference. For example, consider the Michelson interferometer:

\[
\Delta \phi = 2\pi \quad \text{radians, so the phase difference is due only to the optical path difference.}
\]
5.5 DIVISION-OF-AMPLITUDE INTERFEROMETERS

Reflections from beamsplitter in Michelson interferometer: the two components of the normal “output” port have experienced 1 external reflection and 1 internal reflection at the beamsplitter, while light that reflects back out the “input” port have experienced 0 and 2 external reflections. The internal reflection (“dense-to-rare”) adds a phase shift of $\pi$ radians.

If the beamsplitter both reflects and transmits 50% of the irradiance (NOT 50% of the amplitude), then equal portions of the energy are directed toward the mirrors $M_1$ and $M_2$. The amplitude of the electric field of the reflected beam is:

$$E_1 = \sqrt{T_1} = \sqrt{\frac{T_0}{2}} = \sqrt{\frac{E_0^2}{2}}$$

$$= E_0 \cdot \frac{1}{\sqrt{2}} = 0.707 \cdots E_0$$

Because each beam is reflected once and transmitted once before being recombined, the amplitude of each component when recombined is:

$$(E_1)_{out} = (E_2)_{out} = E_0 \cdot \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} E_0$$

Each beam experiences a phase delay proportional to the optical distance traveled in its arm of the interferometer:

$$\Phi_1 = \frac{2\pi}{\lambda_1} \cdot d_1 = \frac{2\pi}{\lambda_1} \cdot n_1d_1 = k \cdot n_1d_1$$

where $d_1$ is the distance traveled by beam #1 and $n_1$ is the refractive index in that path ($n = 1$ in vacuum or air). The beam directed at mirror $M_1$ travels distance $L_1$ from the beamsplitter to the mirror and again on the return, so the total physical path length $d_1 = 2L_1$. Similarly the physical length of the second path is $d_2 = 2L_2$ and the optical path is $n_2d_2 = 2n_2L_2$. After recombination,
the relative phase delay is:

\[ \Delta \Phi = \Phi_1 - \Phi_2 = k (n_1 d_1 - n_2 d_2) \]
\[ = 2k (n_1 L_1 - n_2 L_2) \]
\[ = \frac{4\pi}{\lambda_1} (n_1 L_1 - n_2 L_2) \]

Note that the phase delay is proportional to \( \frac{1}{\lambda_1} \), i.e., longer wavelengths (red light) experience smaller phase differences than shorter wavelengths (blue light) because the ratio of the physical path difference to the wavelength is smaller.

The amplitude at the detector is the sum of the amplitudes:

\[ E(t) = \frac{E_0}{\sqrt{2}} \cos \left( \frac{2\pi}{\lambda_0} (n_1 \cdot 2L_1) - \omega_1 t \right) + \frac{E_0}{\sqrt{2}} \cos \left( \frac{2\pi}{\lambda} (n_2 \cdot 2L_2) - \omega_1 t \right) \]
\[ = \frac{E_0}{\sqrt{2}} \left( \cos \left( \frac{2\pi}{\lambda_0} \cdot n_1 \cdot 2L_1 - \omega_1 t \right) + \cos \left( \frac{2\pi}{\lambda_0} \cdot n_2 \cdot 2L_2 - \omega_1 t \right) \right) \]

which has the form:

\[ \cos [A] + \cos [B] = 2 \cos \left( \frac{A + B}{2} \right) \cdot \cos \left( \frac{A - B}{2} \right) \]

So the amplitude may be rewritten as the product of the average and modulation frequencies:

\[ E(t) = \frac{E_0}{\sqrt{2}} \left( 2 \cos \left( \frac{2\pi (n_1 L_1 + n_2 L_2)}{\lambda_0} - \omega_1 t \right) \cdot \cos \left( \frac{2\pi (n_1 L_1 - n_2 L_2)}{\lambda_0} + 0 \cdot t \right) \right) \]
\[ = \sqrt{2} E_0 \cos \frac{2\pi (n_1 L_1 + n_2 L_2)}{\lambda_0} - \pi \nu_1 t \cdot \cos \frac{2\pi (n_1 L_1 - n_2 L_2)}{\lambda_0} \]

Note that the first term is rapidly oscillating function of time and the second term is the stationary fringe pattern. The time average of squared magnitude is the irradiance:

\[ I = \left< |E(t)|^2 \right> = 2E_0^2 \left\langle \cos^2 \left( \frac{2\pi (n_1 L_1 + n_2 L_2)}{\lambda_0} - \nu_1 t \right) \cdot \cos^2 \left( \frac{2\pi (n_1 L_1 - n_2 L_2)}{\lambda_0} \right) \right\rangle \]
\[ = 2E_0^2 \left\langle \frac{1}{2} + \frac{1}{2} \cos \left( 4\pi \left( \frac{n_1 L_1 + n_2 L_2}{\lambda_0} - \nu_1 t \right) \right) \right\rangle \cdot \left( \frac{1}{2} + \frac{1}{2} \cos \left( 4\pi \frac{(n_1 L_1 - n_2 L_2)}{\lambda_0} \right) \right) \]
\[ = 2E_0^2 \cdot \frac{1}{2} \left( \frac{1}{2} + \frac{1}{2} \cos \left( 4\pi \frac{(n_1 L_1 - n_2 L_2)}{\lambda_0} \right) \right) \]
\[ = \frac{2E_0^2}{4} \left( 1 + \cos \left( \frac{2\pi}{\lambda_0} \cdot 2(n_1 L_1 - n_2 L_2) \right) \right) \]
\[ = I_0 \left( 1 + \cos \left( \frac{2\pi}{\lambda_0} \cdot 2(n_1 L_1 - n_2 L_2) \right) \right) \]
\[ I [L_1, L_2; \lambda_0, n_1, n_2] = I_0 \left( 1 + \cos \left( \frac{2\pi}{\lambda_0} \cdot 2(n_1 L_1 - n_2 L_2) \right) \right) \]

If the two arms are in air, then \( n_1 = n_2 = 1 \) and the irradiance pattern is:

\[ I [L_1, L_2; \lambda_0] = I_0 \left( 1 + \cos \left( \frac{2\pi}{\lambda_0} \cdot 2(L_1 - L_2) \right) \right) \] if \( n_1 = n_2 = 1 \)

The Michelson interferometer with monochromatic plane-wave inputs and “parallel” images of mirrors generates uniform irradiance that depends on ONLY of lengths of arms and on \( \lambda_0 \); not on time or position, so the irradiance is uniform; the Michelson interferometer with monochromatic plane-wave inputs generates a uniform irradiance related to the path difference. If input wavefronts with other shapes are used, then fringes are generated whose period is a function of the local optical
5.5 DIVISION-OF-AMPLITUDE INTERFEROMETERS

path difference, which is directly related to the shape of the wavefronts.

Example: Tilted Mirror

If the incident light is a tilted plane wave, then the resulting pattern is analogous to that from a division-of-wavefront interferometer (Young’s double-slit).

Example: Spherical Waves

If a point source emitting spherical waves is used, then the interferometer may be modeled as two point sources at different distances from the observation plane:

Optical schematic of the Michelson interferometer where the incident wavefront from the source $S$ is spherical. The interferometer may be “unfolded” and modeled as two point sources emitting the same wavefronts but at different distances from the observation plane.

Images of the point source are formed at $S_1$ (due to mirror $M_1$) and at $S_2$ (due to $M_2$). The wavefronts superpose and form interference fringes. The positions of the fringes may be determined from the optical path difference (OPD).

"Unwrapped" schematic of the Michelson interferometer where the wavefronts are spherical. The optical path difference is $2 \cdot (L_1 - L_2) \cos |\theta|$. The OPD is the excess distance that one wavefront has to travel relative to the other before being recombined. For a ray oriented at angle $\theta$ measured relative to the axis of symmetry, the ray reflected
from mirror $M_2$ travels an extra distance:

$$OPD = 2 (L_1 - L_2) \cdot \cos \theta$$

The symmetry about the central axis ensures that the fringes are circular. The phase difference of the waves is the extra number of radians of phase that the wave must travel in that OPD:

$$\Delta \Phi = k \cdot OPD = \frac{2\pi}{\lambda_1} \cdot 2 [L_1 - L_2] \cdot \cos \theta$$

$$= \frac{4\pi \cdot (L_1 - L_2) \cos \theta}{\lambda_1} \text{ [radians]}$$

If the phase difference is a multiple of $2\pi$ radians, then the waves recombine in phase and a maximum of the amplitude results. If the phase difference is an odd multiple of $\pi$ radians, then the waves recombine out of phase and a minimum of the irradiance results due to destructive interference. The locations of constructive interference (irradiance maxima) may be specified by:

$$\Delta \Phi = 2\pi m = \frac{4\pi \cdot (L_1 - L_2) \cos \theta}{\lambda_1} \Rightarrow m\lambda_1 = 2 \cdot (L_1 - L_2) \cos \theta$$

The corresponding angles $\theta$ are specified by:

$$\theta = \cos^{-1} \left( \frac{m\lambda_1}{2 \cdot (L_1 - L_2)} \right)$$

As the physical path difference $L_1 - L_2$ decreases, then $\frac{m\lambda}{2(L_1-L_2)}$ increases and $\theta$ decreases. In other words, if the physical path difference is decreased, the angular size of a circular fringe decreases and the fringes disappear into the center of the pattern. Since a particular fringe occurs at the same angle relative to the optical axis, these are called fringes of equal inclination.

Spherical waves from point source recombined after traveling through arms with different lengths. The curvatures of the two wavefronts are different, which means that the local optical phase difference varies with location in the observation plane.
The OPD is approximately equal to difference in sags between the two spherical wavefronts:

\[ d \cong \frac{r^2}{2R} \quad \text{for } d \ll R \quad \text{(Sag Formula)} \]

\[ d_1 = \frac{r^2}{2R_1} \]

\[ d_2 = \frac{r^2}{2R_2} \]

\[ \text{OPD [m]} = 2(d_1 - d_2) = r^2 \left( \frac{1}{R_1} - \frac{1}{R_2} \right) = \frac{r^2}{R_1 R_2} (R_2 - R_1) \]

\[ \text{OPD [m]} = \frac{r^2}{R_1 R_2} \Delta R \quad \text{[m]} \]

The OPD may be rewritten as a pure number of wavelengths by dividing by \( \lambda_0 \)

\[ \text{OPD [number of wavelengths]} = \frac{r^2}{R_1 R_2} \frac{\Delta R}{\lambda_0} \]

The resulting optical phase difference is obtained by multiplying by \( 2\pi \) radians per wavelength, which is a quadratic function of the radial distance \( r \) from the axis of symmetry:

\[ \Delta \Phi = 2\pi \frac{r^2}{R_1 R_2} \frac{\Delta R}{\lambda_0} = \frac{\pi r^2}{\left( \frac{\lambda_0 R_1 R_2}{2\Delta R} \right)} = \pi \left( \frac{r}{\sqrt{\frac{\lambda_0 R_1 R_2}{2\Delta R}}} \right)^2 = \pi \left( \frac{r}{\alpha} \right)^2 \]

where \( \alpha \) is the radius of the circle measured from the origin where the phase difference is \( \pi \) radians.

The optical phase difference is substituted into the irradiance formula

\[ I [r] \propto 1 + \cos \left[ \pi \left( \frac{r}{\alpha} \right)^2 \right], \text{ where } \alpha = \sqrt{\frac{\lambda_0 R_1 R_2}{2\Delta R}} \quad \text{[m], chirp function} \]

The irradiance changes from a bright fringe where \( r = 0 \) to a dark fringe where \( r = \alpha = \sqrt{\frac{\lambda_0 R_1 R_2}{2\Delta R}} \).

Bright fringes are located at values of \( r \) where the phase difference is an integer multiple of \( 2\pi \) radians, i.e., for \( r = \sqrt{2\alpha}, r = \sqrt{4\alpha} = 2\alpha, r = \sqrt{6\alpha}, \text{ etc.} \) If the projected separation between the two mirrors in the Michelson is \( 0 \), there is no phase difference and so no fringes increased, \( \Delta R \) increases and \( \alpha \) decreases. In words, increasing the mirror separation decreases the radial distance of the phase change of \( \pi \) radians; the fringe spacing decreases.
CHAPTER 5 OPTICAL INTERFERENCE AND COHERENCE

The observed pattern is a “chirp” function \( \cos \left( \pi \left( \frac{x}{2} \right)^2 \right) \), which means that the fringe spacing at the observation plane decreases with increasing radial distance from center. If one of the wavefronts is not spherical, then the fringes are not circular. In fact, the shape of fringes is determined by the shape of wavefront. If one mirror is tilted relative to the other, then the output beams travel in different directions when recombined. The fringes thus obtained are straight and have a constant spacing just like those from a Young’s double-slit experiment.

Applications of Michelson Interferometers

1. Measure refractive index \( n \): Insert a plane-parallel plate of known thickness \( t \) and unknown index \( n \) into one arm of a Michelson interferometer illuminated with light of wavelength \( \lambda \). Count the number of fringes due to the plate:

   \[ OPD = (n - 1) \cdot t = \Delta m \cdot \lambda_1 \]

2. Measure the wavelength of an unknown source or \( \Delta \lambda \) between two spectral lines of a single source. Count the number of fringes that pass a single point as one mirror is moved a known distance.

3. Measure lengths of objects: The standard \( m \) now is defined in terms of a particular spectral line:

   \[ 1 \text{ m} = 1,650,763.73 \text{ wavelengths of emission at } \lambda = 605.8 \text{ nm} \]
   
   from Kr-86 measured in vacuum
This standard has the advantage of being transportable.

4. Measure deflections of objects: The optical phase difference will be significant for very small physical path differences. Place one mirror on an object and count fringes to measure the deflection.

