## Contents

**Preface** ix

**1 Optical Phenomena** 1

1.1 Sources: 2

1.1.1 Physics Texts: 2

1.1.2 Optics Texts: 2

**2 Oscillatory Motion** 5

2.1 Harmonic Oscillations 6

2.2 Harmonic Oscillations – Energy Considerations 7

2.2.1 Anharmonic Oscillations 10

2.3 Representations of Harmonic Oscillations 11

**3 Complex Numbers** 13

3.1 Graphical Representation of Complex Numbers 15

3.2 Euler Relation – Complex Exponentials 18

3.3 Arithmetic of Complex Exponentials 20

3.3.1 De Moivre’s Theorem 21

3.4 Description of Harmonic Oscillations via the Euler relation 22

3.5 Oscillations as Projections of Circular Harmonic Motion 23

3.6 Phasor Notation for Oscillations 23

3.7 Superposition of Oscillations 24

3.7.1 Digression: Nonlinear Optics and Second-Harmonic Generation: 24

3.8 Superposition of Same-Frequency Oscillations 25

3.8.1 Trigonometric Notation: 25

3.8.2 Phasor Representation: 27

3.8.3 Complex Notation: 29

3.9 Superposition of Many Same-Frequency Oscillators 29

3.10 Superposition of Randomly Phased Oscillators 31

3.11 Superposition of Nonrandomly Phased Oscillators 32

3.12 Superposition of Oscillations with Different Frequencies 34

3.12.1 Complex Notation 34

3.13 Introduction to Fourier Analysis 37

3.13.1 Fourier Transform 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.8.4</td>
<td>4. Ampere’s Law</td>
<td>86</td>
</tr>
<tr>
<td>6.8.5</td>
<td>Maxwell’s Equations</td>
<td>86</td>
</tr>
<tr>
<td>6.9</td>
<td>Wave Equation for EM Waves</td>
<td>87</td>
</tr>
<tr>
<td>6.9.1</td>
<td>Electromagnetic Waves from Maxwell’s Equations</td>
<td>90</td>
</tr>
<tr>
<td>6.9.2</td>
<td>Phase Velocity of Electromagnetic Waves</td>
<td>92</td>
</tr>
<tr>
<td>6.10</td>
<td>Consequences of Maxwell’s Equations</td>
<td>95</td>
</tr>
<tr>
<td>6.11</td>
<td>Dispersion Redux</td>
<td>97</td>
</tr>
<tr>
<td>6.11.1</td>
<td>Feynman’s Model of Optical Scattering by Matter</td>
<td>99</td>
</tr>
<tr>
<td>6.11.2</td>
<td>Forces on the Charge</td>
<td>106</td>
</tr>
<tr>
<td>6.11.3</td>
<td>Complete Dispersion Curves</td>
<td>107</td>
</tr>
<tr>
<td>6.11.4</td>
<td>Optical Path Length</td>
<td>112</td>
</tr>
<tr>
<td>6.11.5</td>
<td>Dispersion via Convolution (Optional)</td>
<td>113</td>
</tr>
<tr>
<td>6.12</td>
<td>Dual Nature of Light: Photons</td>
<td>117</td>
</tr>
<tr>
<td>6.12.1</td>
<td>Momentum of Photons</td>
<td>119</td>
</tr>
<tr>
<td>6.13</td>
<td>Optical Frequencies – Detector Response</td>
<td>119</td>
</tr>
<tr>
<td>7</td>
<td>Propagation of Light Waves</td>
<td>121</td>
</tr>
<tr>
<td>7.1</td>
<td>Plane Waves</td>
<td>121</td>
</tr>
<tr>
<td>7.2</td>
<td>Cylindrical Waves</td>
<td>122</td>
</tr>
<tr>
<td>7.3</td>
<td>Spherical Waves</td>
<td>123</td>
</tr>
<tr>
<td>7.4</td>
<td>Huygens’ Principle</td>
<td>124</td>
</tr>
<tr>
<td>8</td>
<td>Interaction of Light and Matter</td>
<td>125</td>
</tr>
<tr>
<td>8.1</td>
<td>Fermat’s Principle</td>
<td>125</td>
</tr>
<tr>
<td>8.1.1</td>
<td>Fermat’s Principle for Refraction:</td>
<td>127</td>
</tr>
<tr>
<td>8.2</td>
<td>Electromagnetic Waves at an Interface</td>
<td>129</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Definition of Vectors for Waves at Interface</td>
<td>129</td>
</tr>
<tr>
<td>8.3</td>
<td>Snell’s Law for Reflection and Refraction of Waves</td>
<td>131</td>
</tr>
<tr>
<td>8.4</td>
<td>Boundary Conditions for Electric and Magnetic Fields</td>
<td>133</td>
</tr>
<tr>
<td>8.5</td>
<td>Transverse Electric Waves, “s” or “⊥” Polarization”</td>
<td>136</td>
</tr>
<tr>
<td>8.6</td>
<td>Transverse Magnetic Waves (“p” or “∥” polarization)</td>
<td>140</td>
</tr>
<tr>
<td>8.7</td>
<td>Comparison of Coefficients for TE and TM Waves</td>
<td>141</td>
</tr>
<tr>
<td>8.7.1</td>
<td>Normal Incidence ($\theta_0 = 0$)</td>
<td>142</td>
</tr>
<tr>
<td>8.7.2</td>
<td>Angular Dependence of Reflection and Transmittance at “Rare-to-Dense” Interface</td>
<td>143</td>
</tr>
<tr>
<td>8.7.3</td>
<td>Brewster’s Angle – Angle of Complete Polarization</td>
<td>144</td>
</tr>
<tr>
<td>8.7.4</td>
<td>Reflectance and Transmittance at “Rare-to-Dense” Interface</td>
<td>146</td>
</tr>
<tr>
<td>8.7.5</td>
<td>Reflection and Transmittance at “Dense-to-Rare” Interface, Critical Angle</td>
<td>146</td>
</tr>
<tr>
<td>8.7.6</td>
<td>Practical Applications for Fresnel’s Equations</td>
<td>147</td>
</tr>
<tr>
<td>9</td>
<td>Polarization</td>
<td>149</td>
</tr>
<tr>
<td>9.1</td>
<td>Plane Polarization = Linear Polarization</td>
<td>149</td>
</tr>
<tr>
<td>9.2</td>
<td>Circular Polarization</td>
<td>150</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Nomenclature for Circular Polarization</td>
<td>151</td>
</tr>
</tbody>
</table>
CONTENTS

9.2.2 Elliptical Polarization, Reflections ........................................ 152
9.2.3 Change of Handedness on Reflection .................................... 152
9.2.4 Natural Light .......................................................................... 152
9.3 Description of Polarization States ............................................. 153
   9.3.1 Jones Vector ......................................................................... 153
9.4 Generation of Polarized Light .................................................... 154
   9.4.1 Selective Emission: ................................................................. 154
   9.4.2 Selective Transmission or Absorption .................................... 155
   9.4.3 Generating Polarized Light by Reflection – Brewster’s Angle ... 155
   9.4.4 Polarization by Scattering ...................................................... 157

10 Birefringence – Double Refraction ........................................ 159
   10.0.5 Examples: ............................................................................ 159
10.1 Phase Delays in Birefringent Materials; Wave Plates .................. 160
   10.1.1 Example: .............................................................................. 160
   10.1.2 Example: .............................................................................. 161
10.2 Circular Polarizer: ................................................................. 161
10.3 Critical Angle – Total Internal Reflection .................................. 161

11 Optical Interference ............................................................... 165
   11.1 Division of Wavefront ............................................................. 165
   11.1.1 Superposition of Two Same-Frequency Plane Waves .......... 170
   11.1.2 Superposition of Two Plane Waves with Different Frequencies 176
   11.2 Fringe Visibility – Coherence .................................................. 179
   11.2.1 Coherence Time and Coherence Length ............................... 184
   11.2.2 Polarization and Fringe Visibility ......................................... 186
11.3 Division-of-Amplitude Interferometers ..................................... 187
   11.3.1 Michelson Interferometer ..................................................... 187
   11.3.2 Mach-Zehnder Interferometer .............................................. 192
   11.3.3 Sagnac Interferometer ......................................................... 193
   11.3.4 Fizeau Interferometer .......................................................... 194

12 Optical Diffraction and Imaging .............................................. 197
   12.1 Diffraction Integrals ............................................................... 198
   12.2 Fresnel Diffraction ................................................................. 200
   12.2.1 Fresnel Diffraction as a Convolution Integral ....................... 202
   12.2.2 Characteristics of Fresnel Diffraction .................................... 205
   12.3 Fraunhofer Diffraction ............................................................ 206
   12.3.1 Computed Examples of Fraunhofer Diffraction .................... 208
   12.4 Fraunhofer Diffraction and Imaging ....................................... 209
Preface

This course considers the link in the imaging chain that collects the radiation and creates the image (optics) using the model that light is a wave.

References

WebTOP (Web resources for “The Optics Project” of Mississippi State University, http://webtop.msstate.edu/)
Chapter 1
Optical Phenomena

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 \( \text{RAY} \) that travels in a straight line until it encounters an interface between media. The wavelength \( \lambda \) and 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) useful for designing imaging systems.