5. Measure the velocity of light (Michelson-Morley experiment)

**Twyman-Green Interferometer**

This is a version of the Michelson interferometer with a point source and a collimating lens that ensures that the incoming light is composed of plane waves

**Fourier-Transform Spectrometry (FTS)**


The Michelson interferometer is the basis for an elegant (though indirect) technique for measuring the wavelength spectrum of a source. Its primary advantage is that it does not “waste” (i.e., discard) light, but the data collection takes time and the spectrum must be reconstructed indirectly.

One traditional method for measuring a spectrum is to interpose a set of bandpass filters between the object and the measurement system, so that the output is a set of images of the source in different passbands. The spectrum of the object is constructed from the different images.

Consider a Michelson interferometer that is configured so that the two arms are exactly equal in length so that the OPD = 0. In theory, the waves emitted by the source are split and reunited exactly in phase and appear as though the interferometer is not present. Now the path length of one arm of the interferometer is changed slowly at a uniform rate, which means that the waves from the source are combined with a slowly increasing OPD.

Consider the case if the source emits only a single wavelength (or, equivalently, a single optical frequency). At the time when OPD = 0, the two waves reunite in phase. As the OPD is increased, the phase difference between the two waves changes from $\Delta \phi = 0$ to $\Delta \phi = \pi$ to $\Delta \phi = 2\pi$, and the irradiance of the combined signals changes from $I = I_{\text{max}}$ at $\Delta \phi = 0$ to $I = 0$ at $\Delta \phi = \pi$ and back to $I = I_{\text{max}}$ at $\Delta \phi = 2\pi$. The last measurement is again a maximum because the source emits only a single wavelength.

If the source emits two wavelengths, the interfering beams at both frequencies are in phase if OPD = 0, so that the resulting irradiance $I[\text{OPD} = 0] = I_{\text{max}}$. If OPD is increased from 0, the phase of the shorter wavelength will change more quickly with distance so that the phase difference at this wavelength will be $\Delta \phi_{\text{shorter}} \lambda = \pi$ when $\Delta \phi_{\text{longer}} \lambda < \pi$. The two beams at the shorter wavelength cancel exactly, but those at the longer wavelength do not quite do so. As the OPD increases further, eventually $\Delta \phi_{\text{shorter}} \lambda = 2\pi$ when $\Delta \phi_{\text{longer}} \lambda < 2\pi$, so the resulting irradiance $I < I_{\text{max}}$.

Now consider the case if the source is completely incoherent, which means that there is the optical phase is perfectly correlated if the beams are reunited after traveling exactly the same path, but the phase is completely uncorrelated for any OPD $\neq 0$. In other words, the correlation of the optical phase of completely incoherent light is a Dirac delta function. In the Michelson, the output irradiance is a maximum value if OPD = 0 and is

5.5.3 **Mach-Zehnder Interferometer**

The M-Z interferometer is very similar to the Michelson, except that a second beamsplitter is used to recombine the beams so that the light does not traverse the same path twice. Therefore there is no factor of 2 in the OPD for the M-Z. Mach-Zehnder interferometers often are used to measure the refractive index of liquids or gases. The container $C_1$ (or $C_2$) is filled with a gas while examining
the fringe pattern. The length of the gas path in the cell is \( \delta \). As the container fills, the refractive index \( n \) increases and so does the optical path length in that arm. The optical path difference is:

\[
OPD = (n - 1) \cdot \delta
\]

The fringes move and each new cycle of the pattern corresponds to an increase in the OPD of \( \lambda \). After \( m \) fringes are counted, the index of refraction is found via:

\[
(n_1 - 1) \cdot \delta \cdot \frac{2\pi}{\lambda_1} = m \cdot 2\pi \rightarrow n_1 = 1 + \frac{m\lambda_1}{\delta}
\]

Optical schematic of the Mach-Zehnder interferometer, where the amplitudes divided by the beamsplitter travel two different paths and are recombined by a second beamsplitter.

5.5.4 Sagnac Interferometer

The Sagnac interferometer is a single-beamsplitter version of the M-Z; the output beamsplitter is exchanged for a mirror which is reversed to create a loop path. Light travels around the loop in both directions so that the optical path difference is zero for a stable configuration. However, if the interferometer (including the illuminator) is rotated as on a turntable, then light in one path will experience a Doppler shift with increasing frequency (blue-shift), while light in the reverse direction will experience a red shift. The phase of the two beams will change in proportion to the frequency shift, and the superposed light will exhibit a sinusoidal variation in the detected signal over time:

\[
\cos [\omega_1 t] + \cos [\omega_2 t] = 2 \cos [\omega_{\text{avg}} t] \cdot \cos [\omega_{\text{mod}} t]
\]

The slower-varying modulation frequency is detectable and linearly proportional to the rotation rate. This device may be used as a gyroscope with no moving parts, and in fact may be constructed from a single optical fiber that forms a loop with counterrotating beams.
5.6 Interference by Multiple Reflections

Division-of-amplitude interference also may occur within a medium by multiple reflections between two surfaces. This is the basis for thin-film interference and the Fabry-Perot interferometer. Thin-film interference is the basis for optical coating design, e.g., for "anti-reflection" ("AR") coatings applied to lenses and for coatings that exhibit near total reflection. The coating is designed so that the multiple reflections within the coating either "cancel out" by destructive interference (in an AR coat) or constructively interfere in phase (in a mirror coating). Fabry-Perot interferometers are used to measure the spectrum of a source or to select a narrow band of wavelengths from a source, as to decrease the bandwidth and extend the coherence length of a laser used to make optical holograms.

5.6.1 Thin Films

Parameters in thin-film interference. The ray with amplitude $a$ (not irradiance) is incident from air onto the medium with index $n$ at angle $\theta$. The Fresnel equations determine the percentage reflected and transmitted at each surface. The front-surface reflection exhibits a phase shift of $\pi$ radians, while there is no phase shift at the "dense-to-rare" interface.
Simple geometry determines the optical path difference (OPD) between the first two reflected rays is:

$$\text{OPD} = 2 \cdot nd \cos [\theta']$$

Since the first ray is reflected at a “rare-to-dense” interface, it includes an intrinsic phase difference of \(\pi\) radians that is not present in the second ray that reflected from the rear surface (“dense-to-rare”).

The total optical phase difference is:

$$\Delta \phi = \frac{2\pi}{\lambda_0} \cdot \text{OPD} + \pi = \frac{4\pi}{\lambda_0} \cdot nd \cos [\theta'] + \pi$$

If \(\Delta \phi = 2\pi \ell\), where \(\ell\) is any integer, the two beams superpose out of phase (due to the front-surface reflection) to produce an interference maximum, which produces high-reflectance coating. The thickness \(d\) is selected for a reflectance coating tuned to a particular wavelength \(\lambda_0\):

$$\Delta \phi = 2\pi \ell = \frac{4\pi}{\lambda_0} \cdot nd \cos [\theta'] - \pi$$

$$\Rightarrow \ell = \frac{2}{\lambda_0} \cdot nd \cos [\theta'] - \frac{1}{2}$$

$$\Rightarrow d = \left( \ell + \frac{1}{2} \right) \cdot \frac{\lambda_0}{2n \cos [\theta']} \text{ for high-reflectance coating}$$

The complementary condition for minimizing the reflection is:

$$\Delta \phi = \pi (2\ell + 1) = \frac{4\pi}{\lambda_0} \cdot nd \cos [\theta'] + \pi$$

$$\Rightarrow d = \frac{\ell \lambda_0}{2n \cos [\theta']} \text{ for antireflectance (AR) coating}$$

All reflections of “transmitted” rays are “dense to rare” with no intrinsic phase change. The OPD is the same as for the reflected rays, so the optical phase difference is:

$$\Delta \phi = \frac{2\pi}{\lambda_0} \cdot \text{OPD} = \frac{4\pi}{\lambda_0} \cdot (nd \cos [\theta'])$$

$$\Rightarrow d = \frac{(2\ell + 1) \lambda_0}{4n \cos [\theta']}$$

so the irradiance of a “transmitted” fringe is a maximum if \(\Delta \phi = 2\pi \ell\) and a minimum if \(\Delta \phi = \pi (2\ell + 1)\). In other words, the reflected and transmissive fringes are complementary; the “bright” and “dark” bands interchange.

Visibility of Thin-Film Fringes

The visibility of the fringes depends on the relative reflectance and transmittance coefficients at the surface, which may be calculated is an elegant way put forth by Sir George Stokes. First consider part (a) of the figure. The reflectance and transmission coefficients are determined by the Fresnel equations for the particular polarization used. The amplitude of the electric field in the incident ray is \(a\) and of the reflected ray is \(ar < 0\) because of the “rare-to-dense” reflection. The amplitude of the electric field of the transmitted ray is \(at > 0\), where \(t\) is the transmission coefficient at a “rare-to-dense” interface. Time has been reversed In part (b) so that there are two incident rays with amplitudes \(ar\) and \(at\). The “output” ray in the time-reversed case that had amplitude \(a\) in part (a) has amplitude \(ar^2 + att\), where \(t'\) is the transmission coefficient in the “dense-to-rare” interaction. The incident wave with amplitude \(ar\) transmits to produce an amplitude \(art\), while the incident amplitude \(at\) produces a reflection with amplitude \(atr'\), where \(r'\) is the “dense-to-rare”
5.6 INTERFERENCE BY MULTIPLE REFLECTIONS

reflection coefficient. Since all four terms must be the same, this establishes two conditions:

\[ a = ar^2 + att' \implies r^2 + tt' = 1 \]
\[ atr' + art = 0 \implies r' = -r \text{ (which we knew already)} \]

Reflectance and transmission coefficients at the surface of a thin film: (a) the incident ray has amplitude \( r \), the amplitude of the reflected ray is \( ar < 0 \), and the transmitted ray is \( at \); (b) the “time-reversed” situation, where the rays with amplitudes \( ar \) and \( at \) are incident on the surface.

We can apply these results to determine the amplitudes of multiply reflected rays. All rays in the following system arise from the single incident ray with field amplitude \( a \), so all rays are mutually coherent. Only the first reflected ray includes a phase shift of \( \pi \) radians, so the negative sign has been included explicitly. We already know that the optical phase difference between adjacent pairs of rays is:

\[ \Delta \phi = \frac{2\pi}{\lambda_0} \cdot 2nd \cos[\phi'] \]
Now consider the amplitude of the sum of all reflected rays:

\[ a_r = -ar + atr' \exp [+i \cdot \Delta \phi] + atr' (r^2) \exp [+i \cdot 2 \cdot \Delta \phi] + atr' (r^4) \exp [+i \cdot 3 \cdot \Delta \phi] + \cdots \]

\[ = -ar + atrt' \exp [+i \cdot \Delta \phi] (1 + r^2 \exp [+i \cdot \Delta \phi] + r^4 \exp [+i \cdot 2 \cdot \Delta \phi] + \cdots) \]

\[ = -ar + atrt' \exp [+i \cdot \Delta \phi] \sum_{n=0}^{\infty} (r^2 e^{+i \cdot \Delta \phi})^n \]

Now use the identity:

\[ \sum_{n=0}^{\infty} t^n = \frac{1}{1 - t} \text{ where } |t| < 1 \]
\[ a_r = -ar + artt' e^{+i \Delta \phi} \left( \frac{1}{1 - r^2 e^{+i \Delta \phi}} \right) \]

\[ = -ar \left( 1 - \frac{tt' e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right) \]

\[ = -ar \left( 1 - \frac{r^2 e^{+i \Delta \phi} - tt' e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right) \]

\[ = -ar \left( 1 - \frac{(r^2 + tt') e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right) \]

\[ a_r = -ar \left( \frac{1 - e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right) \]

The ratio of the irradiances is:

\[ \frac{I_r}{I_{in}} = \frac{|a_r|^2}{|a|^2} = r^2 \left| \frac{1 - e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right|^2 \]

\[ = r^2 \left( \frac{1 - e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right) \left( \frac{1 - e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right)^* \]

\[ = r^2 \left( \frac{1 - e^{+i \Delta \phi}}{1 - r^2 e^{+i \Delta \phi}} \right) \left( \frac{1 - e^{-i \Delta \phi}}{1 - r^2 e^{-i \Delta \phi}} \right) \]

\[ = r^2 \left( \frac{2 - 2 \cos [\Delta \phi]}{1 + r^4 - 2r^2 \cos [\Delta \phi]} \right) \]

\[ = r^2 \left( \frac{2 (1 - \cos [\Delta \phi])}{1 + r^4 - 2r^2 \cos [\Delta \phi]} \right) \]

Now use the identities:

\[ 1 - \cos [\theta] = 2 \sin^2 \left( \frac{\theta}{2} \right) \]

\[ 1 + r^4 = (1 - r^2)^2 + 2r^2 \]

\[ \frac{I_r}{I_{in}} = r^2 \left[ \frac{4 \sin^2 \left( \frac{\Delta \phi}{2} \right)}{(1 - r^2)^2 + 2r^2 - 2r^2 \cos [\Delta \phi]} \right] \]

\[ = \frac{4r^2 \sin^2 \left( \frac{\Delta \phi}{2} \right)}{(1 - r^2)^2 + 2r^2 \left( 1 - \cos [\Delta \phi] \right)} \]

\[ = \frac{4r^2 \sin^2 \left( \frac{\Delta \phi}{2} \right)}{(1 - r^2)^2 + 4r^2 \sin^2 \left( \frac{\Delta \phi}{2} \right)} \]

\[ = \frac{\left( \frac{2r}{\pi/r^2} \sin \left( \frac{\Delta \phi}{2} \right) \right)^2}{1 + \left( \frac{2r}{\pi/r^2} \sin \left( \frac{\Delta \phi}{2} \right) \right)^2} \]
We also know that

\[
\frac{I_t}{I_{in}} = 1 - \frac{I_r}{I_{in}} \\
= 1 - \frac{\left(\frac{2r}{(1-r^2)} \sin \left(\frac{\Delta \phi}{2}\right)\right)^2}{1 + \left(\frac{2r}{(1-r^2)} \sin \left(\frac{\Delta \phi}{2}\right)\right)^2} \\
= \frac{1}{1 + \left(\frac{2r}{(1-r^2)} \sin \left(\frac{\Delta \phi}{2}\right)\right)^2}
\]
5.6 INTERFERENCE BY MULTIPLE REFLECTIONS

Example: Small Reflectance Coefficient \( r \)

Now consider the case of \( r^2 \ll 0 \), e.g., \( r = 0.2 \) \( \implies r^2 = 0.04 \)

\[
\left( \frac{2r}{1 - r^2} \right)^2 = \left( \frac{0.4}{0.96} \right)^2 = 0.1736 \ll 1
\]

\[
\implies \frac{I_r}{I_{in}} = \frac{0.1736 \sin^2 \left[ \frac{\Delta \phi}{2} \right]}{1 + 0.1736 \sin^2 \left[ \frac{\Delta \phi}{2r} \right]}
\]

\[
\implies \frac{I_t}{I_{in}} = \frac{1}{1 + 0.1736 \sin^2 \left[ \frac{\Delta \phi}{2r} \right]}
\]

\( \frac{I_t}{I_{in}} \) (black) and \( \frac{I_r}{I_{in}} \) (red) for \( r = 0.2 \)
Example: Large Reflectance Coefficient $r$

If the surfaces have reflective coatings with $r = 0.95 \implies r^2 = 0.95^2 \approx 0.9$

$$
\left( \frac{2r}{1 - r^2} \right)^2 = \left( \frac{1.9}{1 - 0.95^2} \right)^2 \approx 380 >> 1
$$

$$
\implies \frac{I_r}{I_{in}} = \frac{380 \sin^2 \left( \frac{\Delta \phi}{2} \right)}{1 + 380 \sin^2 \left( \frac{\Delta \phi}{2} \right)}
$$

$$
\implies \frac{I_t}{I_{in}} = \frac{1}{1 + 380 \sin^2 \left( \frac{\Delta \phi}{2} \right)}
$$

Again, the reflective and transmissive patterns are complementary. As $r^2$ increases, the transmitted maxima become “sharper” while the reflected maxima become “broader.” This observation is the basis for the Fabry-Perot etalon, which is a division-of-amplitude instrument that acts by multiple reflections from two highly (though not totally) reflecting surfaces many times before recombining.
5.6 INTERFERENCE BY MULTIPLE REFLECTIONS

5.6.2 Fabry-Perot Interferometer

The Fabry-Perot interferometer uses two highly-reflecting surfaces that amplify the internal reflections and “narrow” the fringes in transmitted light. The width of the transmissive peaks are characterized by the “finesse,” which is defined:

$$F = \frac{\text{period of } I_n}{\text{FWHM}} = \frac{2\pi}{W}$$

where “FWHM” is the “full width at half maximum.” As $r$ increases, the widths of the peaks decrease and the finesse increases. The FWHM is determined from the increment $\varepsilon$ of the optical
phase difference that satisfies the conditions:

\[
\frac{I_t[\Delta \phi = 2\pi \ell]}{I_{in}} = 1
\]

\[
\frac{I_t[\Delta \phi = 2\pi (\ell + \varepsilon)]}{I_{in}} = \frac{1}{2}
\]

\[\implies F = \frac{2\pi}{2\varepsilon}\]

\[F = \frac{2\pi}{2\varepsilon} = \frac{\pi}{\varepsilon} = \frac{\pi r}{1 - r^2}\]

If the source included a second wavelength, the spacing between the fringe maxima would be scaled and the fringes from the two wavelengths with the same order $\ell$ would be positioned at different spots. The two wavelengths can be distinguished if $F$ is large. For quasimonochromatic light so that $\Delta \lambda << \lambda_0$, the maximum of order $\ell$ is spread over the

**Optical Path Difference:**

\[OPD = 2nd \cos[\theta_i] = m\lambda\] for constructive interference

\[\rightarrow 2d \cos[\theta_i] \text{ in air } (n = 1)\]

- All rays that are incident at same angle $\theta_i$ arrive at $P$

- Add collimating lens on input side (distance from lens to interferometer = $f$)
5.6 INTERFERENCE BY MULTIPLE REFLECTIONS

Effect of varying amplitude reflectance coefficient $r$ on fringes from a Fabry-Perot interferometer:
(a) $r \gg 0$ so that the maxima are “wide;” (b) $r \ll 1$ producing “narrow” maxima.

Resolving power of Fabry-Perot etalon: Two fringes from different wavelengths will overlap significantly if the FWHM $W_\frac{1}{2}$ is large. The resolving power $R \equiv \frac{\Delta \lambda}{\Delta \phi} = \frac{\lambda}{W_\frac{1}{2}} = mF$. 
Chapter 6

Optical Diffraction and Imaging

In geometrical (ray) optics, light is assumed to propagate in straight lines from the source (rectilinear propagation), and cannot penetrate into the shadow of an opaque obstruction. However, Grimaldi observed in the 1600s that light is able to “bend” into the shadow region and named this phenomenon diffraction (from Latin, dis + frangere, “to break apart”). This spreading of a bundle of rays affects the sharpness of shadows cast by opaque objects by making edges become “fuzzy.” A second observation of diffraction is the comparison of diameters $D$ of the light beams through an aperture of diameter $d$. In the regime where geometrical explanations apply, the observed diameter $D \propto d$, if $d$ is made sufficiently small, then $D \propto d^{-1}$.

\begin{itemize}
  \item[(a)] If the diameter $d_1$ of the aperture is “large,” the diameter of the projection is $D_1 \propto d_1$ (larger hole $\Rightarrow$ larger “image”);
  \item[(b)] If the aperture diameter is decreased, eventually the diameter of the projection is $D_2 \propto (d_2)^{-1}$ (smaller hole $\Rightarrow$ larger “image”).
\end{itemize}

Diffraction is really a generalization of interference that differs only in the number of sources. Like interference, the observed effects of diffraction also depend on the state of coherence of the

\begin{itemize}
  \item Coherent light:
    \begin{itemize}
      \item The intensity pattern is a set of circular fringes.
      \item The angular spread of the diffraction pattern is $\Delta \theta \propto \frac{1}{d}$.
    \end{itemize}
  \item Incoherent light:
    \begin{itemize}
      \item The intensity pattern is a set of concentric rings.
      \item The angular spread of the diffraction pattern is $\Delta \theta \propto \frac{1}{\sqrt{d}}$.
    \end{itemize}
\end{itemize}
field. In both interference and diffraction, light from a source is combined after traveling different paths (few paths for interference, many for diffraction). The amplitude of the combination depends on the relative phase of the component beams; the amplitude is increased if the beams are in phase (or nearly so) and decreases if the vector sum is small. In other words, the wave character of light creates stationary regions of constructive and destructive interference that may be observed as bright and dark regions.

In the simplest case of two sources of infinitesimal size, the superposition wave may be determined by summing spherical wave contributions from the sources. If the number of sources generating beams to be added, the effect is considered to be interference. In diffraction, the sources are large compared to the wavelength $\lambda_0$, and light with the same wavelength emerging from the large number of “subsources” at different locations on the source will interfere constructively or destructively depending on the relative phases of the light waves. The spherical-wave contributions from the large number of subsources are summed (by integrating over the area of the aperture) to determine the total electric field. The superposition electric field vector (magnitude and phase) is the vector sum of the fields due to these spherical-wave subsources.

Interference is often considered first in courses in optics because it is arguably easier to analyze and understand the interactions of small numbers of sources. However, this tactic often shortchanges the study of diffraction, even though it is more applicable to imaging. You already have developed the necessary mathematical tools for the analysis of diffraction in your study of Fourier methods, so we can transfer directly into this study; there is no need to pass “Go” beforehand.

The mathematical model for diffraction is straightforward to develop, though computations may be tedious. It is based on Huygen’s principle:

### 6.1 Huygens’ Principle

[\textit{PON, §1}]

In 1678, Christiaan Huygens theorized a model for light propagation that claimed that each point on a propagating wavefront (regardless of the “shape” of the wavefront) acted as the source of a new spherical wave. The sum of these secondary spherical “wavelets” produced the subsequent wavefronts. Huygens’ principle had the glaring problem that these secondary spherical wavefronts propagated “backwards” as well as forwards, but this problem was addressed by Fresnel and Kirchhoff in the 19th century by introducing an “obliquity” factor that attenuates the parts of the wavefront that propagate “off axis.” With this correction, the Huygens’ principle provides a very useful model for light propagation that naturally leads to expressions for “diffracted” light.

Consider the mathematical expression for a spherical wave emitted by a source located at the origin of coordinates; energy conservation requires that the energy density of the electric field of a spherical wave decrease as the square of the distance from the source. Correspondingly, the electric field decreases as the distance from the source. The electric field observed at location $[x_1, y_1, z_1]$ due to a spherical wave emitted from the origin is:

$$s [x_1, y_1, z_1, t] = \frac{E_0}{\sqrt{x_1^2 + y_1^2 + z_1^2}} \cos [k_x \cdot x + k_y \cdot y + k_z \cdot z - \omega_0 t]$$

where $\sqrt{x_1^2 + y_1^2 + z_1^2} = r$ is the distance from the source. This observation that light from a point source generates a spherical wave is the first step towards Huygen’s principle, which states that every point on a wavefront may be modeled as a “secondary source” of spherical waves. The summation of the waves from the secondary sources (sometimes called “wavelets”) produces a new wavefront that is “farther downstream” in the optical path.

In the more general case of a spherical wave emitted from a source located at coordinates $[x_0, y_0, z_0]$ and observed at $[x_1, y_1, z_1]$ has the form:

$$s [\mathbf{r}, t] = \frac{E_0}{|\mathbf{r}|} \cos [\mathbf{k}_0 \cdot \mathbf{r} - \omega_0 t] \to \frac{E_0}{|\mathbf{r}|} \exp [\mathbf{k}_0 \cdot \mathbf{r} - \omega_0 t]$$
where 

\[ |\mathbf{r}| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2} \]

and \( |\mathbf{k}| = \frac{2\pi}{\lambda_0} \)

If the propagation distance \( |\mathbf{r}| \) is large, then the spherical wave may be approximated accurately as a paraboloidal wave, and for VERY large \( |\mathbf{r}| \) the sphere becomes a plane wave. The region where the first approximation is acceptable defines Fresnel diffraction, while the more distant region where the second approximation is valid is the Fraunhofer diffraction region.

### 6.2 Diffraction Integrals

Consider the electric field emitted from a point source located at \([x_0, y_0, z_0 = 0]\). The wave propagates in all directions. The electric field of that wave is observed on an observation plane centered about coordinate \(z_1\). The location in the observation plane is described by the two coordinates \([x_1, y_1]\). The electric field at \([x_1, y_1]\) at this distance \(z_1\) from a source located in the plane \([x_0, y_0]\) centered about \(z = z_0 = 0\) is:

\[
E[x_1, y_1; z_1, x_0, y_0, 0] = \frac{E_0}{|\mathbf{r}|} \cos |\mathbf{k}| \cdot \mathbf{r} - \omega_0 t, \text{ where } |\mathbf{r}| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}
\]

Though this may LOOK complicated, it is just an expression of the electric field propagated as a spherical wave from a source in one plane to the observation point in another plane; the amplitude decreases as the reciprocal of the distance and the phase is proportional to the distance and time. Diffraction calculations based on the superposition of spherical waves is Rayleigh-Sommerfeld diffraction.

Now, observe the electric field at that same location \([x_1, y_1, z_1]\) that is generated from many point sources located in the \(x - y\) plane located at \(z = 0\). The summation of the fields is computed as an integral of the electric fields due to each point source. The integral is over the area of the source plane. If all sources emit the same amplitude \(E_0\), then the integral is simplified somewhat:

\[
E_{\text{total}}[x_1, y_1; z_1] = \alpha^2 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} E[x_1, y_1; z_1, x_0, y_0, 0] \; dx_0 \; dy_0
\]

\[
= \int_{\text{aperture}} \frac{E_0}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}} \times \cos \left[ \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} - \omega_0 t \right] \; dx_0 \; dy_0
\]

where \(\alpha\) is a constant that is evaluated from physical considerations. It can be shown that \(\alpha = (\sqrt{\lambda_0})^{-1} \Rightarrow \alpha^2 = \frac{1}{i\lambda_0}\), which includes the constant phase of \(i^{-1} = \exp(-i\pi/2)\). This integral may be recast into a different form by defining the shape of the aperture to be a 2-D function \(f[x_0, y_0]\) in the source plane:

\[
E_{\text{total}}[x_1, y_1; z_1] = \frac{E_0}{i\lambda_0} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x_0, y_0] \left. \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} - \omega_0 t \right] \; dx_0 \; dy_0
\]

This expression is the diffraction integral. Again, this expression may LOOK complicated, but really represents just the summation of the electric fields due to all point sources. Virtually the entire study of optical diffraction is the application of various schemes to simplify and apply this equation. We will approximate the diffraction integral for two distinct cases, as shown in the figure:
CHAPTER 6 OPTICAL DIFFRACTION AND IMAGING

1. Observation plane $z_1$ is located “close to” the source plane $z_0 = 0$; this is “near-field” or “Fresnel diffraction” (the $s$ in “Fresnel” is silent – the name is pronounced Frey’ nel).

2. Observation plane $z_1$ located a large distance the source plane at $z_0 = 0$ so that $z_1 \gg \infty$. This is called “far-field,” or “Fraunhofer diffraction,” which is particularly interesting (and easier to compute results) because the diffraction integral is proportional to the Fourier transform of the object distribution (the shape of the planar aperture).