(d) 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 \( \text{WAVE} \);

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

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

(d) leads to explanations of \( \text{reflection, refraction, diffraction, interference, polarization, dispersion} \). Useful for assessing the quality of the images.

III. Quantum Optics: Atomic-scale Phenomena

(a) light is a \( \text{photon} \), has 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 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.

1.1 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.

1.1.1 Physics Texts:


1.1.2 Optics Texts:

1.1 SOURCES:

Chapter 2

Oscillatory Motion

Sources: HR §15, C$§1$

Before discussing the wave nature of light, it is important to review the salient characteristics of oscillations and waves from basic physics.

**Oscillation** – periodic variation of any characteristic of a physical system about some equilibrium (mean) value, e.g.,

1. position angle of a pendulum bob in a gravitational field,
2. voltage across the capacitor in an LC circuit,
3. position of a mass on a spring.

Three systems that exhibit oscillation: the pendulum in a gravitational field, the L-C oscillator, and a mass on a spring.

The position angle $\theta$ of the pendulum bob and the voltage across the capacitor plates (or current in the circuit or the magnetic field generated by the inductor, ...) oscillate as functions of time. The position angle of the pendulum varies about its mean value (the vertical defined by the gravitational field) as a periodic function of time.

Oscillations result from the joint and balanced presence of two forces:

1. **Inertia**: displaces the physical quantity (e.g., the position angle of the pendulum or the voltage in the LC oscillator) from its equilibrium value.
2. **Restoring (or return) force**: opposes changes in the physical quantity, acts to return it to equilibrium. The greater the deviation from equilibrium, the larger the restoring force. (acts as negative feedback whose restoring force increases with deviation from equilibrium)

Oscillations of matter can be either *transverse* or *longitudinal* (or some combination):

1. **Longitudinal oscillation**: the vectors describing the opposing forces are parallel, e.g., as for a mass attached to a coil spring that is restricted to motion toward or away from the anchoring point of the spring;

2. **Transverse oscillation**: the vectors describing the two forces are not parallel, e.g., for the pendulum, where inertial force is horizontal and restoring force is directed vertically downward or for unrestricted motion of mass on a helical spring.

### 2.1 Harmonic Oscillations

The simplest oscillations are *harmonic*, which means that the function describing the oscillation is composed of a single sinusoidal frequency (usually defined as a *cosine* rather than a *sine* because this is more compatible with complex notation). For example, consider the position angle of a pendulum in a gravitational field as a function of time:

\[
y [t] = y_0 + A_0 \cos \left( \omega_0 t + \phi_0 \right)
\]

![Sinusoidal oscillation over time.](image)

1. \( y \) is the coordinate (or position) of the oscillating quantity of the system, (e.g., angle, voltage, etc.)

2. \( y_0 \) is the equilibrium value of the measured quantity;
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 \( y \): \([A] = [y]\);

4. \( \omega_0 \) is the *angular temporal frequency* of the oscillation, measured in units of \([\omega_0] = \text{radians}, \text{second}^{-1}\);

5. \( v_0 \) is the *temporal frequency* of the oscillation, measured in units of \([v_0] = \text{cycles}, \text{second} = \text{Hertz} \text{[Hz]}\); \( v_0 = \frac{\omega_0}{2\pi} \);

6. \( T_0 = \frac{1}{v_0} = \frac{2\pi}{\omega_0} \) is the temporal *period* of the oscillation, units of \([T_0] = \text{s}\);

7. \( \Phi \) is the *phase angle* of the oscillation (the argument of the sinusoid), measured in radians;

8. \( \phi_0 \) is the *initial phase angle* of the oscillation, i.e., the phase angle in radians measured at the origin of coordinates, i.e., at \( t = 0 \).

### 2.2 Harmonic Oscillations – Energy Considerations

RW §2.4

Given the equation of motion of a simple oscillatory system in the time domain, e.g., \( y(t) = A_0 \cos(\omega_0 t + \phi_0) \), the velocity, acceleration, and force exerted by the system may be calculated by differentiating. Note that physicists often use a shorthand notation for time derivatives with one overscored dot for the first derivative and two dots for the second.

**Velocity:**

\[
\begin{align*}
v &= \frac{dy}{dt} \equiv \dot{y} = \frac{d}{dt} (A_0 \cos(\omega_0 t + \phi_0)) \\
&= A_0 \frac{d}{dt} \left( \cos(\omega_0 t + \phi_0) \right) \\
&= A_0 \cdot (-\omega_0 \sin(\omega_0 t + \phi_0)) \\
&= -A_0 \omega_0 \sin(\omega_0 t + \phi_0) = A_0 \omega_0 \cos \left( \omega_0 t + \phi_0 + \frac{\pi}{2} \right)
\end{align*}
\]

Because of the time derivative, the velocity of the measurable quantity is “out of phase” relative to the motion by one-quarter of a cycle (+\( \frac{\pi}{2} \) radians). Because the phase difference between the velocity and the displacement is +\( \frac{\pi}{2} \) radians, the velocity “leads.” Therefore the largest positive value of the velocity occurs when the motion is at its equilibrium displacement and the displacement is increasing, as shown in the figure. The velocity is zero when the motion is at an extremum. Also note that the “amplitude” of the velocity is scaled by the factor of \( \omega_0 \); the larger the angular (and thus the temporal) frequency, the larger the velocity.
**CHAPTER 2 OSCILLATORY MOTION**

Acceleration:

\[
\begin{align*}
a &= \frac{d^2y}{dt^2} \equiv \ddot{y} = -A_0\omega_0^2 \cos [\omega_0 t + \phi_0] \\
&= A_0\omega_0^2 \cos [(\omega_0 t + \phi_0) \pm \pi]
\end{align*}
\]

The acceleration is the second derivative of the motion, and thus is out of phase by one-half cycle, so that when the motion is at its positive extremum the acceleration is at its negative extremum. The factor of \(\omega_0^2\) indicates that the extrema of the acceleration are larger if the temporal frequency is larger.

**Harmonic motion and its derivatives for two temporal frequencies:** (a) harmonic motion with a small temporal frequency (black), its velocity (red) is out of phase by one-quarter cycle, and the acceleration (blue) is out of phase from the motion by one-half cycle; (b) same graphs for a larger spatial frequency, showing that the velocity and acceleration are larger because of the scaling by \(\omega_0\) and \(\omega_0^2\), respectively.

**Inertial Force**

The force of motion in the system is the familiar product of the mass (or its analogue) and the acceleration:

\[
ma = m\ddot{y} = -m (\omega_0^2 A_0 \cos [\omega_0 t + \phi_0])
\]

\[
= -m\omega_0^2 (A_0 \cos [\omega_0 t + \phi_0])
\]

\[
= -m\omega_0^2 \cdot y [t] \equiv -ky
\]

where the “Force Constant” of Restoring Force is:

\[k \equiv m\omega_0^2\]
2.2 HARMONIC OSCILLATIONS – ENERGY CONSIDERATIONS

The force equation may be transposed to:

\[ m \ddot{y} + ky = 0 \]

this is the equation of motion of the simple harmonic oscillator. This equation is satisfied by all systems that exhibit simple harmonic motion and may be used as the defining equation for a sinusoidal function. In words, sinusoidal motion results if the displacement and the corresponding acceleration have opposing signs and the appropriate proportional relationship.

It is easy to derive the potential and kinetic energies of the harmonic oscillator from the motion and its derivatives:

**Kinetic Energy** \( E_k [t] \):

The energy of motion is determined from the familiar equation:

\[
E_k [t] = \frac{1}{2} m v^2 = \frac{m}{2} (-\omega_0 A_0 \sin [\omega_0 t + \phi_0])^2
\]

\[
E_k [t] = \frac{m A_0^2 \omega_0^2}{2} \sin^2 [\omega_0 t + \phi_0]
\]

**Potential Energy** \( E_p [t] \):

The energy of position is the integral of the force multiplied by the distance:

\[
E_p [t] = -\int_0^y F \cdot ds = -\int_0^y (-ky) \, dy = +\frac{ky^2}{2} = +\frac{m \omega_0^2 y^2}{2}
\]

\[
E_p [t] = \frac{m A_0^2 \omega_0^2}{2} \cos^2 [\omega_0 t + \phi_0]
\]

**Total Energy** \( E [t] \):

The energy is the sum of kinetic and potential energies:

\[
E [t] = E_k [t] + E_p [t] = \frac{m A_0^2 \omega_0^2}{2} \sin^2 [\omega_0 t + \phi_0] + \frac{m A_0^2 \omega_0^2}{2} \cos^2 [\omega_0 t + \phi_0]
\]

\[
E [t] = \frac{m A_0^2 \omega_0^2}{2} [\sin^2 [\omega_0 t + \phi_0] + \cos^2 [\omega_0 t + \phi_0]]
\]

\[
E [t] = \frac{m A_0^2 \omega_0^2}{2}
\]

which is constant over time. This is an expression of the law of conservation of energy.

\[
E = \frac{m A_0^2 \omega_0^2}{2}
\]

**Observations:**

1. \( E \) is not a function of time, i.e., the total energy is constant
2. $E_k[t]$ and $E_p[t]$ are both always greater than 0.

3. $\mathcal{E} \propto A_0^2$, the energy is proportional to the square of the amplitude

4. $\mathcal{E} \propto \omega_0^2$, the energy is proportional to the square of the frequency: higher frequency $\implies$ more energy for the same amplitude

5. $\omega_0^2$ is the return force per unit displacement per unit mass.

2.2.1 Anharmonic Oscillations

Now consider the more general case of nonharmonic (or anharmonic) oscillations, which simply means that the characteristic of the physical system varies in a nonsinusoidal manner. An example is shown in the figure:

Anharmonic motion can result simply by adding two oscillations with different frequencies (and possibly different amplitudes and phases). This case is relatively easy to analyze. However, more complicated motions (such as the “sawtooth” wave $y[t]$ in the figure) are produced by adding more oscillations. We can decompose a waveform into its constituent sinusoidal waves by applying Fourier analysis, which determines the amplitude, frequency, and phase of each sinusoidal component that sum up to make $y[t]$. Conversely, we can sum up sinusoidal components with specified amplitude, frequency, and initial phase to “synthesize” a waveform via Fourier synthesis.

The mathematical formulas for the motion and energy of the anharmonic oscillator are identical to those for the harmonic oscillator, but the derivatives and integrals are much more complicated to calculate. Fortunately, as we shall see, virtually any periodic function can be decomposed into a sum of harmonic functions. Recall that differentiation is linear, i.e.,

$$\text{if } f[x] = f_1[x] + f_2[x] \text{ then } \frac{df}{dx} = \frac{df_1}{dx} + \frac{df_2}{dx}$$

Therefore, the derivatives of each component may be taken separately and summed to find the derivatives of the result.

The decomposition of a function into its component frequencies is known as Fourier analysis, and will be discussed in more detail later.
2.3 Representations of Harmonic Oscillations

Since harmonic oscillators exhibit sinusoidal motion, they may certainly be described by trigonometric functions as above.

\[ y(t) = A_0 \sin(\omega_0 t + \phi_0) = A_0 \cos\left(\frac{\pi}{2} - (\omega_0 t + \phi_0)\right) = A_0 \cos\left((\omega_0 t + \phi_0) - \frac{\pi}{2}\right) \]

where the second expression arises because

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

where the last expression results from the symmetry of the cosine (i.e., \(\cos(-\theta) = \cos(\theta)\)). This description of oscillations is perfectly ok—it leads to all the correct results—but it can be awkward to keep a math handbook handy to recall the necessary expressions for the cosine and/or sine of sums, differences, and/or products of angles. The notation becomes even more complicated when considering the superposition (sum) of many oscillations or waves. For example, how easy is it to find the resultant of the sum of two oscillators, \(y_1(t) + y_2(t)\), where \(y_i = A_i \sin(\omega_i t + \phi_i)\)? You can look this up to find:

\[ y_1(t) + y_2(t) = \sin(\omega_1 t + \phi_1) + \sin(\omega_2 t + \phi_2) = 2 \sin\left(\frac{\omega_1 + \omega_2}{2}\right) t + \left(\frac{\phi_1 + \phi_2}{2}\right) \cdot \cos\left(\frac{\omega_1 - \omega_2}{2}\right) t + \frac{\phi_1 - \phi_2}{2} \]

but this result is easy to derive by using complex notation, as shown in the next section.
Chapter 3

Complex Numbers

H §2

The complex representation offers many mathematical advantages over trigonometric expression for oscillators.

Complex numbers arise from imaginary numbers. Since there is no real number solution for \(\sqrt{-1}\), the imaginary number \(i\) is arbitrarily assigned as the solution, i.e.,

\[ i = \sqrt{-1} \implies i^2 = -1 \]

**Complex Number:** A complex number \(z\) is an ordered pair of real numbers \([a, b] \equiv a + ib\):

- \(a\) is the real part \((\text{Re}\{z\})\) and \(b\) is the imaginary part \((\text{Im}\{z\})\).

**Complex Conjugate:** The complex conjugate of a complex number \(z = a + ib\) is defined as \(z^* = a - ib\). To evaluate the complex conjugate, simply replace \(i\) wherever it appears.

**Complex Arithmetic:** Given two complex numbers \(z_1 = a_1 + ib_1\) and \(z_2 = a_2 + ib_2\), the following arithmetic rules apply:

1. **Equality:** \(z_1 = z_2\) if and only if \(a_1 = a_2\) and \(b_1 = b_2\);

2. **Addition:** \(z_1 + z_2 = (a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2)\); (add real and imaginary parts separately, \(\text{Re}\{z_1 + z_2\} = a_1 + a_2\), \(\text{Im}\{z_1 + z_2\} = b_1 + b_2\));

3. **Multiplication:**

\[ z_1 \cdot z_2 = (a_1 + ib_1) \cdot (a_2 + ib_2) \]
\[ = a_1a_2 + a_1(ib_2) + a_2(ib_1) + (ib_1)(ib_2) \]
\[ = (a_1a_2 - b_1b_2) + i(a_1b_2 + a_2b_1) \]
\[ \text{Re}\{z_1 \cdot z_2\} = a_1a_2 - b_1b_2 \]
\[ \text{Im}\{z_1 \cdot z_2\} = a_1b_2 + a_2b_1; \]
Note that the product of \( z_1 \) with its complex conjugate \( z_1^* \) is:

\[
z_1 \cdot z_1^* = (a_1 + ib_1) \cdot (a_1 + ib_1)^*
= (a_1 + ib_1) \cdot (a_1 - ib_1)
= a_1a_1 + a_1(-ib_1) + a_1(+ib_1) + (+ib_1)(-ib_1)
= (a_1^2 + b_1^2) + i(a_1b_1 - a_1b_1)
\]

\[
\text{Re} \{z_1 \cdot z_1^*\} = a_1^2 + b_1^2
\]
\[
\text{Im} \{z_1 \cdot z_1^*\} = 0
\]
\[
z_1 \cdot z_1^* \equiv |z_1|^2
\]

The \textit{squared magnitude} of the complex number \( z_1 \) is the square of the length of the 2-D vector with components \([\text{Re} \{z_1\}, \text{Im} \{z_1\}]\).

4. Reciprocal: \( z_2^{-1} \) \textit{Use this trick:} multiply \( z_2^{-1} \) by unity specified as the ratio of its complex conjugate:

\[
1 = \frac{z_2^*}{z_2^*} = \frac{a_2 - ib_2}{a_2 - ib_2}, \text{ assuming that } z_2^* \neq 0, \text{ and thus that } z_2 \neq 0
\]

The reciprocal of \( z_2 \) is:

\[
\frac{1}{z_2} = \frac{z_2^*}{|z_2|^2} = \frac{1}{a_2 + ib_2} \cdot \frac{a_2 - ib_2}{a_2 - ib_2}
= \frac{a_2 - ib_2}{a_2^2 + b_2^2}
= \left( \frac{a_2}{a_2^2 + b_2^2} \right) + i \left( \frac{-b_2}{a_2^2 + b_2^2} \right)
\Rightarrow \text{Re} \left\{ \frac{1}{z_2} \right\} = \frac{a_2}{a_2^2 + b_2^2} = \frac{a_2}{|z_2|^2}
\Rightarrow \text{Im} \left\{ \frac{1}{z_2} \right\} = -\frac{b_2}{a_2^2 + b_2^2} = -\frac{b_2}{|z_2|^2}
\]

The magnitude and phase of the reciprocal of \( z_2 \) are:

\[
\left| \frac{1}{z_2} \right| = \sqrt{\left( \frac{a_2}{a_2^2 + b_2^2} \right)^2 + \left( \frac{-b_2}{a_2^2 + b_2^2} \right)^2}
= \sqrt{\frac{a_2^2 + b_2^2}{(a_2^2 + b_2^2)^2}} = \sqrt{\frac{1}{a_2^2 + b_2^2}} \text{ if } a_2 \neq 0 \text{ or } b_2 \neq 0
\]

\[
\Phi \left\{ \frac{1}{z_2} \right\} = \tan^{-1} \left[ \frac{-b_2}{a_2} \right] = \tan^{-1} \left[ \frac{-b_2}{a_2} \right] = -\tan^{-1} \left[ \frac{b_2}{a_2} \right] = -\Phi \left\{ z_2 \right\}
\]
5. Division: Apply multiplication and the reciprocal to obtain:

\[
\frac{z_1}{z_2} = \frac{a_1 + ib_1}{a_2 + ib_2}
\]

\[
\frac{z_1}{z_2} \cdot \frac{z_2^*}{z_2^*} = \frac{a_1 + ib_1}{a_2 + ib_2} \cdot \frac{a_2 - ib_2}{a_2 - ib_2} = \frac{(a_1a_2 + b_1b_2) + i(a_2b_1 - a_1b_2)}{a_2^2 + b_2^2}
\]

\[
\text{Re}\left\{ \frac{z_1}{z_2} \right\} = \frac{a_1a_2 + b_1b_2}{a_2^2 + b_2^2}
\]

\[
\text{Im}\left\{ \frac{z_1}{z_2} \right\} = \frac{a_2b_1 - a_1b_2}{a_2^2 + b_2^2}
\]

6. The real and imaginary parts of \(z\) may be expressed in terms of \(z\) and \(z^*\)

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

\[
\text{Im}\{z\} = \frac{1}{2} (z - z^*)
\]

The magnitude of \(z\) is defined as

\[
|z| \equiv \sqrt{z \cdot z^*}
\]

### 3.1 Graphical Representation of Complex Numbers

As you learned in high-school algebra, any ordered pair of numbers can be located on a two-dimensional (2-D) graph, e.g., using the Cartesian coordinates \([x, y]\). The \(y\)-axis becomes the imaginary axis, i.e., all values along \(y\) are implicitly multiplied by \(i = \sqrt{-1}\). Such a plot is sometimes called an Argand diagram.