The diffraction regions for spherical waves emitted by a point source. Rayleigh-Sommerfeld diffraction is based on the spherical waves emitted by the source. Fresnel diffraction is an approximation based on the assumption that the wavefronts are parabolic and with unit amplitude for distances from some limit outward to $\infty$. The “width” of the quadratic phase is indicated by $\alpha_n$; this is the off-axis distance from the origin where the phase change is $\pi$ radians. Fraunhofer diffraction assumes that the spherical wave has traveled a large distance and the wavefronts may be approximated by planes.

Also note that the diffraction integral depends on the wavelength $\lambda_0$, which means that it strictly applies ONLY to monochromatic light and thus cannot be used to describe the effect in temporally incoherent.

6.3 Fresnel Diffraction

Consider the first case of the diffraction integral where the observation plane is near to the source plane, where the concept of near must be defined. Note that the distance $|\mathbf{r}|$ appears twice in the expression for the electric field due to a point source – once in the denominator and once in the phase of the cosine. The first term affects the size (magnitude) of the electric field, and the scalar product of the second with the wavevector $\mathbf{k}$ is computed to determine the rapidly changing phase angle of the sinusoid. The optical phase changes very quickly with time (because $\omega_0$ is very large, $\omega_0 \cong 10^{15}$ radians/second) and with distance (because $|\mathbf{k}|$ is very small, $|\mathbf{k}| \cong 10^{-7}$ m), so the phase difference
of light observed at one point in the observation plane but generated from two points in the source plane may differ by MANY radians. Simply put, small changes in the propagation distance $|r|$ are very significant in the computation of the phase, but much less so when computing the amplitude of the electric field. Therefore, the distance may be approximated more crudely in the denominator than in the phase.

Now consider the approximation of the distance $|r|$. The exact expression is:

$$|r| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}$$

$$= \sqrt{z_1^2 \cdot \left(1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right)}$$

$$= z_1 \cdot \sqrt{1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}}$$

$$= z_1 \cdot \left(1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right)^\frac{1}{2}$$

This expression may be expanded into a power series by applying the binomial theorem. Recall that the general binomial expansion is:

$$(1 + \alpha)^n = 1 + n\alpha + \frac{n \cdot (n-1)}{2!} \alpha^2 + \frac{n \cdot (n-1) \cdot (n-2)}{3!} \alpha^3 + \cdots + \frac{n!}{(n-r)!r!} \alpha^r + \cdots$$

which converges to the correct value if $\alpha^2 < 1$. In the current example of $n = \frac{1}{2}$ (square root), the series is:

$$(1 + \alpha)^{\frac{1}{2}} = 1 + \frac{\alpha}{2} - \frac{1}{8} \alpha^2 + \frac{1}{16} \alpha^3 - \cdots$$

which may be applied to find an approximation for the distance $|r|$:

$$|r| = z_1 \cdot \left(1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right)^\frac{1}{2}$$

$$= z_1 \left(1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2} - \frac{1}{8} \left(\frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right)^2 + \cdots\right)$$

If $z_1$ is "sufficiently" large, terms of second and larger order are assumed to be sufficiently close to zero to be ignored. In other words, we assume that:

$$(x_1 - x_0)^2 + (y_1 - y_0)^2 << z_1^2$$

in the third term. The resulting approximation for the propagation distance is:

$$|r| \approx z_1 \left(1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right) = z_1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{2z_1}$$

In words, the propagation distance from the source to the observation point is the sum of the "longitudinal" distance $z_1$ (down the optical axis) and a quadratic "correction" arising from the "off-axis" distance of the observation point.
The expression for the electric field may be simplified further by recasting into complex notation:

\[ E[x_1, y_1; z_1, x_0, y_0, 0] \approx E_0 \frac{1}{i \lambda z_1} \text{Re} \left\{ \exp \left[ +\frac{2\pi i}{\lambda_0} \left( z_1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{2z_1} \right) \right] \right\} \]

\[ = \frac{E_0}{i \lambda z_1} \text{Re} \left\{ \exp \left[ +\frac{2\pi i}{\lambda_0} z_1 \right] \cdot \exp [-2\pi i n t] \cdot \exp \left[ +\frac{i \pi}{\lambda z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \right\} \]

The phase of this approximation of the spherical wave includes a constant phase \( \frac{2\pi z_1}{\lambda_0} \), a phase term that is a linear function of time, and the last phase term that is proportional to the square of the distance off-axis from the source point at \([x_0, y_0]\) to the observation point \([x_1, y_1]\). This last spatial quadratic-phase term shows that the approximate wavefront emitted by a point source is a paraboloid instead of the sphere resulting from Huygens' principle.

Note that one part of the assumption of Fresnel diffraction is clearly unreasonable: that the wavefront is assumed to have constant magnitude regardless of the off-axis location \([x_1, y_1]\) where the field is measured. In other words, the paraboloidal wave in Fresnel diffraction has the same "brightness" regardless of where it is measured in the observation plane.

For larger values of \(z_1\) (observation plane still farther from the source), the radius of curvature of the approximate paraboloidal waves increases, so the change in phase measured for nearby points in the observation plane decreases. As \(z_1\) increases still further, the paraboloid approaches a plane wave.

This electric field is substituted into the diffraction integral to obtain the approximate expression in the near field:

\[ E_{total}[x_1, y_1; z_1] = \int_{-\infty}^{+\infty} E[x_1, y_1; z_1; x_0, y_0, 0] \ dx_0 \ dy_0 \]

\[ \approx \frac{1}{i \lambda z_1} \exp \left[ \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ +\frac{i \pi}{\lambda z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \ dx_0 \ dy_0 \]

Again, this LOOKS complicated, but really is just a collection of the few parts that we have considered already. In words, the integral says that the electric field downstream but near to the source function is the summation of paraboloidal fields from the individual sources. The paraboloidal approximation significantly simplifies the computation of the diffracted light.

### 6.3.1 Fresnel Diffraction as a Convolution Integral

The Fresnel diffraction integral

\[ g[x_1, y_1] = \frac{1}{i \lambda z_1} \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ +\frac{i \pi}{\lambda z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \ dx_0 \ dy_0 \]

may be rewritten by defining the exponential term as a function \(h\) that depends on the four variables in a particular way:

\[ h[x_1 - x_0, y_1 - y_0] = \frac{1}{i \lambda z_1} \exp \left[ +\frac{i \pi}{\lambda z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \]

In other words, the Fresnel diffraction integral may be written as:

\[ g[x_1, y_1] = \int_{-\infty}^{+\infty} f[x_0, y_0] \ h[x_1 - x_0, y_1 - y_0] \ dx_0 \ dy_0 \]

This is a convolution integral of the form that abounds in all areas of physical science, and particularly in imaging. The function \(h\) is the impulse response of the integral operator, and often is called the point-spread function in imaging and particularly in optics. It has other names in other disci-
6.3 Fresnel Diffraction

The integral operator often is given a shorthand notation, such as the asterisk “∗.” The variables of integration also often are renamed as dummy variables, such as α, β:

\[
g [x, y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[\alpha, \beta] h[x - \alpha, y - \beta] \ d\alpha \ d\beta
\]

\[
\equiv f [x, y] * h [x, y]
\]

The impulse response for Fresnel diffraction is:

\[
h [x, y] = \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i \frac{\pi (x^2 + y^2)}{\lambda_0 z_1} \right]
\]

which is the product of a constant and a quadratic-phase “chirp” function, whose real and imaginary parts are sinusoids with varying spatial frequency that depends on the propagation distance \(z_1\). The parameters of the chirp often are combined into the chirp rate \(\alpha_0 \equiv \sqrt{\lambda_0 z_1}\) that has dimensions of length:

\[
h [x, y] = \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i \frac{\pi x^2}{\alpha_0^2} \right]
\]

The chirp rate is the radial distance \(r = \sqrt{x^2 + y^2} = \alpha_0\) measured from the origin of coordinates over which the phase changes by \(\pi\) radians, which means that the sign of the impulse response “flips.”

Again note that the magnitude of the impulse response is the unit constant:

\[
|h [x, y]| = \frac{1}{\lambda_0 z_1} \cdot 1 [x, y]
\]

which indicates that the assumed illumination from a point source in the Fresnel diffraction region is constant off axis; there is no “inverse square law.” This obviously unphysical assumption limits the usefulness of calculated diffraction patterns to the immediate vicinity of the optical axis of symmetry.

Profiles of the impulse response along a radial axis are shown for \(\alpha_0 = \sqrt{\lambda_0 z_1} = 1\) and \(\alpha_0 = 2\). In the second case, \(\sqrt{z_1}\) is larger by a factor of two, so \(z_1\) is larger by a factor of four.
1-D profiles of the impulse response of Fresnel diffraction for (a) $\sqrt{\lambda z_1} = 1$ and (b) $\sqrt{\lambda z_1} = 2$, so that $z_1$ is four times larger in (b). Note that the phase increases less rapidly with $x$ for increasing distances from the source.

1-D profiles of the impulse response of Fresnel diffraction for $\lambda_0 z_1 = 1$. Note that the magnitude of the impulse response is 1 and the phase is a quadratically increasing function of $x$.

The convolution integral is straightforward to implement.

Computed Examples of Fresnel Diffraction

The convolution formulation of diffraction allows easy calculation of profiles of diffraction patterns.
Fresnel Diffraction from a Point Source

\[ f[x, y] = \delta[x - x_0, y - y_0] \]

\[
* \left( \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \left( \exp \left[ +i\pi \frac{x^2}{\lambda_0 z_1} \right] \cdot \exp \left[ +i\pi \frac{y^2}{\lambda_0 z_1} \right] \right) \right)
\]

\[
= \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \delta[x - x_0] \exp \left[ +i\pi \frac{x^2}{\lambda_0 z_1} \right] \delta[y - y_0] \exp \left[ +i\pi \frac{y^2}{\lambda_0 z_1} \right]
\]

\[
= \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i\pi \frac{(x - x_0)^2}{\lambda_0 z_1} \right]
\]

The irradiance is proportional to the squared magnitude:

\[
|f[x, y]|^2 = \left| \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i\pi \frac{(x - x_0)^2 + (y - y_0)^2}{\lambda_0 z_1} \right] \right|^2
\]

\[
= \left( \frac{1}{\lambda_0 z_1} \right)^2
\]

The impulse response is translated and now centered about the location of the point source.
Fresnel Diffraction from Two Point Sources

\[ f[x, y] = \delta [x + x_0, y] + \delta [x - x_0, y] = (\delta [x + x_0] + \delta [x - x_0]) \cdot \delta [y] \]

The separability of the quadratic-phase kernel allows the convolution to be evaluated easily because the "input function" \( f[x, y] \) also is separable:

\[
g[x, y] = f[x, y] * h[x, y] = (\delta [x + x_0] + \delta [x - x_0]) \cdot (\delta [x + x_0] \cdot \exp \left[ +i \pi \frac{x^2}{\lambda_0 z_1} \right] \cdot \exp \left[ +i \pi \frac{y^2}{\lambda_0 z_1} \right])
\]

\[
= \left( \frac{1}{i \lambda_0 z_1} \exp \left[ 2 \pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cdot \left( \delta [x + x_0] \cdot \exp \left[ +i \pi \frac{x^2}{\lambda_0 z_1} \right] \cdot \exp \left[ +i \pi \frac{y^2}{\lambda_0 z_1} \right] \right)
\]

\[
= \left( \frac{1}{i \lambda_0 z_1} \exp \left[ 2 \pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cdot \left( \delta [x + x_0] \cdot \exp \left[ +i \pi \frac{x^2}{\lambda_0 z_1} \right] \cdot \exp \left[ +i \pi \frac{y^2}{\lambda_0 z_1} \right] \right)
\]

\[
= \left( \frac{1}{i \lambda_0 z_1} \exp \left[ 2 \pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cdot \left( \exp \left[ +i \pi \frac{(x + x_0)^2}{\lambda_0 z_1} \right] + \exp \left[ +i \pi \frac{(x - x_0)^2}{\lambda_0 z_1} \right] \right) \cdot \exp \left[ +i \pi \frac{y^2}{\lambda_0 z_1} \right]
\]
\[
g [x, y] = \left( \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \\
\times \left[ \exp \left[ +i\pi \frac{x^2 + 2xx_0 - 2xx_0}{\lambda_0 z_1} \right] + \exp \left[ +i\pi \frac{x^2 + x_0^2 - 2xx_0}{\lambda_0 z_1} \right] \right] \\
\times \exp \left[ +i\pi \frac{y^2}{\lambda_0 z_1} \right] \\
= \left( \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \\
\times \left[ \exp \left[ +i\pi \frac{2xx_0}{\lambda_0 z_1} \right] + \exp \left[ -i\pi \frac{2xx_0}{\lambda_0 z_1} \right] \right] \\
\times \exp \left[ +i\pi \frac{x_0^2}{\lambda_0 z_1} \right] \exp \left[ +i\pi \frac{y^2}{\lambda_0 z_1} \right] \\
= \left( \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \\
\times \left( 2 \cos \left[ +2\pi \frac{xx_0}{\lambda_0 z_1} \right] \right) \exp \left[ +i\pi \frac{x_0^2}{\lambda_0 z_1} \right] \exp \left[ +i\pi \frac{(x^2 + y^2)}{\lambda_0 z_1} \right] \\
= \left( \frac{2}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cos \left[ +2\pi \frac{xx_0}{\lambda_0 z_1} \right] \exp \left[ +i\pi \frac{x_0^2}{\lambda_0 z_1} \right] \exp \left[ +i\pi \frac{(x^2 + y^2)}{\lambda_0 z_1} \right] \\
\text{The irradiance is proportional to squared magnitude:} \\
|g [x, y]|^2 = \left| \left( \frac{2}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \right|^2 \cdot \cos^2 \left[ +2\pi \frac{xx_0}{\lambda_0 z_1} \right] \\
\times \left| \exp \left[ +i\pi \frac{x_0^2}{\lambda_0 z_1} \right] \right|^2 \cdot \left| \exp \left[ +i\pi \frac{(x^2 + y^2)}{\lambda_0 z_1} \right] \right|^2 \\
= \frac{4}{(\lambda_0 z_1)^2} \cos^2 \left[ +2\pi \frac{x}{(\lambda_0 z_1)} \right] \\
= \frac{4}{(\lambda_0 z_1)^2} \left( \frac{1}{2} \left( 1 + \cos \left[ 2 \left( +2\pi \frac{x}{(\lambda_0 z_1)} \right) \right] \right) \right) \\
= \frac{2}{(\lambda_0 z_1)^2} \left( 1 + \cos \left[ +2\pi \frac{x}{(\lambda_0 z_1)} \right] \right) \\
= \frac{2}{(\lambda_0 z_1)^2} \left( 1 + \cos \left[ +2\pi \frac{x}{D} \right] \right) \\
\text{where the period of the sinusoidal fringe pattern is:} \\
D = \frac{\lambda_0 z_1}{2x_0} = \frac{\lambda_0}{2 \sin \theta} \\
\text{where } \theta \text{ is the angle of each aperture measured from the optical axis at the observation plane. This} \\
\text{is the same pattern from interference of two plane waves that are tilted relative to the } z \text{-axis by angle } \theta. \\
\text{The length of the period increases with increasing propagation distance } z_1, \text{ increasing wavelength } \lambda_0, \text{ or decreasing separation of the point sources } x_0.
**Fresnel Diffraction from a Knife Edge** that would be generated from a knife edge at the same distances from the source as shown above. Note the “ringing” at the edges and that the fringes are farther apart when observed farther from the source. Compare these images to actual Fresnel diffraction patterns in Hecht.