For example, consider the case:

\[
z_1 = 1 + i
\]

\[
z_2 = 2 + i
\]

\[
\Rightarrow z_3 = z_1 + z_2 = 3 + 2i
\]

\[
\Rightarrow z_4 = z_1 - z_2 = -1
\]
Argand diagram of complex numbers $z_1$ and $z_2$.

Just as in algebra, we can also represent the Cartesian ordered pair $[a, b]$ in a polar notation $z_1 = (A_1, \phi_1)$, where $A_1$ is the magnitude of the vector $[a_1, b_1]$ and $\phi_1$ is its polar angle (or phase angle):

\[
\text{magnitude: } A_1 = |z_1| \equiv \sqrt{z_1 \cdot \bar{z_1}^*} = \sqrt{(a_1 + ib_1)(a_1 - ib_1)} = \sqrt{a_1^2 + b_1^2}
\]

Again, the magnitude is the length of the 2-D vector whose components are the real and imaginary parts of $z_1$, respectively.

\[
\text{phase: } \phi_1 = \tan^{-1} \left( \frac{b_1}{a_1} \right) = \tan^{-1} \left[ \frac{\text{Im} \{z_1\}}{\text{Re} \{z_1\}} \right]
\]

Though the range of valid phase angles is $-\infty < \phi < +\infty$, calculations of the inverse tangent of the ratio $\frac{\text{Im}(z_1)}{\text{Re}(z_1)}$ will return a value only within the interval $-\frac{\pi}{2} \leq \phi < +\frac{\pi}{2}$ because the information about the algebraic signs of the imaginary and real parts is lost within the ratio. A phase angle with a range of $2\pi$ radians ($-\pi \leq \phi < +\pi$) may be recovered by considering the signs of the real and imaginary parts to determine the appropriate quadrant. Some computer languages do this automatically via an arctangent function with arguments for both the real and imaginary parts. By convention, one interval of $2\pi$ radians is selected as the principal value of the phase that describes the possible values of $\phi$. Two semiclosed intervals are commonly used: $[-\pi, +\pi)$ and $[0, 2\pi)$. The first choice is used in this discussion because it is symmetric about the origin.

Phases over a larger range of angles may be inferred by determining where the calculated phase jumps by a large increment at adjacent coordinates due to the discontinuity and adding or subtracting the required increment of $2\pi$ radians at those coordinates; the process is called unwrapping the phase, as shown in the figure.
process is quite straightforward in 1-D but much more complicated in 2-D. This has the effect of constraining the derivative of the phase and thus the range of possible instantaneous frequencies.

Demonstration of phase “unwrapping for \( f[x] = \exp[+i\pi x^2] \).” (a) \( \text{Re}\{f[x]\} = \cos[\pi x^2] \) (in black) and \( \text{Im}\{f[x]\} = \sin[\pi x^2] \) (in red); (b) phase of \( f[x] \) without unwrapping (in black) is constrained to \(-\pi \leq \Phi[x] < +\pi\), while increments of \(2\pi\) radians are added to the unwrapped phase at the discontinuities to create the “smooth” curve.

The real and imaginary parts may be evaluated from the magnitude and phase via the trigonometric conversions:

\[
\text{Re}\{z_1\} = \text{Re}\{a_1 + ib_1\} = a_1 = A_1 \cos[\phi] \\
\text{Im}\{z_1\} = \text{Im}\{a_1 + ib_1\} = b_1 = A_1 \sin[\phi]
\]

Magnitude (a real number) \( \equiv |z_1| = \sqrt{a_1^2 + b_1^2} = \sqrt{A_1^2 \cos^2\phi_1 + A_1^2 \sin^2\phi_1} = \sqrt{A_1^2} = A_1 \)

\[
z_1 = \text{Re}\{z_1\} + i \ \text{Im}\{z_1\} \\
= A_1 \cos[\phi_1] + A_1(i \sin[\phi_1]) \\
= A_1(\cos[\phi_1] + i \sin[\phi_1])
\]
3.2 Euler Relation – Complex Exponentials

Complex numbers are very conveniently denoted as exponentials; makes multiplication easy. Represent $z$ in its polar form:

$$z = (r, \phi) = r (\cos \phi + i \sin \phi) = r e^{i\phi}$$

This expression arises from the Euler relation:

$$\cos \theta + i \sin \theta = e^{+i\theta}$$

**Proof.** Consider $z = [r \cos \theta, r \sin \theta] = (r, \theta)$

$$dz = (\cos \theta + i \sin \theta) \, dr + r (-\sin \theta \, d\theta + i \cos \theta \, d\theta)$$

$$= r (\cos \theta + i \sin \theta) \, dr + r (-\sin \theta + i \cos \theta) \, d\theta$$

$$= r (\cos \theta + i \sin \theta) \frac{dr}{r} + r (i^2 \sin \theta + i \cos \theta) \, d\theta$$

$$= z \frac{dr}{r} + i r (\cos \theta + i \sin \theta) \, d\theta = z \left( \frac{dr}{r} + i d\theta \right)$$

$$\frac{dr}{r} + id\theta = \frac{dz}{z} \implies \frac{dz}{z} = \int_0^r \frac{dr}{r} + i \int_0^\theta \, d\theta = \log_e r + i \theta$$

$$\log_e z = \log_e r + i \theta \implies \exp \left[ \log_e z \right] = \exp \left[ \log_e r + i \theta \right] = \exp \left[ \log_e r \right] \cdot \exp \left[ i \theta \right] = r \cdot e^{i \theta}$$

$$\implies z = [r \cos \theta, r \sin \theta] = r e^{i \theta}$$
Power-series expansions also may be used to derive the Euler relation:

\[ e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = x^0 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \]

\[ = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \quad [\text{where } 0! \equiv 1] \]

\[ e^{i\theta} = 1 + i\theta + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \cdots \]

\[ = 1 + i\theta + \frac{i^2\theta^2}{2!} + \frac{i^3\theta^3}{3!} + \frac{i^4\cdot i\theta^4}{4!} + \cdots \]

\[ = 1 + i\theta + \frac{\theta^2}{2!} - \frac{i\theta^3}{3!} + \frac{\theta^4}{4!} + \frac{i\theta^5}{5!} - \cdots \]

\[ = \left(1 - \frac{i\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots \right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots \right) \]

\[ = \cos \theta + i \sin \theta \]

where:

\[ \cos \theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots \]

\[ \sin \theta = \frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots \]

The complex conjugate of the Euler relation is:

\[ (e^{+i\theta})^* = e^{-i\theta} = \cos [\theta] - i \sin [\theta] \]

Clearly the Euler relation leads to other useful expressions for the cosine and sine:

\[ e^{+i\theta} = \cos \theta + i \sin \theta \]

\[ e^{-i\theta} = (e^{+i\theta})^* = \cos \theta - i \sin \theta \]

\[ e^{+i\theta} + e^{-i\theta} = 2 \cdot \cos [\theta] \implies \cos [\theta] = \frac{e^{+i\theta} + e^{-i\theta}}{2} \]

\[ e^{+i\theta} - e^{-i\theta} = i \cdot (2 \cdot \sin [\theta]) \implies \sin [\theta] = \frac{e^{+i\theta} - e^{-i\theta}}{2i} = \frac{i}{2} \left(e^{-i\theta} - e^{+i\theta}\right) \]
Examples of Euler’s relation:

\[ \phi = 0 \Rightarrow e^{0} = 1 \quad \text{because} \quad \cos [0] = 1 \quad \text{and} \quad \sin [0] = 0 \]
\[ \phi = \frac{\pi}{2} \Rightarrow e^{i\frac{\pi}{2}} = \cos \left[ \frac{\pi}{2} \right] + i \sin \left[ \frac{\pi}{2} \right] = i \]
\[ \phi = \pi \Rightarrow e^{i\pi} = \cos [\pi] + i \sin [\pi] = -1 \]

As another aside, the approximations for cosine, sine, and tangent of small angles may be evaluated from the series:

\[
\lim_{\theta \to 0} \{\cos [\theta]\} \simeq 1 - \frac{\theta^2}{2!} \to 1
\]
\[
\lim_{\theta \to 0} \{\sin [\theta]\} = \theta - \frac{\theta^3}{3!} \to \theta
\]
\[
\lim_{\theta \to 0} \{\tan [\theta]\} = \lim_{\theta \to 0} \left\{ \frac{\sin [\theta]}{\cos [\theta]} \right\} = \theta
\]

Graphs of \( \theta \), \( \sin [\theta] \), and \( \tan [\theta] \) are compared in the figure for \( \theta \) in the interval \( -\frac{\pi}{10} \leq \theta \leq +\frac{\pi}{10} \), which demonstrates the approximate equality for small \( \theta \).