One feature to note is that the irradiance at the projected location of the edge is 0.25 of the “average” maximum value, regardless of the propagation distance.
6.3 Fresnel Diffraction

1-D profiles of the irradiance (squared magnitude) of diffraction patterns from a sharp “knife edge” (modeled as the STEP function shown) for the same distances from the origin: (a) $\sqrt{\lambda z_1} = 1$; (b) $\sqrt{\lambda z_1} = 2 \implies z$ is four times larger. Note that the “period” of the oscillation has increased with increasing distance from the source and that the irradiance at the origin is not zero but rather $\frac{1}{4}$.

Fresnel Diffraction from a Rectangular Aperture (“slit”) Recall that a 1-D rectangle function may be constructed by subtracting appropriately translated STEP functions:

$$RECT[x] = \text{STEP}[x + \frac{1}{2}] - \text{STEP}[x - \frac{1}{2}]$$

Since convolution is linear and shift invariant, the Fresnel diffraction pattern of a rectangular aperture may be calculated by convolving the impulse response for the appropriate propagation distance with appropriately translated step functions, subtracting the amplitude distributions, and calculating the squared magnitudes.
Fresnel diffraction of rectangle as difference of two edges: (a) \( f_1 \left[ x \right] = \text{STEP} \left[ x + 4 \right] \) and \( f_2 \left[ x \right] = \text{STEP} \left[ x - 4 \right] \); (b) \( \text{Re} \left\{ f_1 \left[ x \right] \exp \left[ +i \pi \left( \frac{x}{4} \right)^2 \right] \right\} \) and \( \text{Re} \left\{ f_2 \left[ x \right] \exp \left[ +i \pi \left( \frac{x}{4} \right)^2 \right] \right\} \); (c) \( \text{Im} \left\{ f_1 \left[ x \right] \exp \left[ +i \pi \left( \frac{x}{4} \right)^2 \right] \right\} \) and \( \text{Im} \left\{ f_2 \left[ x \right] \exp \left[ +i \pi \left( \frac{x}{4} \right)^2 \right] \right\} \); (d) real and imaginary parts of difference; (e) squared magnitude \( \left[ \left( f_1 \left[ x \right] - f_2 \left[ x \right] \right) \exp \left[ +i \pi \left( \frac{x}{4} \right)^2 \right] \right]^2 \) compared to \( \text{RECT} \left[ \frac{x}{8} \right] \).
Fresnel diffraction of “narrower” rectangle as difference of two edges: (a) $f_1[x] = \text{STEP}[x+2]$ and $f_2[x] = \text{STEP}[x-2]$; (b) $\text{Re}\left\{f_1[x] \ast \exp \left[+i\pi \left(\frac{x}{4}\right)^2\right]\right\}$ and $\text{Re}\left\{f_2[x] \ast \exp \left[+i\pi \left(\frac{x}{4}\right)^2\right]\right\}$; (c) $\text{Im}\left\{f_1[x] \ast \exp \left[+i\pi \left(\frac{x}{4}\right)^2\right]\right\}$ and $\text{Im}\left\{f_2[x] \ast \exp \left[+i\pi \left(\frac{x}{4}\right)^2\right]\right\}$; (d) real and imaginary parts of difference; (e) squared magnitude $\left|(f_1[x] - f_2[x]) \ast \exp \left[+i\pi \left(\frac{x}{4}\right)^2\right]\right|^2$ compared to $\text{RECT}\left[\frac{x}{4}\right]$. 
6.3.2 Characteristics of Fresnel Diffraction

Fresnel diffraction of rectangle farther “downstream”: (a) \( f_1[x] = \text{STEP}[x + 4] \) and \( f_2[x] = \text{STEP}[x - 4] \); (b) \( \text{Re} \left\{ f_1[x] * \exp \left[ +i\pi \left( \frac{x}{2} \right)^2 \right] \right\} \) and \( \text{Re} \left\{ f_2[x] * \exp \left[ +i\pi \left( \frac{x}{2} \right)^2 \right] \right\} \); (c) \( \text{Im} \left\{ f_1[x] * \exp \left[ +i\pi \left( \frac{x}{2} \right)^2 \right] \right\} \) and \( \text{Im} \left\{ f_2[x] * \exp \left[ +i\pi \left( \frac{x}{2} \right)^2 \right] \right\} \); (d) real and imaginary parts of difference; (e) squared magnitude \( (f_1[x] - f_2[x]) * \exp \left[ +i\pi \left( \frac{x}{2} \right)^2 \right] \) compared to \( \text{RECT} \left[ \frac{x}{8} \right] \).

The parabolic approximation to the spherical impulse response of light propagation produces “images” of the original object that have “fuzzy edges” and oscillating amplitude on the bright side of an edge. At a fixed distance from the object, the width of the diffraction pattern is proportional to the width of the original object; if the object becomes wider, so does the “image” in the Fresnel.
6.3 Fresnel Diffraction

diffraction pattern.

The Fresnel diffraction patterns of at the same distance from the origin for two rectangles with different widths, showing that the “width” of the Fresnel pattern is proportional to the width of the object. Note that the “period” of the “ringing” is larger for larger observation distance $z$.

One of the more important aspects of imaging using the Fresnel approximation to diffraction is that the “imaging system” is linear and shift invariant; the action of light propagation may be described by an impulse response and a corresponding transfer function. Note that this is different from reality where light propagates as spherical waves, which is linear but not shift invariant because the amplitude generated at an observation point from individual identical sources changes due to the different propagation distances.

6.3.3 Transfer Function of Fresnel Diffraction

Fresnel diffraction is described by a convolution with a quadratic-phase impulse response, which is the product of a spatial constant and a separable quadratic-phase factor:

$$h[x, y] = \frac{1}{i \lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i \pi \left( \frac{x^2 + y^2}{(\sqrt{\lambda_0 z_1})^2} \right) \right]$$

$$\equiv A_0 \exp \left[ +i \pi \frac{x^2}{(\sqrt{\lambda_0 z_1})^2} \right] \exp \left[ +i \pi \frac{y^2}{(\sqrt{\lambda_0 z_1})^2} \right]$$

where the leading constants were collapsed into $A_0$:

$$A_0 \equiv \frac{1}{i \lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right]$$

We may describe the same system by an equivalent representation in the frequency domain via the known 1-D Fourier transform of the quadratic-phase factor:

$$\mathcal{F}_1 \left\{ \exp \left[ [\pm i \pi x^2] \right] \right\} = \exp \left[ \pm \frac{\pi}{4} \right] \cdot \exp \left[ \mp i \pi \xi^2 \right]$$

and the scaling theorem:

$$\mathcal{F}_1 \left\{ f \left[ \frac{x}{\alpha} \right] \right\} = |\alpha| F \left[ \alpha \xi \right] = |\alpha| F \left[ \frac{\xi}{\alpha^{-1}} \right]$$
Therefore, we have:

\[ H [\xi, \eta] = \mathcal{F}_2 \{ h [x, y] \} = \mathcal{F}_2 \left\{ A_0 \exp \left[ +i\pi \frac{x^2}{(\sqrt{\lambda_0 z_1})^2} \right] \exp \left[ +i\pi \frac{y^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right\} \]

\[ = A_0 \cdot \left( \mathcal{F}_1 \left\{ \exp \left[ +i\frac{\pi}{4} \frac{x^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right\} \right) \left( \mathcal{F}_1 \left\{ \exp \left[ +i\frac{\pi}{4} \frac{y^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right\} \right) \]

\[ = A_0 \cdot \left( \sqrt{\frac{\lambda_0 z_1}{\lambda_0 z_1}} \exp \left[ +i\frac{\pi}{4} \frac{x^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right) \left( \sqrt{\frac{\lambda_0 z_1}{\lambda_0 z_1}} \exp \left[ -i\frac{\pi}{4} \frac{y^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right) \]

\[ = A_0 \cdot \left( \frac{\lambda_0 z_1}{\lambda_0 z_1} \exp \left[ +i\frac{\pi}{4} \frac{x^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right) \left( \frac{\lambda_0 z_1}{\lambda_0 z_1} \exp \left[ -i\frac{\pi}{4} \frac{y^2}{(\sqrt{\lambda_0 z_1})^2} \right] \right) \]

\[ = \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot (i\lambda_0 z_1 \exp \left[ -i\pi \lambda_0 z_1 (\xi^2 + \eta^2) \right]) \]

\[ = \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \exp \left[ -i\pi \lambda_0 z_1 (\xi^2 + \eta^2) \right] \]

The transfer function of light propagation in the Fresnel region is circularly symmetric:

\[ H [\xi, \eta] = \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \exp \left[ -i\pi \lambda_0 z_1 \rho^2 \right] \]

\[ = \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \exp \left[ -i\pi \frac{\rho^2}{\left( \frac{1}{\lambda_0 z_1} \right)^2} \right] \]

\[ = \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \exp \left[ -i\pi \frac{\rho^2}{\alpha_1^2} \right] \]

where \( \rho \equiv \sqrt{\xi^2 + \eta^2} \). The leading factor has “disappeared” into the complex factor that carries the linear phase dependence on propagation distance \( z_1 \) and on time \( t \). This modulates a “downchirp” factor (leading negative sign) whose chirp rate is \( \alpha_1 \equiv \sqrt{\frac{1}{\lambda_0 z_1}} = \frac{1}{\alpha_0} \).

Some authors call \( H [\xi, \eta] \) the transfer function of free space, but to me this name does not convey the idea that it applies to propagation by the specific distance \( z_1 \).

### 6.4 Fraunhofer Diffraction

The diffraction integral may be further simplified for the case where the distance from the source to the observation plane is sufficiently large to allow the electric field from an individual source to be approximated by a plane wave. The process may be considered for one of the paraboloidal waves:

\[ \exp \left[ \frac{i\pi}{\lambda_0 z_1} \left( (x_0 - x_0)^2 + (y_1 - y_0)^2 \right) \right] \]

\[ = \exp \left[ \frac{i\pi (x_0^2 + y_0^2)}{\lambda_0 z_1} \right] \cdot \exp \left[ \frac{i\pi (x_0^2 + y_0^2)}{\lambda_0 z_1} \right] \cdot \exp \left[ -\frac{2\pi i}{\lambda_0 z_1} (x_0 x_1 + y_0 y_1) \right] \]

If the source is restricted to be near to the optical axis so that \( x_0, y_0 \approx 0 \) (or, more rigorously, \( x_0^2 + y_0^2 << \lambda_0 z_1 \)), then

\[ \frac{(x_0^2 + y_0^2)}{\lambda_0 z_1} \approx 0 \]
Considered in the observation plane as functions of magnitudes of the field:

The irradiance pattern in the Fraunhofer region again is proportional to the time-averaged squared relationships transform by mapping the frequency coordinates back to the space domain via the proportional in the Fresnel region. In fact, the Fraunhofer diffraction formula may be interpreted as a Fourier the resulting diffraction patterns are very different from those from the same aperture measured of plane waves generated by each source point. This is called the Fraunhofer diffraction formula, and the resulting diffraction patterns are very different from those from the same aperture measured in the Fresnel region. In fact, the Fraunhofer diffraction formula may be interpreted as a Fourier transform by mapping the frequency coordinates back to the space domain via the proportional relationships \( \xi = \frac{x_1}{\lambda z_1}, \eta = \frac{y_1}{\lambda z_1} \):

\[
E_{total} [x_1, y_1; z_1] = \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f [x_0, y_0] \exp \left[ -2\pi i \left( x_0 \left( \frac{x_1}{\lambda_0 z_1} \right) + y_0 \left( \frac{y_1}{\lambda_0 z_1} \right) \right) \right] dx_0 dy_0
\]

In words, the diffracted light measured far from the source is approximately equal to the summation of plane waves generated by each source point. This is called the Fraunhofer diffraction formula, and the resulting diffraction patterns are very different from those from the same aperture measured in the Fresnel region. In fact, the Fraunhofer diffraction formula may be interpreted as a Fourier transform by mapping the frequency coordinates back to the space domain via the proportional relationships \( \xi = \frac{x_1}{\lambda z_1}, \eta = \frac{y_1}{\lambda z_1} \):

\[
E_{total} [x_1, y_1; z_1] = \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f [x_0, y_0] \exp \left[ -2\pi i \left( x_0 \left( \frac{x_1}{\lambda_0 z_1} \right) + y_0 \left( \frac{y_1}{\lambda_0 z_1} \right) \right) \right] dx_0 dy_0
\]

The irradiance pattern in the Fraunhofer region again is proportional to the time-averaged squared magnitudes of the field:

\[
|E_{total} [x_1, y_1; z_1]|^2 \propto \frac{1}{\lambda_0^2 z_1^2} \cdot \left| F \{ f [x, y] \} \right|_{\xi = \frac{x_1}{\lambda z_1}, \eta = \frac{y_1}{\lambda z_1}}^2
\]

The amplitude of the field evaluated in the Fraunhofer diffraction region is proportional to the
appropriately scaled Fourier transform of the original amplitude, which therefore demonstrates that the
process is linear and shift \textit{variant}, rather than the shift-invariant result in the Fresnel region. Propagation into the Fraunhofer region in this model cannot be described by an impulse response or transfer function. However, also note that the Fresnel region includes the Fraunhofer region, so it is possible to calculate the amplitude by that means. The fact that the two calculations will not agree is a byproduct of the additional approximation in the Fraunhofer region.

All of the Fourier transform relationships apply to Fraunhofer diffraction patterns. The scaling theorem of the Fourier transform shows that the patterns (the “images”) scale in inverse proportion to the original functions, so that larger input functions $f[x, y]$ produce smaller (and brighter) diffraction patterns. The shift theorem indicates that movement of the input object “off axis” produces a linear phase in the diffraction pattern; since the irradiance is (generally) viewed, the translation off axis does not generally have a visible effect on the diffraction pattern, but its effect is visible when two such patterns are added.

6.4.1 Examples of Fraunhofer Diffraction

No Aperture

We can now easily evaluate the Fraunhofer diffraction patterns produced by simple apertures. First, consider the case if the aperture does not exist, so that:

$f_1[x, y] = 1 [x, y] = 1 [x] \cdot 1 [y] \implies F_1[\xi, \eta] = \delta [\xi] \cdot \delta [\eta]$

Apply the Fourier transform to evaluate the spectrum in the irradiance in the Fraunhofer diffraction region:

$|E_{\text{total}}[x_1, y_1; z_1]|^2 \propto \frac{1}{\lambda_0^2 z_1^2} \cdot \left| \delta \left[ \frac{x_1}{\lambda_0 z_1}, \frac{y_1}{\lambda_0 z_1} \right] \right|^2$

This does not make sense! The reason is actually quite simple; the size of the (nonexistent) aperture does not obey the assumed constraint for Fraunhofer diffraction, that $x_0, y_0 \cong 0$, which means that the aperture is small.

Rectangular Aperture

The profiles of 2-D square apertures and computed simulations of the resulting profiles of the amplitude diffraction pattern in the Fraunhofer diffraction region are easy to calculate:

$f[x, y, z = 0] = RECT \left( \frac{x}{a}, \frac{y}{b} \right)$

$E_{\text{total}}[x_1, y_1; z_1] \propto \frac{|ab|}{\lambda_0^2 z_1^2} \cdot SINC \left[ \frac{x_1}{\lambda_0 z_1}, \frac{y_1}{\lambda_0 z_1} \right]$

The irradiance is the squared magnitude of the plotted amplitude. The original point source located at $z_1 \cong \infty$ so that the light “fills” the aperture with zero phase. Two examples are shown for the same fixed (large) distance from the object; the “images” “narrow” and get “taller” as the aperture width increases.
1-D profiles of Fraunhofer diffraction patterns: (a) input objects are two rectangles that differ in width; (b) amplitude (NOT irradiance) of the corresponding Fraunhofer diffraction patterns, showing that the wider aperture produces a “brighter” and “narrower” amplitude distributions.

A useful measure of the “width” of the Fraunhofer pattern is labeled for both cases; this is the distance from the center of symmetry to the first zero. This “width” provides a measure of the ability of the system to resolve fine detail, because the individual diffraction patterns generated by two point sources may overlap and become difficult to distinguish as individual sources as the angular separation of the point sources decreases.
Illustrations of resolution in Fraunhofer diffraction: (a) individual images of two point sources in the Fraunhofer domain; (b) sum of the two images, showing that they may be distinguished easily; (c) images of two sources that are closer together; (d) sum showing that it is much more difficult to distinguish the sources.

6.5 Coherence of Light and Imaging

We should also emphasize that the analysis and the “images” just considered assumed illumination with a single wavelength $\lambda_0$, which means that light can “interfere” constructively and destructively after traveling different paths (distances and/or times) from the source. Of course, this constraint does not apply in realistic imaging systems; the range of wavelengths $\Delta\lambda$ typically is large (for visible light, $\Delta\lambda \cong 300$ nm). It often is convenient to specify the bandwidth in frequency:

$$\Delta\nu \equiv \nu_{\text{max}} - \nu_{\text{min}} = \frac{c}{\lambda_{\text{min}}} - \frac{c}{\lambda_{\text{max}}} = c \left( \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}} \cdot \lambda_{\text{min}}} \right) \cong \frac{c}{\lambda_0^2} \cdot \Delta\lambda$$

Therefore the frequency bandwidth of white visible light is:

$$\Delta\nu \cong 3 \times 10^8 \text{ m s}^{-1} \cdot \frac{700 \text{ nm} - 400 \text{ nm}}{400 \text{ nm} - 700 \text{ nm}} \cong 3.2 \times 10^{14} \text{ Hz}$$
which is of the same order as a wavelength of light, because

\[ \nu_{red} = \frac{3 \times 10^8 \text{ m s}^{-1}}{700 \text{ nm}} = \frac{3 \times 10^8 \text{ m s}^{-1}}{700 \cdot 10^{-9} \text{ m}} \approx 4.3 \times 10^{14} \text{ Hz} \]

The bandwidth of the light source has a profound effect on the “quality” of the image produced by the system. If, as we have assumed thus far, the source generates the single wavelength \( \lambda_0 \), then the bandwidth \( \Delta \nu = 0 \text{ Hz} \) and the light is “perfectly coherent,” which means that the phase of the light at all points in space is deterministic (this is an idealization, because a traveling wave with \( \Delta \nu = 0 \) must be infinitely long). When the amplitude of light is combined from different points on the source or after traveling different paths, the deterministic phase means that the interference effects are important at the image. For example, the light can destructively interfere to create a null amplitude, or it may even produce a “negative” amplitude. In imaging systems used in coherent light, the amplitudes of the light waves add and phase effects are paramount. In short, an optical imaging system acting in coherent light is linear in the complex-valued amplitude.

At the opposite extreme of a very wide bandwidth, the light is “incoherent,” which means that the phase of the light measured over a long time scale is random. Incoherent light does not interfere in the conventional sense to create negative amplitudes, though it is still possible to construct meaningful interferograms (as demonstrated by Brown and Twiss). The quantity of importance when incoherent light waves are added is the irradiance. The irradiance cannot be negative and the importance of the relative phase of the light is buried in the calculation. An optical imaging system acting in incoherent light is linear in irradiance, which is the real-valued squared magnitude of the complex amplitude.

The quality of the light source cannot be either perfectly coherent or perfectly incoherent, but the action of the imaging system may be calculated rather easily in these two extreme cases, whereas the calculation is difficult or impossible in intermediate cases.

### 6.6 Transmissive Optical Elements

The imaging “systems” that we have considered so far is as simple as you can get; the light in the first cases merely propagates from one plane to another by Huygens’ principle, and the amplitude is calculated using one of the two approximations: the Fresnel transform in the near field and the Fourier transform in the far field. The shift-variant expression for the diffraction integral simplified to a convolution with a quadratic-phase factor in the Fresnel region, and to a scaled Fourier transform in the Fraunhofer region. In the second case, light propagated from the source to the aperture to the observation plane.

At this point, we consider methods for manipulating or modifying the propagating light waves by inserting optical elements at the pupil plane that are based on the physical interaction of refraction; elements may also be constructed based on the mechanisms of reflection and diffraction. The physical mechanism of refraction is the result of the fact that light waves (or rays) within a medium slow down in inverse proportion to a parameter known as the refractive index of the medium, i.e.,

\[ v_{\text{glass}} = \frac{c}{n}; \quad c \approx 3 \cdot 10^8 \text{ m s}^{-1} \]

If the glass has thickness \( \tau \) (measured in mm, say), then the “arrival time” of a fixed point (point of constant phase) on a light wave that travels through the glass will be delayed when compared to that for a wave traveling the same distance through vacuum. The time delay is:

\[ \Delta t = t_{\text{glass}} - t_{\text{vacuum}} = \frac{\tau}{v_{\text{glass}}} - \frac{\tau}{c} = \frac{\tau (n - 1)}{c} \text{ seconds} \]

This time delay produces a difference in the phase of the light that passes through the glass when
compared to that traveling through vacuum:

$$\Delta \phi = \phi_{\text{glass}} - \phi_{\text{vacuum}} = (c \cdot \Delta t) \frac{2\pi}{\lambda_0} = 2\pi (n - 1) \frac{\tau}{\lambda_0} \text{ radians.}$$

An ideal nonabsorptive transmissive optical element of refractive index $n$ and circularly symmetric physical thickness $\tau(r)$ may be described by a complex-valued transmittance function:

$$t(r) = e^{+i[\Delta \phi(r)]} = e^{+2\pi i \frac{n \cdot \tau(r)}{\lambda_0}} = e^{+2\pi i (n-1) \frac{\tau(r)}{\lambda_0}} e^{+2\pi i \frac{\tau(r)}{\lambda_0}}$$

Our task now is to evaluate this transmittance function for different phase functions.

### 6.6.1 Optical Elements with Constant or Linear Phase

If the optical element is a slab of glass with uniform thickness $\tau$, then the phase incident on the glass measured at all points after the glass additional constant factor of $\exp\left[+2\pi i \frac{n \cdot \tau}{\lambda_0}\right]$. Since the phase change is constant, the effect of the glass plate on the observed irradiance is undetectable. An example is shown in the figure.

![Phase shift induced by a flat piece of glass compared to the phase of light propagating in vacuum.](image)

*The locus of points of constant phase on the entering wavefront is a vertical plane. The relative phase shift of light traveling through the glass is constant (not a function of $x$ or $y$ and the locus of points of constant phase is now two discontinuous planes.*

In a slightly more complicated case where the glass thickness varies with position $[x, y]$, so will the phase delay. For example, if the thickness is a linear function of $x$, then the transmittance function is:

$$\tau[x, y] = \alpha x \cdot 1[y] \rightarrow t[x, y] = e^{+2\pi i(n-1) \frac{\alpha x}{\lambda_0}} 1[y]$$

Such an optical element is called a *prism*, as shown in the figure. The term $\frac{(n-1)\alpha}{\lambda_0}$ has dimensions of reciprocal length and thus may be identified as a spatial frequency:
Schematic of the phase shift induced by a prism with thickness function \( \tau [x, y] = \alpha x \). The locus of the points with constant phase on the entering wavefront is a vertical plane. Distances between several points where the optical path difference is \( 6 \lambda \) (phase difference of \( 12\pi \) radians) are shown, which confirm that the wavefront was tilted “down” by the action of the prism.

\[
t[x, y] = \exp [+2\pi \xi_0 x] \cdot 1[y] \quad \text{where} \quad \xi_0 = (n - 1) \frac{\alpha}{\lambda_0}
\]

Assume that the prism located at \( z_1 >> 0 \) (in the Fraunhofer region) is illuminated by a point source at \( z = 0 \). The light emerging from the prism propagates to a plane at \( z = z_1 + z_2 \) that is in the Fraunhofer region measured from \( z_1 \). The “image” generated at the observation plane is proportional to the Fourier transform of the unit-magnitude “aperture” function \( t[x_1, y_1] \):

\[
E[x, y; z_1 + z_2] = \mathcal{F}_2 \{ t[x_1, y_1] \} \left( \xi - \frac{x}{\lambda_0 z_2}, \eta - \frac{y}{\lambda_0 z_2} \right)
\]

\[
= \mathcal{F}_2 \{ \exp [+2\pi \xi_0 x] \cdot 1[y] \} \left( \xi - \frac{x}{\lambda_0 z_2}, \eta - \frac{y}{\lambda_0 z_2} \right)
\]

\[
= \delta \left[ \frac{x}{\lambda_0 z_2} - (n - 1) \frac{\alpha}{\lambda_0} \right] \cdot \delta \left[ \frac{y}{\lambda_0 z_2} \right]
\]

\[
= |\lambda_0 z_2|^2 \delta [x - (n - 1) \alpha z_2] \cdot \delta [y]
\]

The action of the prism is a change in position of the displayed “image” from the origin to the off-axis location \( [x_2, y_2] = [(n - 1) \alpha z_2, 0] \).

### 6.6.2 Lenses with Spherical Surfaces

Optical systems typically are used to form images of the source distribution by redirecting the radiation from point sources to point images. For example, consider a slab of glass with maximum thickness \( \tau \) and with a convex spherical face with radius of curvature \( R \) (i.e., a plano-convex lens). The thickness function of the glass is found by using the Pythagorean formula in the following picture:

\[
(R - s(r))^2 + r^2 = R^2 \quad \Rightarrow \quad r^2 - 2R \cdot s(r) + s^2(r) = 0
\]

If both \( R \) and \( r \) are much larger than \( s(r) \), then the factor \( s^2(r) \) may be ignored. The result is a simple formula, the sag formula for a spherical lens:

\[
s(r) \approx \frac{r^2}{2R}
\]
as shown in the figure. In words, the sag formula approximates the spherical surface by a paraboloid. This process is analogous to the approximation of a spherical wavefront as a paraboloid in the Fresnel diffraction region.