Plots of \( \theta \), \( \sin [\theta] \), and \( \tan [\theta] \) for \( |\theta| \leq 1 \) radian, showing that the three functions are approximately equal for \( |\theta| \lesssim \frac{\pi}{10} \approx 0.31 \) radians.

### 3.3 Arithmetic of Complex Exponentials

1. equality:

   \[ z_1 = A_1 e^{i\phi_1} \text{ is equal to } z_2 = A_2 e^{i\phi_2} \text{ if and only if } A_1 = A_2 \text{ and } \phi_1 = \phi_2 \]
2. addition:

\[ z_1 \pm z_2 = A_1 e^{i\phi_1} \pm A_2 e^{i\phi_2} = A_1 e^{i\phi_1} \pm \left( A_1 e^{i\phi_2} + (A_2 - A_1) e^{i\phi_2} \right) = A_1 \left( e^{i\phi_1} \pm e^{i\phi_2} \right) \pm (A_2 - A_1) e^{i\phi_2} \]

where the decomposition is sometimes convenient to collect the phase terms together. Consider the case \( \phi_2 = -\phi_1 \):

\[ z_1 + z_2 = A_1 e^{i\phi_1} + A_2 e^{-i\phi_1} = A_1 \left( e^{i\phi_1} + e^{-i\phi_1} \right) + (A_2 - A_1) e^{i\phi_2} = 2A_1 \cos [\phi_1] + (A_2 - A_1) e^{i\phi_2} \]

3. multiplication:

\[ z_1 z_2 = A_1 e^{+i\phi_1} \cdot A_2 e^{+i\phi_2} = A_1 A_2 e^{+i(\phi_1 + \phi_2)} \]

4. division:

\[ \frac{z_1}{z_2} = \frac{A_1 e^{i\phi_1}}{A_2 e^{i\phi_2}} = \frac{A_1}{A_2} e^{+i(\phi_1 - \phi_2)} \]

### 3.3.1 De Moivre’s Theorem

Generalization of multiplication of complex exponentials:

\[ z^n = [r(\cos \phi + i \sin \phi)]^n = r^n [\cos \phi + i \sin \phi]^n = r^n \left( \cos [n\phi] + i \sin [n\phi] \right) \]

A representation of \( e^{-i\theta} \) can be derived using De Moivre’s theorem or the series expansion of \( e^{i\theta} \):

\[ e^{-i\theta} = e^{i(-\theta)} = 1 + i(-\theta) + \frac{i^2(-\theta)^2}{2!} + \frac{i^3(-\theta)^3}{3!} + \cdots = 1 - i\theta - \frac{\theta^2}{2!} + \frac{i\theta^3}{3!} + \frac{\theta^4}{4!} - \cdots = \left( 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots \right) + i \left( -\theta + \frac{\theta^3}{3!} - \frac{\theta^5}{5!} + \cdots \right) = \left( 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots \right) - i \left( \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots \right) = \cos [\theta] - i \sin [\theta] \]

As will be discussed, products and sums of same-frequency harmonic oscillations are easily computed.
3.4 Description of Harmonic Oscillations via the Euler relation

To illustrate the utility of complex exponentials for describing harmonic oscillators, consider the action of $z(t) = Ae^{i\omega t}$ as a function of time:

- $z(t = 0) = Ae^{i\cdot0} = A$
- $z\left[t = \frac{\pi}{4\omega} = \frac{T}{8}\right] = A \exp \left[i\frac{\pi}{4}\right] = A\left(\cos \left[\frac{\pi}{4}\right] + i \sin \left[\frac{\pi}{4}\right]\right) = A\left(\frac{1}{\sqrt{2}}\right)(1 + i)$
- $z\left[t = \frac{\pi}{2\omega} = \frac{T}{4}\right] = A \exp \left[i\frac{\pi}{2}\right] = A\left(\cos \left[\frac{\pi}{2}\right] + i \sin \left[\frac{\pi}{2}\right]\right) = A \cdot i$
- $z\left[t = \frac{\pi}{\omega} = \frac{T}{2}\right] = A \exp \left[i\pi\right] = A\left(\cos \left[\pi\right] + i \sin \left[\pi\right]\right) = -A$
- $z\left[t = \frac{3\pi}{2\omega} = \frac{3T}{4}\right] = A \exp \left[i\frac{3\pi}{2}\right] = A\left(\cos \left[\frac{3\pi}{2}\right] + i \sin \left[\frac{3\pi}{2}\right]\right) = A \cdot -1 = -A$

As $t$ increases, the complex function describes a circle of radius $A$ about the origin.

If the vector rotates in the direction of $+\phi$ with increasing time, then the oscillation frequency is positive; if the vector rotates in direction of $-\phi$ with increasing time, the frequency is negative. The temporal frequency is proportional to the rate of change of phase. In words, the faster the oscillation, the more rapidly the phase changes:

$$\omega = \frac{\partial \Phi [t]}{\partial t}$$

where $\Phi [t]$ is the phase of the complex function. Since the phase has dimensionless “units” of radians, its temporal derivative has dimensions of radians per unit time. The quantity $\omega$ is the angular temporal frequency. Since there are $2\pi$ radians per cycle, the angular temporal frequency may be converted to temporal frequency $\nu$ via:

$$\frac{\partial \Phi [t]}{\partial t} \left[\text{radians/sec}\right] \times \frac{1}{2\pi \left[\text{radians/cycle}\right]} = \frac{1}{2\pi} \frac{\partial \Phi [t]}{\partial t} \left[\text{cycles/sec}\right] \equiv \nu \left[\text{Hz}\right]$$

The temporal is proportional to the time derivative of the phase, which shows directly
that the temporal frequency $\nu_0$ is negative if the phase decreases with increasing time.

### 3.5 Oscillations as Projections of Circular Harmonic Motion

The sum of these two harmonic oscillations $\cos(\omega t)$ and $i \sin(\omega t)$ yields uniform circular motion. Because the sine term is imaginary, it is oriented at right angles to the (real) cosine term. The imaginary part of the motion can also be rewritten:

$$\sin(\omega t) = \cos \left[ \frac{\pi}{2} - \omega t \right] = \cos \left[ - \left( \frac{\pi}{2} - \omega t \right) \right] = \cos \left[ + \omega t - \frac{\pi}{2} \right] \quad \text{(because cosine is even)}$$

$$\implies y(t) = \cos(\omega t) + i \cos \left[ \omega t - \frac{\pi}{2} \right]$$

Thus: uniform circular motion results from the addition of two harmonic oscillations at right angles and with a phase difference of $\frac{\pi}{2}$ radians = 90°.

Conversely: the projection of uniform circular motion in any direction yields harmonic motion. The initial phase of the harmonic motion is determined by the azimuth of projection.

For example, when projecting onto the real axis, the information about variation along the imaginary axis is ignored:

$$\text{Re} \{ y(t) \} = \cos(\omega t).$$

Projection onto the imaginary axis discards information about variation along the real axis, and the result:

$$\text{Im} \{ y(t) \} = \sin(\omega t) = \cos \left[ \omega t - \frac{\pi}{2} \right]$$

### 3.6 Phasor Notation for Oscillations

The interpretation of harmonic motion as a projection of uniform circular motion leads to a third method for representing oscillations – the phasor. Its use is quite popular in electrical engineering applications.
The phasor with magnitude $A$ and phase $\Phi [t_0]$ is denoted by the polar vector $(A, \Phi [t_0])$ that describes the instantaneous position of the oscillator on the 2-D plot (Argand diagram). As time progresses, the phasor of an oscillator rotates with period $T = \frac{1}{\nu} = \frac{2\pi}{\omega}$. Generally, the phasor picture portrays the amplitude and phase of the oscillator at a particular time $t_0$ (generally $t_0 = 0$ seconds).

Since the phasors of same-frequency oscillators rotate at the same rate, their relative phase is invariant. Therefore, the phasor picture is useful for describing the relative amplitudes and phases of two or more oscillators with the same frequency. Also, it is useful for finding the resultant of the superposition of the same-frequency oscillators, as will be shown.

### 3.7 Superposition of Oscillations

H§7, §14

When two (or more) oscillations (or waves) are present at the same location in a medium at the same time, the resultant motion is (obviously) some combination of the two component oscillations (or waves). The simplest combination of the components (and the most common for electromagnetic oscillations or waves) is the superposition, or sum. When the principle of superposition holds, the response is said to be linear, i.e., the resultant $y [t]$ is the linear combination of the components $y_1 [t] + y_2 [t]$. The principle of superposition holds for acoustic and electromagnetic waves in most common situations (e.g., EM waves in a vacuum).