The thickness of glass as a function of radial distance from the optical axis is the difference between the maximum thickness and the parabolic sag:

\[ \tau (r) = \tau - s (r) \cong \tau - \frac{r^2}{2R} \]

which measures the thickness of glass as a function of radial distance \( r \) from the optical axis. The phase delay as a function of radial distance from the optic axis is the sum of the phase delay through the glass and the phase delay through air in the sag region. The delay due to the glass is the number of wavelengths multiplied by the refractive index

\[
\phi (r) = \left( \frac{2\pi}{\lambda_0} \times \text{distance traveled in glass} \right) + \left( \frac{2\pi}{\lambda_0} \times \text{distance traveled in vacuum} \right) \\
= \left( \frac{2\pi}{\lambda_0} \right) \cdot \tau (r) + \left( \frac{2\pi}{\lambda_0} \right) \cdot s (r) \\
= 2\pi \frac{n}{\lambda_0} \left( \tau - \frac{r^2}{2R} \right) + 2\pi \frac{r^2}{\lambda_0} \\
= 2\pi \frac{n}{\lambda_0} \tau - \pi (n - 1) \frac{r^2}{\lambda_0 R}
\]

The first (constant) term is the fixed phase delay of light rays traveling through the maximum thickness \( \tau \) of glass, while the second term is a quadratic function that decreases as a function of the radial distance \( r \) measured from the optical axis due to the more rapid velocity of light in the sag region. The quadratic approximation to the spherical surface is the source of the quadratic phase contribution of a spherical lens.

The electric field exiting the lens as a function of radial position \( r \) from the optical axis is the product of the incident field and the phase delays, leading to constant- and quadratic-phase terms:

\[ E [r, t] = E_0 [r] \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ -i\pi (n - 1) \frac{r^2}{\lambda_0 R} \right] \]

In the common case of a lens has two convex spherical surfaces with radii \( R_1 \) and \( R_2 \), the radius of curvature of the one surface must have the opposite sign because it is oriented in the opposite direction. By convention, a surface with center of curvature to the right of the surface has positive
radius. The phase transmittance of the lens has contributions from each sag:

\[
\begin{align*}
    t[r] &= \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ -i\pi (n-1) \frac{r^2}{\lambda_0 R_1} \right] \cdot \exp \left[ +i\pi (n-1) \frac{r^2}{\lambda_0 (-R_2)} \right] \\
    &= \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ -i\pi (n-1) \frac{r^2}{\lambda_0} \left( \frac{1}{R_1} - \frac{1}{R_2} \right) \right]
\end{align*}
\]

The expression may be simplified by combining terms to define the factor \( f \) that has dimensions of length:

\[
\frac{1}{f} = (n-1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right)
\]

This possibly is a familiar equation; it is the *lens-maker’s formula* that relates the *focal length* \( f \) of the lens to the index of refraction \( n \) and the radii of curvature \( R_1 \) and \( R_2 \) of the two surfaces. The effect of the lens is to change the phase of the transmitted light due to the optical transmission function \( t(r) \):

\[
\begin{align*}
    t(r) &= \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ -i\pi \frac{r^2}{\lambda_0 f} \right] \\
    &= \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ -i\pi \frac{r^2}{\lambda_0 |f|} \right] \text{ if } f > 0
\end{align*}
\]

Note that if \( f \) is negative, the algebraic sign of the quadratic phase due to the lens is positive:

\[
\begin{align*}
    t(r) &= \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ +i\pi \frac{r^2}{\lambda_0 |f|} \right] \text{ if } f < 0
\end{align*}
\]

which means that the phase of the transmitted light *increases* with increasing radial distance from the optical axis. At this point, the transmittance is a pure phase function and thus has infinite support. The finite diameter of a realistic lens may be introduced as a real-valued function \( p[x, y] \) (which may be circularly symmetric or not) that defines the *pupil* of the lens. The pupil function usually is binary \((p[x, y] = 0, 1)\) but we can (and will) use pupil functions whose transmission varies with radial distance. A circularly symmetric spherical lens with positive focal length (converging lens) will be modeled by the transmittance function:

\[
\begin{align*}
    t(r) &= p(r) \exp \left[ +2\pi i \frac{n\tau}{\lambda_0} \right] \cdot \exp \left[ -i\pi \frac{r^2}{\lambda_0 f} \right]
\end{align*}
\]

as shown in the figure. In words, a lens changes the radius of curvature of the incident parabolic wave by adding or subtracting a phase factor that is a quadratic function of the distance from the optical axis. The variable part of the phase contribution from a spherical lens is quadratic with a negative sign, while the (approximate) impulse response of light propagation in the Fresnel diffraction region is a quadratic-phase factor with a positive sign.
CHAPTER 6 OPTICAL DIFFRACTION AND IMAGING

The negative quadratic phase induced by a positive lens changes the radius of curvature of incident plane waves to converging spherical waves that focus at the distance $f$.

6.7 Fraunhofer Diffraction and Imaging

A monochromatic point object located a long distance away from an imaging system produces a set of wavefronts that are approximately planar. The entrance pupil of the optical system (the image of the aperture stop in object space) collects a section of each plane wavefront and the optical elements convert it to a spherical (actually, only approximately spherical) wavefront that converges to an image “point.” We can use the concept of Fraunhofer diffraction to define the “angular resolution” of the imaging system.

As an introduction, consider an optical system that consists of only an aperture, which we can think of as the entrance pupil or the aperture stop, because they coincide if no other optics are involved. The pieces of the wavefront from the object source that are collected by the stop would continue to propagate “downstream.” If observed a long distance from the stop/pupil, the irradiance pattern from each point source is the Fraunhofer diffraction pattern of the stop, and thus proportional to the appropriately scaled Fourier transform of the stop. Recall that the Fourier transform has a “reciprocal” relationship: the smaller the stop, the larger the diffraction pattern and vice versa. Of course, the observed irradiance is the time average of the squared magnitude of the amplitude, and thus is nonnegative.

Now consider the situation if the object instead consists of two monochromatic point sources that appear to be displaced by the small angle $\theta_0$ when viewed from the aperture/pupil. The light from each source propagates from the aperture at the same angle $\theta_0$ to create the Fraunhofer diffraction pattern of the aperture at the observation plane. The light from the “off-axis” source forms a replica of the same Fraunhofer diffraction pattern but at the same angular separation $\theta_0$. The observed irradiance again is the time average of the squared magnitude of this amplitude. If the aperture stop is “wide” in some sense, then the diffraction patterns will be “narrow” because of the reciprocal relation between the widths of the aperture and its Fraunhofer diffraction pattern. If the diffraction patterns are “narrow”, the fact that the object consisted of two point sources may be apparent. Conversely, if the stop is “narrow” and the diffraction patterns are “wide,” the patterns from the two sources may overlap and be difficult to distinguish. Thus the ability of the system to resolve two point sources separated by the angle $\theta$ depends on the size of the entrance pupil.
Fraunhofer diffraction of aperture stop with no optic: the distance from the monochromatic point object to the stop is \( z_1 \approx \infty \) and the light diffracted by the stop forms a Fraunhofer irradiance pattern on the observation screen at \( z_2 \approx \infty \). If the object consists of two point sources at the same distance \( z_1 \) separated in angle by \( \theta_0 \), the diffracted amplitude is the sum of the individual amplitudes separated by the same angle \( \theta_0 \).

Of course, real optical imaging systems consist of more than just an aperture; they also include lenses and/or mirrors that change the curvature of that part of the wavefront that passes through the stop/pupil. The section of the (approximately) spherical wave emerging from the optical system converges approximately to a point in the image plane. Because the distance \( z_1 \) from the object to the system is large, the image plane is approximately located at the focal plane, i.e., the distance \( z_2 = f \), the focal length of the lens. We can interpret the action of the lens as “bringing infinity up close (and personal?)”.

In other words, the lens creates a scaled replica of the light pattern that would have been generated at a large distance from the stop if the lens were not present; the image of a point object is created by the imaging system is a scaled replica of the Fraunhofer diffraction pattern of the entrance pupil (and thus of the aperture stop); the scale factor is the ratio of the original propagation distance \( z_1 \) to the focal length \( f \) (or to the “effective focal length \( f_{eff} \) if the optic is composed of multiple elements). This scaled replica of the Fraunhofer diffraction pattern of the stop is the impulse response of the optical imaging system in monochromatic (coherent) light. Note that the lens also adds a phase factor to the wavefront due to the curvature of the converging wave. This phase factor may be approximated by an additional quadratic-phase factor, but this effect is not generally significant in imaging because the sensor measures the irradiance of the light, which is proportional to the squared magnitude of the amplitude, eliminates this factor. If we ignore this additional quadratic-phase factor, the amplitude produced by a single on-axis point source in
monochromatic light is proportional to the Fourier transform of the entrance pupil $p[x, y]$:

$$E_{\text{total}}[x_1, y_1; f] = h[x_1, y_1; f] \propto P\left[\frac{x_1}{\lambda_0 \cdot f}, \frac{y_1}{\lambda_0 \cdot f}\right]$$

So, again, a large-diameter entrance pupil produces a “small” impulse response and “better” angular resolution, if the performance of the system is limited by diffraction (which often is not true; performance is frequently limited by aberrations in the optical system).

If the object consists of two point sources separated by the apparent angle $\theta_0$ as before, the amplitude pattern at the image plane is the sum of the two Fraunhofer diffraction patterns displaced by same angle but that have been “minified” by a constant factor equal to the ratio of the image distances in the two cases: $\frac{f}{z_2}$.

Lord Rayleigh derived a metric for the angular resolution of an imaging system if the wavelength is $\lambda_0$ and the two objects are separated by the distance $d$ in the object plane, the distance from the object to the imaging system is $z_1 = L$, and the diameter of the entrance pupil is $D$, then the Rayleigh criterion states that the two sources are just resolved in the image if the separation in the image plane is equal to the radius of the diffraction spot. For square apertures, the “half-width” of the diffraction spot is:

$$d \approx \frac{L\lambda_0}{D}$$

For circular apertures, the “half-width” (radius) of the diffraction spot is:

$$d \approx 1.22 \frac{L\lambda_0}{D}$$

(note the similarity to the formula for the separation between interference fringes in Young’s exper-
The angular radius of the image spot is a good measure of the angular resolution:

\[ \Delta \theta \cong \frac{d}{L} \cong \frac{\lambda_0}{D}, \text{ for square apertures} \]

To find the image pattern from a point source through a circular entrance pupil of diameter \( D \), we may use Gaskill’s notation for the so-called cylinder function:

\[ f[x, y] = CYL \left( \frac{r}{D} \right) \]

Its Fourier transform is:

\[
F[\xi, \eta] = \pi \left( \frac{D}{2} \right)^2 \text{SOMB} (D \rho) \\
= \pi \frac{D^2}{4} \cdot \frac{2J_1(\pi D \rho)}{\pi D \rho} = \frac{D}{2 \rho} J_1(\pi D \rho)
\]

We make the substitution that takes the equation from the frequency domain to the space domain by dividing by the product of the wavelength and the effective focal length:

\[ \rho = \sqrt{x^2 + y^2} \rightarrow \sqrt{\frac{x^2}{\lambda_0 f_{eff}} + \frac{y^2}{\lambda_0 f_{eff}}} = \frac{r}{\lambda_0 f_{eff}} \]

\[ \implies D \rho \rightarrow D \cdot \frac{r}{\lambda_0 f_{eff}} \]

The first zero of \( J_1(\pi u) \) occurs at \( u_0 \approx 1.22 \), which corresponds to a radial distance \( r_0 \) that satisfies the condition:

\[ \frac{Dr_0}{\lambda_0 f_{eff}} \approx 1.22 \]

\[ \implies r_0 \approx 1.22 \cdot \frac{\lambda_0 f_{eff}}{D} \]

The ratio of the effective focal length \( f_{eff} \) and the diameter of the entrance pupil \( D \) is the “f-number” (often abbreviated “f/#”) of the imaging system, which is a convenient measure of the angle of the cone of rays from the optical system to the image. The radius of the diffraction spot measured to the first zero may be written in concise form:

\[ r_0 \approx 1.22 \cdot \lambda_0 \cdot f/# \]

It often is “easier” to think of the diameter of the resolution “spot”:

\[ d_0 \approx 2.44 \cdot \lambda_0 \cdot f/# \]

This leads to a useful mnemonic that applies for visible light \((\lambda_0 \cong 0.5 \mu m)\). The product \(2.44 \cdot \lambda_0 \cong 2.44 \cdot 0.5 \mu m = 1.22 \cong 1.0\), so that:

\[ d_0 \text{ [\mu m]} \cong f/# \]

In words, the diameter (in micrometers) of the resolution spot of an imaging system operating in visible light is approximately equal to the f-number. This is the diffraction limit of the imaging system; it is a measure of the smallest distance on the image plane where two point objects may be resolved.