#### 3.7.1 Digression: Nonlinear Optics and Second-Harmonic Generation:

To help illustrate linear media and the principle of superposition, we will first consider an example where superposition is not valid. There are situations and media which can generate a resultant that is not a linear combination (weighted sum) of the components, but rather a power of the sum. Such a process has developed into the field of nonlinear optics. For example, if a two high-energy lasers are focused onto one of a class of crystals (such as quartz or potassium dihydrogen phosphate – KDP), the interaction with the crystal generates some emerging energy that is proportional
3.8 Superposition of Same-Frequency Oscillations

3.8.1 Trigonometric Notation:

Consider the linear superposition of two oscillations with the same frequency and different amplitudes and phases:

\[
y_1 [t] = A_1 \cos [\omega_0 t + \phi_1] \\
y_2 [t] = A_2 \cos [\omega_0 t + \phi_2] \\
\implies y [t] = y_1 [t] + y_2 [t]
\]
The trigonometric solution of the resultant $y[t]$ can be found as follows:

$$y_1 + y_2 = A_1 (\cos \omega_0 t \cos \phi_1 - \sin \omega_0 t \sin \phi_2) + A_2 (\cos \omega_0 t \cos \phi_2 - \sin \omega_0 t \sin \phi_2) = \cos \omega_0 t (A_1 \cos \phi_1 + A_2 \cos \phi_2) - \sin \omega_0 t (A_1 \sin \phi_1 + A_2 \sin \phi_2)$$

Since real parts add to real parts, etc., we can define the real and imaginary parts of the resultant:

$$\text{Re} \{(A, \phi)\} = A \cos \phi = A_1 \cos \phi_1 + A_2 \cos \phi_2$$
$$\text{Im} \{(A, \phi)\} = A \sin \phi = A_1 \sin \phi_1 + A_2 \sin \phi_2$$

The squared magnitude of the result is:

$$(A \sin \phi)^2 + (A \cos \phi)^2 = A^2 = (A_1 \sin \phi_1 + A_2 \sin \phi_2)^2 + (A_1 \cos \phi_1 + A_2 \cos \phi_2)^2$$

$$\implies A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2 \cos (\phi_1 - \phi_2)}$$

and phase:

$$\frac{A \sin \phi}{A \cos \phi} = \tan \phi = \frac{A_1 \sin \phi_1 + A_2 \sin \phi_2}{A_1 \cos \phi_1 + A_2 \cos \phi_2}$$

$$\implies \phi = \tan^{-1} \left[ \frac{A_1 \sin \phi_1 + A_2 \sin \phi_2}{A_1 \cos \phi_1 + A_2 \cos \phi_2} \right].$$

Consider some simple cases:

1. $A_1 = A_2, \phi_1 = \phi_2 \implies$ same amplitude, same phase:

$$A_1 \cos \omega_0 t + \phi_1 + A_1 \cos \omega_0 t + \phi_1 = 2A_1 \cos \omega_0 t + \phi_1 \implies A = 2A_1, \ \phi = \phi_1$$

$$A^2 = A_1^2 + A_2^2 + 2A_1A_2 \cos (\phi_1 - \phi_1) = 2A_1^2 + 2A_1^2 \cos (0) = 4A_1^2$$

$$\implies A = 2A_1$$

$$\tan \phi = \left[ \frac{2A_1 \sin \phi_1}{2A_1 \cos \phi_1} \right]$$

$$\implies \phi = \phi_1$$

Addition of two identical oscillations gives a resultant with twice the amplitude and the same phase, as expected.

2. $A_1 = A_2, \phi_2 = (\phi_1 - \pi) \implies$ same amplitude, phase difference of $\pi$ radians:

$$A^2 = A_1^2 + A_2^2 + 2A_1A_2 \cos [(\phi_1 - (\phi_1 - \pi)]$$

$$= 2A_1^2 + 2A_1^2 \cos [\pi] = 2A_1^2 - 2A_1^2 = 0$$

$$\implies A = 0$$
3.8 SUPERPOSITION OF SAME-FREQUENCY OSCILLATIONS

\[ \phi = \tan^{-1}[\pi] = \pm \infty, \text{ but } \phi \text{ is irrelevant since amplitude } A = 0 \]

The convention is that \( \phi = 0 \) if \( A = 0 \).

Addition of two oscillations with same amplitude but out of phase by \( \pm \pi \) radians gives \( A = 0 \), also as expected.

3. \( A_1 = A_2, \phi_2 = \phi_1 + \frac{\pi}{2} \implies \) same amplitude, phase difference of \( +\frac{\pi}{2} \) radians.

The resultant has magnitude:

\[ A^2 = A_1^2 + A_2^2 + 2A_1A_2 \cos \left[ \phi_1 - \phi_1 - \frac{\pi}{2} \right] \]

\[ = 2A_1^2 \left( 1 + \cos \left[ \frac{\pi}{2} \right] \right) = 2A_1^2 \implies A = \sqrt{2}A_1, \]

and phase:

\[ \tan \phi = \frac{A_1 \sin \left[ \phi_1 \right] + A_1 \sin \left( \phi_1 + \frac{\pi}{2} \right)}{A_1 \cos \left[ \phi_1 \right] + A_1 \cos \left( \phi_1 + \frac{\pi}{2} \right)} \]

Since: \( \cos \left[ \phi_1 + \frac{\pi}{2} \right] = -\sin \left[ \phi_1 \right], \sin \left[ \phi_1 + \frac{\pi}{2} \right] = \cos \left[ \phi_1 \right] \),

\[ \implies \tan \phi = \frac{\cos \left[ \phi_1 \right] + \sin \left[ \phi_1 \right]}{\cos \left[ \phi_1 \right] - \sin \left[ \phi_1 \right]} = \frac{\cos \left[ \phi_1 \right] \left( 1 + \tan \left[ \phi_1 \right] \right)}{\cos \left[ \phi_1 \right] \left( 1 - \tan \left[ \phi_1 \right] \right)} \]

\[ = \left[ \tan \left[ \frac{\pi}{4} \right] + \tan \phi_1 \right] \left[ 1 - \tan \left[ \frac{\pi}{4} \right] \tan \phi_1 \right], \text{ since } \tan \left[ \frac{\pi}{4} \right] = 1. \]

Now we cheat, from a table of trigonometric properties, we can find that:

\[ \tan \left[ \alpha + \beta \right] = \frac{\tan \alpha + \tan \beta}{1 - \tan \alpha \tan \beta} \]

which leads to the observation:

\[ \left[ \tan \left[ \frac{\pi}{4} \right] + \tan \left[ \phi_1 \right] \right] \left[ 1 - \tan \left[ \frac{\pi}{4} \right] \tan \left[ \phi_1 \right] \right] = \tan \left[ \phi_1 + \frac{\pi}{4} \right]. \]

and the phase of the resultant is:

\[ \phi = \phi_1 + \frac{\pi}{4} \text{ (if } \phi_1 = 0, \text{ then } \phi = +\frac{\pi}{4}) \]

If you add two oscillations with the same amplitude and a phase difference of \( + \), the resultant has the Pythagorean amplitude \( A = \sqrt{A_1^2 + A_2^2} \) and a phase angle midway between those of the components.

3.8.2 Phasor Representation:

Phasors are useful for computing the magnitude and phase resulting of the superposition (sum) of two (or more) oscillators with the same frequency. The resultant of
the superposition of two oscillators is the vector sum of the phasors defining the two oscillators:

\[ y_1 [t] = A_1 \sin [\omega_0 t + \phi_1] \equiv (A_1, \phi_1) \]
\[ y_2 [t] = A_2 \sin [\omega_0 t + \phi_2] \equiv (A_2, \phi_2). \]

The resultant phasor is \((A, \phi) = (A_1, \phi_1) + (A_2, \phi_2)\). The magnitude can be computed by adding the real and imaginary parts separately:

\[
\begin{align*}
\text{Re} \{A\} &= \text{Re} \{A_1\} + \text{Re} \{A_2\} = A_1 \cos [\phi_1] + A_2 \cos [\phi_2] \\
\text{Im} \{A\} &= \text{Im} \{A_1\} + \text{Im} \{A_2\} = A_1 \sin [\phi_1] + A_2 \sin [\phi_2]
\end{align*}
\]

Since the two oscillators have the same frequency \(\omega_0\), the relative phase of the two oscillators is invariant, and thus the relative initial phase is sufficient to compute the relative phase of the resultant.

\[ \phi = \tan^{-1} \left[ \frac{A_1 \sin [\phi_1] + A_2 \sin [\phi_2]}{A_1 \cos [\phi_1] + A_2 \cos [\phi_2]} \right]. \]

\textit{n.b.}, if the oscillators have different frequencies, the relative phase \(\Phi_1 [t] - \Phi_2 [t]\) varies with time and the phasor picture is not useful.

The magnitude also may be computed by using the law of cosines:

\[
\begin{align*}
\mathcal{A}^2 &= A_1^2 + A_2^2 - 2A_1A_2 \cos [\phi_1 - (\phi_2 - \pi)] \\
&= A_1^2 + A_2^2 - 2A_1A_2 \cos [\pi - (\phi_2 - \phi_1)] \\
&= A_1^2 + A_2^2 + 2A_1A_2 \cos [\phi_2 - \phi_1] \\
&= A_1^2 + A_2^2 + 2A_1A_2 \cos [\phi_1 - \phi_2],
\end{align*}
\]

where the last step follows because \(\cos [\theta] = \cos [-\theta]\).

Argand diagram of two phasors that rotate with the same angular frequency \(\omega_0\) shown at \(t = 0\). The phasors rotate about the origin at the same rate and therefore maintain the same relative angle.
3.9 SUPERPOSITION OF MANY SAME-FREQUENCY OSCILLATORS

3.8.3 Complex Notation:

Consider the complex representation of two oscillators with the same frequency $\omega_0$:

\[
y_1 [t] = A_1 e^{+i(\omega_0 t + \phi_1)} \\
y_2 [t] = A_2 e^{+i(\omega_0 t + \phi_2)}
\]

\[
y [t] \equiv A e^{+i(\omega_0 t + \phi)} = e^{+i\omega_0 t} A e^{+i\phi} \\
= y_1 [t] + y_2 [t] = A_1 e^{+i(\omega_0 t + \phi_1)} + A_2 e^{+i(\omega_0 t + \phi_2)} \\
= e^{+i\omega_0 t} [A_1 e^{+i\phi_1} + A_2 e^{+i\phi_2}] = e^{+i\omega_0 t} A e^{+i\phi} \\
\Rightarrow A e^{+i\phi} = A_1 e^{+i\phi_1} + A_2 e^{+i\phi_2}
\]

\[n.b., \text{ The resultant oscillation has the same frequency as the components frequency } \omega_0.\]

The last line represents the sum of two phasors: $(A_1, \phi_1)$ and $(A_2, \phi_2)$. This was solved on the previous page:

\[
A = \sqrt{A_1^2 + A_2^2 + 2A_1 A_2 \cos(\phi_1 - \phi_2)}
\]

\[
\tan[\phi] = \left[ \frac{A_1 \sin[\phi_1] + A_2 \sin[\phi_2]}{A_1 \cos[\phi_1] + A_2 \cos[\phi_2]} \right]
\]

**THE SUPERPOSITION OF TWO SAME-FREQUENCY OSCILLATIONS IS AN OSCILLATION OF THAT FREQUENCY**

(The principles of Fourier analysis can be applied to make this statement obvious.

3.9 Superposition of Many Same-Frequency Oscillators

Since the sum of two same-frequency oscillations is a harmonic oscillation of that frequency, clearly the sum of $N$ same-frequency oscillations must also be a harmonic oscillation of that frequency. This is easy to prove using complex notation:

\[
y_n [t] = A_n e^{+i(\omega_0 t + \phi_n)} \\
y [t] = \sum_{n=1}^{N} A_n e^{+i(\omega_0 t + \phi_n)} = e^{+i\omega_0 t} \sum_{n=1}^{N} A_n e^{+i\phi_n} \equiv e^{+i\omega_0 t} (A e^{i\Phi}).
\]

The resultant oscillation has amplitude $A$ and phase $\Phi$, and hence may be specified by the phasor $(A, \Phi)$. 

\[ A e^{+i\Phi} = \sum_{n=1}^{N} A_n e^{+i\phi_n} \]

\[
\text{Re}\{A e^{+i\Phi}\} = \text{Re}\left\{ \sum_{n=1}^{N} A_n e^{+i\phi_n} \right\} = \sum_{n=1}^{N} A_n \cos[\phi_n]
\]

\[
\text{Im}\{A e^{+i\Phi}\} = \text{Im}\left\{ \sum_{n=1}^{N} A_n e^{+i\phi_n} \right\} = \sum_{n=1}^{N} A_n \sin[\phi_n]
\]

By the Pythagorean theorem:

\[ A^2 = [\text{Re}\{A\}]^2 + [\text{Im}\{A\}]^2 = \left[ \sum_{n=1}^{N} A_n \cos[\phi_n] \right]^2 + \left[ \sum_{n=1}^{N} A_n \sin[\phi_n] \right]^2 \]

Look at the square of the real part:

\[
\left[ \sum_{n=1}^{N} A_n \cos[\phi_n] \right]^2 = \left[ \sum_{j=1}^{N} A_j \cos[\phi_j] \right] \cdot \left[ \sum_{k=1}^{N} A_k \cos[\phi_k] \right]
\]

\[ = \sum_{j=1}^{N} \sum_{k=1}^{N} A_j A_k \cos[\phi_j] \cos[\phi_k]\]

\[ = A_1^2 \cos^2[\phi_1] + A_1 A_2 \cos[\phi_1] \cos[\phi_2] + A_2 A_1 \cos[\phi_2] \cos[\phi_1] + A_2^2 \cos^2[\phi_2] + \cdots\]

\[ = A_1^2 \cos^2[\phi_1] + A_2^2 \cos^2[\phi_2] + \cdots + 2A_1 A_2 \cos[\phi_1] \cos[\phi_2] + 2A_2 A_3 \cos[\phi_2] \cos[\phi_3] + \cdots\]

This may be rewritten as the sum of the squared terms for \( j = k = 1, \ldots, N \) (which we will index by \( n = j = k \)) and the sum of the terms for which \( j \neq k \). The second set includes two identical terms, which may be combined by considering the values of \( j > k \):

\[
\left[ \sum_{n=1}^{N} A_n \cos[\phi_n] \right]^2 = \left[ \sum_{j=1}^{N} A_j \cos[\phi_j] \right] \left[ \sum_{k=1}^{N} A_k \cos[\phi_k] \right]
\]

\[ = \sum_{n=1}^{N} A_n^2 \cos^2[\phi_n] + \sum_{j \neq k}^{N} A_j A_k \cos[\phi_j] \cos[\phi_k]\]

\[ = \sum_{n=1}^{N} A_n^2 \cos^2[\phi_n] + 2 \sum_{j > k}^{N} A_j A_k \cos[\phi_j] \cos[\phi_k],\]

so that: \( j = [2, N], \ k = [1, N - 1] \)
3.10 Superposition of Randomly Phased Oscillators

The treatment for the imaginary part is identical:

\[
\left( \sum_{n=1}^{N} A_n \sin [\phi_n] \right)^2 = \sum_{j=1}^{N} A_j^2 \sin^2 [\phi_j] + \sum_{j \neq k} A_j A_k \sin [\phi_j] \sin [\phi_k]
\]

\[
= \sum_{j=1}^{N} A_j^2 \sin^2 [\phi_j] + 2 \sum_{j > k} A_j A_k \sin [\phi_j] \sin [\phi_k]
\]

Therefore the square of the resulting magnitude may be written as the sum of these four sums:

\[
A^2 = \left[ \sum_{n=1}^{N} A_n \cos [\phi_n] \right]^2 + \left[ \sum_{n=1}^{N} A_n \sin [\phi_n] \right]^2
\]

\[
+ 2 \sum_{j > k} A_j A_k \cos [\phi_j] \cos [\phi_k] + \sum_{j > k} A_j A_k \sin [\phi_j] \sin [\phi_k]
\]

\[
= \sum_{n=1}^{N} A_n^2 \left( \cos^2 [\phi_n] + \sin^2 [\phi_n] \right) + 2 \sum_{j > k} A_j A_k \left( \cos [\phi_j] \cos [\phi_k] + \sin [\phi_j] \sin [\phi_k] \right)
\]

Now apply the trigonometric identity \( \cos [\phi_j] \cos [\phi_k] + \sin [\phi_j] \sin [\phi_k] = \cos [\phi_j - \phi_k] \):

\[
A^2 = \sum_{n=1}^{N} A_n^2 + 2 \sum_{j > k} A_j A_k \cos [\phi_j - \phi_k]
\]

Since the phase angles are randomly distributed, the phase angle of the resultant is randomly distributed as well – therefore, no prediction of the phase can be made.

3.10 Superposition of Randomly Phased Oscillators

Special Case I: The oscillators have identical amplitudes (\( A_j = A_k = A_0 \)) and phases that are randomly distributed over the full domain of possible phase angles.

Random phases \( \iff [\phi_j] \) is randomly distributed in the interval \([0, 2\pi)\) (i.e., \( 0 \leq \phi < 2\pi \)) or equivalently in the interval \([-\pi, +\pi)\) (so that \(-\pi \leq \phi < +\pi\)).

\( \iff [\phi_j] - [\phi_k] \) is randomly distributed in \([-2\pi, 2\pi)\), and so is randomly distributed in \([0, 2\pi)\)

\( \iff \cos [\phi_j - \phi_k] \) is randomly distributed over the interval \([-1, 1]\)

\( \iff A^2 = \sum_{n=1}^{N} A_n^2 + 2 \cdot A_0^2 \sum_{j > k} \cos [\phi_j - \phi_k] \)

Since \( \cos [\phi_j - \phi_k] \) is randomly distributed over the interval \([-1, 1]\), we expect that
the sum of many terms likely will approximate zero:

\[
\sum_{j>k}^{N} \cos [\phi_j - \phi_k] \approx 0
\]

so that the amplitude of the resultant should be.

\[
A^2 = \sum_{n=1}^{N} A_0^2 = N \cdot A_0^2
\]

\[\Rightarrow A = (\sqrt{N}) A_0\]

In words, the total amplitude is \(\sqrt{N}\) times as large as the amplitude of the individual oscillators. The phase of the sum of the random-phase oscillators cannot be predicted, but can be any angle in the interval \([-\pi, +\pi]\). Recall that the energy of the oscillator is proportional to \(A^2\), so if the phases are random, the total energy is the sum of the individual energies, as expected. Randomly phased oscillators are said to be incoherent.

The results of summing \(N\) randomly phased oscillators with identical amplitudes for two cases, showing that the magnitudes of the resultants are both approximately equal to \(\sqrt{N}\).