The angular diameter of the image spot produced by a system with a circular aperture is:

\[ 2 \cdot \Delta \theta \cong 2.44 \cdot \frac{\lambda_0}{D}. \]
Check this for the most accessible imaging system you have: your eyes. At night, the diameter of the pupil of a dark-adapted eye is \( D \approx 7 \text{ mm} \), so the Rayleigh limit of angular resolution near the middle of its sensitivity in green light (\( \lambda_0 = 550 \text{ nm} \)) is approximately:

\[
2 \cdot \Delta \theta \approx 2.44 \cdot \frac{0.55 \mu \text{m}}{7 \text{ mm}} \approx 1.92 \cdot 10^{-4} \text{ radians} \approx 0.192 \text{ milliradians}
\]

To convert this to seconds of arc, we need to know the number of arc-seconds in a radian:

\[
1 \text{ radian} = \frac{360^\circ}{2\pi \text{ radians}} \cdot \frac{60 \text{ arc-minutes}}{1 \circ} \cdot \frac{60 \text{ arc-seconds}}{1 \text{ arc-minute}} = 206,265 \text{ arc-seconds}
\]

\[
\implies 1 \text{ radian} \approx 2 \cdot 10^6 \text{ arc-seconds}
\]

\[
\implies 1 \text{ arc-second} \approx 5 \cdot 10^{-6} \text{ radians} = 5 \text{ microradians}
\]

which is a useful factor to remember. The Rayleigh limit to the resolution of the dark-adapted human eye is approximately:

\[
2.44 \cdot \frac{\lambda_0}{D} \approx 0.192 \cdot 10^{-3} \text{ radians} \cdot 206,265 \frac{\text{arc-seconds}}{\text{radians}}
\]

\[
\approx 40 \text{ arc-seconds} = \frac{2}{3} \text{ arc-minute}
\]

which is approximately \( \frac{1}{36} \) of the diameter of the moon. Since the eye is not “perfect,” we more often assume that the resolution of the eye is about 1 minute of arc, or \( \frac{1}{30} \) of the moon’s diameter.

\[
\Delta \theta_{\text{human eye}} \approx 1 \text{ arc-minute}
\]

The diameter of the primary mirror of the Hale telescope on Palomar Mountain is \( D \approx 5 \text{ m} \), so the theoretical minimum angular separation is \( \Delta \theta \approx 0.24 \cdot \lambda_0 \), for \( \lambda_0 \) measured in meters. The angular separation is:

\[
\Delta \theta \approx 1.32 \cdot 10^{-7} \text{ radians} = 0.132 \mu \text{radians} \approx 0.027 \text{ arc-seconds}
\]

Again, this is the theoretical limit that would apply to a perfectly figured mirror with no other “optical elements” in the system. The first condition is approximately true, but the second clearly is not. The ground-based Hale telescope views astronomical objects through the atmosphere, which is composed of volumes of gases with different temperatures (and thus different refractive indices) at different locations that move rapidly in time. The variation in refractive index over time has the effect of reducing the resolution of the telescope from its theoretical limit of about 0.3 arc-seconds to angles of the order of arc-seconds, i.e., a factor of 30 or so. This approximate angular resolution of the Hale telescope is that of an optical system with \( D \approx 100 \text{ mm} \approx 4 \text{ in} \). The additional aperture of a large telescope does have the notable benefit of collecting more light to allow imaging of fainter objects, but increasing the ability of the telescope to resolve smaller separations requires other technology, e.g., interferometric methods (stellar spectral interferometry) to correct the images after collection, or adaptive optics (“rubber mirrors”) to correct the images in “real time.” Adaptive optical systems have to respond to phase changes within a few hundredths of a second.

The diameter of the primary mirror of the Hubble space telescope is approximately half that of the Hale telescope (\( D \approx 2.4 \text{ m} \)), so the angular resolution of the HST optics at 550 nm is approximately twice as large (0.26 \( \mu \text{radian} \approx 0.054 \text{ arc-seconds} \)). Of course, there is no atmosphere to mess up the Hubble images, though readers may well recall some issues (to put it mildly) with mirror fabrication.

That the dimension of the resolution “spot” is directly proportional to the wavelength means that a system with fixed focal length and fixed entrance pupil diameter can resolve smaller separations at shorter wavelengths. Also, systems with smaller focal ratios have better resolution (“faster systems have better resolution”).

In real life, it is very difficult (meaning very expensive) to fabricate an optical imaging system with a large entrance pupil and whose performance is limited by diffraction. The performance of the
system is much more often constrained by the limitations of the optical design and fabrication so that the outgoing wavefront is not spherical, but has another shape. This behavior is characterized by various optical aberrations (e.g., spherical aberration, astigmatism, distortion) that quantify the difference from the ideal spherical wave. Generally speaking, aberrations become more difficult and expensive to correct as the size of the entrance pupil increases.

Of course, the ultimate resolution of the Hale telescope actually is limited by atmospheric turbulence, which creates random variations in local air temperature and thus in the refractive index of “patches” of the atmosphere. These variations are often decomposed into the aberrations introduced into the wavefront by the phase errors. The constant phase (“piston”) error has no effect on the irradiance (the squared magnitude of the amplitude). Linear phase errors move the image from side to side and or top to bottom (“tip-tilt”). Quadratic phase errors (“defocus”) act like additional power in the lens that moves the image plane backwards or forwards along the optical axis. In general, tip-tilt error is the most significant, which means that correcting this aberration significantly improves the image quality. The field of correcting atmospheric aberrations is called “adaptive optics,” and is an active research area.

6.7.1 “Diffraction-Limited” and “Detector-Limited” Imaging Systems

The “blur” of an imaging and/or sensor system discussion just concluded on the effect of Fraunhofer diffraction in imaging shows that images generated by ideal optical imaging systems exhibit “blur” due to unavoidable diffraction of the light focused by the optics. The linear dimension of this diffraction spot defines one possible measure of the spatial “resolution” of the images, i.e., the smallest separation between point objects (such as stars) that may be “just” distinguished by the system. It is important to note a couple of points about resolution. First, the range of wavelengths in the light beam being imaged has an important impact on resolution. In other words, the resolution of an imaging system used in “natural” (incoherent) light that has a wide range of wavelengths is different from the resolution of the same imaging system used in “coherent” light (as from a laser) where the range of wavelengths is small. Also note that several metrics of resolution are commonly used, including the Rayleigh and Dawes limits, so it is important to understand which is being used to specify the quality of the system.

In most optical systems of interest, the image is focused onto an imaging sensor with its own resolution properties. This concept perhaps has become more familiar with the widespread availability of discrete electronic imaging sensors, such as the charge-coupled device (CCD). The concept of resolution of a discrete sensor is quite straightforward, because we usually assume that the individual sensor elements have identical uniform response over their entire area. We should note that some these assumptions are not sufficient for some critical applications. For our purposes, the simplifying assumptions are sufficient. The transverse magnification of the imaging system and the linear dimension of the sensor elements determines the smallest separation between two point sources that may be resolved by this sensor; the resolution limit requires that the image field measured by the sensor of two equally bright point objects must be separated by a sensor element that responds less.

6.8 Depth of Field and Depth of Focus

In a system consisting of both optics and sensor, the system is “diffraction-limited” if the pixel size is smaller than the linear dimension of the diffraction spot. The system is “detector-limited” or “sensor-limited” if the linear dimension of the individual sensor elements is larger than the diffraction spot.

The discussion of the limiting “blur” of an imaging system may be extended to characterize the range of “distances” (or “depths”) over which images of point objects exhibit the “same” (or at least “similar”) blur dimensions. If specified in object space, the distance range is called the “depth of field;” the same metric in image space is the “depth of focus.”

There is no one way to define the depths of field and focus, but we can rather easily derive a “hybrid” metric (based on both ray and wave optics). The measurement is based on the linear dimension of the blur spot, which is the diffraction spot in a diffraction-limited system or on the size
CHAPTER 6 OPTICAL DIFFRACTION AND IMAGING

of the sensor element in a detector-limited system. Until the arrival of digital imaging systems, the former case of diffraction-limited optics has been most often considered because diffraction provided the fundamental limit to resolution. As we shall see, the depth of field is determined by the f-number of the optical system.

Assume that the linear dimension of image blur has been measured for a particular imaging system at the specific pair of object and image distances \((z, z')\) of interest:

\[
\text{Blur in a diffraction-limited system with aperture diameter } D. \text{ The image of the point source is a diffraction pattern at the image plane whose linear dimension (using some criterion) is } B_0.
\]

For example, the image of a point source located a distance \(z\) from the system could be measured to find this limiting “blur diameter” \(B'\), where the prime indicates that the measurement is made in image space. In a diffraction-limited system, the discussion of Fraunhofer diffraction in imaging shows that one possible measure for \(B'\) is the diameter of the central lobe of the diffraction spot:

\[
B' = 2.44 \cdot \lambda_0 \cdot \frac{z'}{D} \approx 2.44 \cdot \lambda_0 \cdot \frac{f_{eff}}{D} = 2.44 \cdot \lambda_0 \cdot f/\#
\]

From \(B'\), it is easy to evaluate the corresponding dimension \(B\) in object space from the transverse magnification \(M_T = -\frac{z'}{z} = \frac{y'}{y}\).
6.8 DEPTH OF FIELD AND DEPTH OF FOCUS

The calculation of depth of field: \( B' \) is the linear dimension of the blur for the system, either the diameter of the diffraction spot in a diffraction-limited system or the dimension of the sensor element in a detector-limited system. The locations \( z' \pm \delta' \) specify locations in image space where the geometrical blur has the same size. The corresponding locations in object space are the limits of the “depth of field.”

As shown in the figure, the “blur” due to geometrical rays from the point object has the same dimension \( B' \) as the blur spot at two locations equidistant from the “in-focus” image. We assign the name \( \delta' \) to the distance between the “in-focus” image and the geometrically blurred images, so these two planes are located at \( z' \pm \delta' \). The depth of focus in this model is twice \( \delta' \):

\[ \Delta z' = 2 \cdot \delta' \]

In the ray model, the drawing shows that:

\[ \frac{D}{z'} = \frac{B'}{\delta'} \implies \delta' = B' \frac{z'}{D} \approx B' \cdot f/\# \]

If \( B' \) is small, so must be \( \delta' \); if the \( f/\# \) is large, so must be \( \delta' \).

The corresponding locations in object space (and the depth of field) may be determined by substituting these distances and the focal length \( f_{eff} \) into the imaging equation. For example

\[ \frac{1}{z_1} = \frac{1}{f_{eff}} - \frac{1}{z'_1} = \frac{1}{f_{eff}} - \frac{1}{z' - \delta'} \]

The distances \( z_1 \) and \( z_2 \) may be evaluated from the imaging equation for the corresponding image distances \( z'_1 = z' - \delta' \) and \( z'_2 = z' + \delta' \). It is easy to see that the absolute magnification \( |M_T| \) is smaller for the smaller image distance, i.e., \( M_T \) for \( z'_1 = z' - \delta' \) is smaller than \( M_T \) for the larger object distance \( z'_2 = z' + \delta' \). The nonlinearity of the imaging equation ensures that the distances between the in-focus object distance \( z \) and the extrema are not equal, i.e., \( z_1 - z \neq z - z_2 \), thus
requiring labels for both: \( z_1 = z + \delta_1 \) and \( z_2 = z - \delta_2 \). However, if \( \delta' \) is small, then the concept of longitudinal magnification allows simple approximate expressions for the object distances:

\[
\begin{align*}
    z_1 &= z + \delta_1 \simeq z + \frac{\delta'}{|M_L|} = z + \frac{\delta'}{M_T^2} \\
    z_2 &= z - \delta_2 \simeq z - \frac{\delta'}{|M_L|} = z - \frac{\delta'}{M_T^2}
\end{align*}
\]

The depth of field may be written as:

\[
\Delta z = \delta_1 + \delta_2 \simeq 2 \cdot \frac{\delta'}{M_T^2} = 2 \cdot \frac{B' \cdot f/\#}{M_T^2}
\]

In the diffraction-limited case, we can substitute the known width of the diffraction pattern for \( B' \) to obtain:

\[
\Delta z \simeq 2 \cdot (2.44 \cdot \lambda_0 \cdot f/\#) \cdot \frac{f/\#}{M_T^2} = 4.88 \cdot \frac{\lambda_0 \cdot (f/\#)^2}{M_T^2}
\]

In the detector-limited case with pixel dimension \( b \), the depth of field is:

\[
\Delta z \simeq 2 \cdot \frac{b \cdot f/\#}{M_T^2}
\]

In words, the depth of field in image space is proportional to the wavelength and the square of the \( f \)-number of the system, which means that “slower” systems have longer depths of field. This is rather obvious from the drawing, because the angle of the extreme rays is smaller for a slow system.

Another important metric in imaging is the hyperfocal distance, which is the shortest object distance where the depth of field extends to infinity. The corresponding image distance \( z' \) is the sum of the focal length and \( \delta' \):

\[
\begin{align*}
    z_1 &= z + \delta_1 = \infty \implies z' - \delta' = f_{e\text{ff}} \\
    \implies z' &= f_{e\text{ff}} + \delta'
\end{align*}
\]

The hyperfocal object distance satisfies the imaging equation:

\[
\frac{1}{z} + \frac{1}{z'} = \frac{1}{f_{e\text{ff}}}
\]

where \( z' = f_{e\text{ff}} + \delta' \), so that:

\[
\begin{align*}
    z &= \left( \frac{1}{f_{e\text{ff}}} - \frac{1}{f_{e\text{ff}} + \delta'} \right)^{-1} \\
    &= \left( \frac{f_{e\text{ff}} + \delta' - f_{e\text{ff}}}{(f_{e\text{ff}})(f_{e\text{ff}} + \delta')} \right)^{-1} \\
    &= \frac{(f_{e\text{ff}} + \delta')(f_{e\text{ff}})}{\delta'} \\
    &\approx \frac{f_{e\text{ff}}^2}{2.44 \cdot \lambda_0 \cdot f/\#} = \frac{f_{e\text{ff}}^2}{2.44 \cdot \lambda_0 \cdot \frac{f_{e\text{ff}}}{D}} \propto f_{e\text{ff}} \cdot D
\end{align*}
\]