3.11 Superposition of Nonrandomly Phased Oscillators

**Special Case II**: Amplitudes AND phases are equal, i.e., \(A_j = A_k = A_0\) and \([\phi_j] = [\phi_k] = \phi_0\)
3.11 Superposition of Nonrandomly Phased Oscillators

\[ I = A^2 = N \cdot A_0^2 + 2 \cdot A_0^2 \sum_{j=2}^{N} \cos(\phi_0 - \phi_j) \]

\[ = N \cdot A_0^2 + 2 \cdot A_0^2 \sum_{j=2}^{N} 1 \]

\[ = A_0^2(N + 2(N - 1)) \]

\[ = (3N - 2)A_0^2 \]

Examples:

\[ N = 1 \Rightarrow I = A_0^2 \equiv I_0, \text{ for one oscillator} \]

\[ N = 2 \Rightarrow I = 4A_0^2 = 4I_0 \Rightarrow 4 \times \text{energy of one oscillator} \]

\[ N = 3 \Rightarrow I = 7A_0^2 = 7I_0 \]

\[ N = 4 \Rightarrow I = 10A_0^2 = 10I_0 \]

n.b., \( I > NI_0 \), the intensity of the sum of \( N \) in-phase oscillators is larger than expected, i.e., the noise is louder, or the light is brighter. Of course, energy must be conserved, so if the signal is “louder” or “brighter” at some locations, it must be “less loud” or “dimmer” at other locations.

If the phase relationship between the component oscillators is well-defined, the oscillators are coherent.

Two examples of the sum of nonrandomly phased oscillators. In one case (shown in black), the sum yields a “null output” (resulting magnitude is 0). In the other case, the sum is nonzero.
CHAPTER 3 COMPLEX NUMBERS

3.12 Superposition of Oscillations with Different Frequencies

3.12.1 Complex Notation

H§7.5, HR §20

We have just seen that the superposition of any number of same-frequency oscillators is an oscillation with that frequency. When superposing a number of oscillators with different frequencies, the situation is quite different – (almost) any periodic function can be synthesized from the summation of harmonic terms. This is the principle of Fourier analysis.

Simple Example: Addition of two oscillators of same amplitude $A_0$, same phase $\phi_0$, different frequencies $\omega_1$ and $\omega_2$:

$$y[t] = y_1[t] + y_2[t] = A_0 \cos[\omega_1 t] + A_0 \cos[\omega_2 t]$$

$$= A_0 (\cos[\omega_1 t] + \cos[\omega_2 t])$$

$$= \text{Re} \left\{ A_0 \left[ e^{+i\omega_1 t} + e^{+i\omega_2 t} \right] \right\}$$

Note that we can do both the sum of the cosines and of the sines at the same time:

$$\Rightarrow e^{+i\omega_1 t} + e^{+i\omega_2 t} = \left( e^{+i\frac{\omega_1}{2} t} \cdot e^{+i\frac{\omega_1}{2} t} \right) \cdot 1 + \left( e^{+i\frac{\omega_2}{2} t} \cdot e^{+i\frac{\omega_2}{2} t} \right) \cdot 1$$

$$= \left( e^{+i\frac{\omega_1}{2} t} \cdot e^{+i\frac{\omega_1}{2} t} \right) \left( e^{+i\frac{\omega_1}{2} t} \cdot e^{-i\frac{\omega_1}{2} t} \right) + \left( e^{+i\frac{\omega_2}{2} t} \cdot e^{-i\frac{\omega_2}{2} t} \right) \left( e^{+i\frac{\omega_2}{2} t} \cdot e^{-i\frac{\omega_2}{2} t} \right)$$

$$= \left( e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t} \right) \left( e^{+i\left(\frac{\omega_1 - \omega_2}{2}\right) t} \right) + \left( e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t} \right) \left( e^{-i\left(\frac{\omega_1 - \omega_2}{2}\right) t} \right)$$

$$= \left( e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t} + e^{-i\left(\frac{\omega_1 - \omega_2}{2}\right) t} \right) \left( e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t} \right)$$

$$= 2 \cos\left(\frac{\omega_1 - \omega_2}{2}\right) \cdot e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t}$$

$$\Rightarrow y[t] = \text{Re} \left\{ A_0 \left[ e^{i\omega_1 t} + e^{i\omega_2 t} \right] \right\} + i \text{Im} \left\{ A_0 \left[ e^{i\omega_1 t} + e^{i\omega_2 t} \right] \right\}$$

$$A_0 (\cos[\omega_1 t] + \cos[\omega_2 t]) = A_0 \text{Re} \left\{ e^{i\omega_1 t} + e^{i\omega_2 t} \right\}$$

$$= A_0 \text{Re} \left\{ 2 \cos\left(\frac{\omega_1 - \omega_2}{2}\right) t \left( e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t} \right) \right\}$$

$$= 2A_0 \cos\left(\frac{\omega_1 - \omega_2}{2}\right) t \text{Re} \left\{ e^{+i\left(\frac{\omega_1 + \omega_2}{2}\right) t} \right\}$$

$$A_0 (\cos[\omega_1 t] + \cos[\omega_2 t]) = 2A_0 \cos\left(\frac{\omega_1 - \omega_2}{2}\right) t \cdot \cos\left(\frac{\omega_1 + \omega_2}{2}\right) t$$
Section 3.12: Superposition of Oscillations with Different Frequencies

\[ A_0 (\sin [\omega_1 t] + \sin [\omega_2 t]) = A_0 \text{Im} \left\{ e^{i\omega_1 t} + e^{i\omega_2 t} \right\} \]
\[ = A_0 \text{Im} \left\{ 2 \cos \left[ \left( \frac{\omega_1 - \omega_2}{2} \right) t \right] \left( e^{i\left(\frac{\omega_1 + \omega_2}{2}\right)t} \right) \right\} \]
\[ = 2A_0 \cos \left[ \left( \frac{\omega_1 - \omega_2}{2} \right) t \right] \text{Im} \left\{ e^{i\left(\frac{\omega_1 + \omega_2}{2}\right)t} \right\} \]

By defining an average and a modulation (angular) frequency:

\[ \Omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} \]
\[ \Omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2}, \]

We obtain the expression for the sum of two sinusoids with angular frequencies \( \omega_1 \) and \( \omega_2 \) as the product of two sinusoids with the average and modulation frequencies:

\[ y [t] = 2A_0 \cos [\Omega_{\text{avg}} t] \cdot \cos [\Omega_{\text{mod}} t] \]

In words, the sum of two harmonic oscillations with different frequencies \( \omega_1 \) and \( \omega_2 \) yields the product of two harmonic oscillations, one with the average frequency \( \Omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} \), and one with the so-called modulation frequency \( \Omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2} \).

Both the product and sum of different-frequency sinusoids yield results that are not harmonic. The former is equivalent to the sum of sinusoids at the sum and difference frequencies, while the sum is equivalent to the product of sinusoids at \( \Omega_{\text{avg}} \) and \( \Omega_{\text{mod}} \). The periods of the superposition are \( T_{\text{avg}} \) and \( T_{\text{mod}} \), where \( T_{\text{mod}} > T_{\text{avg}} \). The slower period \( T_{\text{mod}} \) is the source of the phenomenon known commonly as beats, from its musical context, though this kind of pattern is seen (heard?) in many other situations as well. Low-frequency Moiré fringes are seen when two periodic patterns are overlaid are examples. The phenomenon of aliasing in digital signal/image processing is closely related.

The converse is also true: the product of two periodic signals can be expressed as the sum of two other oscillations: the heterodyning operation in radio is an example. AM radio signals are broadcast at frequencies 560 kHz \( \leq \nu_1 \leq 1600 \) kHz. To render the signals audible, they are beat down by multiplying by an intermediate frequency (IF) \( \nu_2 \). Two signals result: one with frequency \( \nu_1 + \nu_2 \) and one with \( \nu_1 - \nu_2 \). Judicious choice of \( \nu_2 \) puts the lower-frequency sideband in the audible range. The upper sideband is removed by a filter which passes only low frequencies (low-pass filter).

**Example:**

Consider the product and sum of two harmonic oscillations with angular frequencies \( \nu_1 = \frac{1}{50} \) and \( \nu_2 = \frac{1}{60} \) cycles per unit length, so the corresponding temporal periods are \( T_1 = 50 \) and \( T_2 = 60 \). These are illustrated below:
CHAPTER 3 COMPLEX NUMBERS

36

Sum and product of oscillations: (a) \( f_1 [t] = \cos \left( \frac{2 \pi}{50} t \right) \), (b) \( f_2 [t] = \cos \left( \frac{2 \pi}{60} t \right) \), (c) \( f_1 [t] + f_2 [t] \), also showing modulation wave, (d) \( f_1 [t] \times f_2 [t] \), showing different-frequency wave.

The sum of these two oscillations is:

\[
\cos \left( 2 \pi \nu_1 t \right) + \cos \left( 2 \pi \nu_2 t \right) = 2 \cos \left[ 2\pi \left( \frac{\nu_1 + \nu_2}{2} \right) t \right] \cdot \cos \left[ 2\pi \left( \frac{\nu_1 - \nu_2}{2} \right) t \right] \\
= 2 \cos \left[ 2\pi \left( \frac{\frac{1}{50} + \frac{1}{60}}{2} \right) t \right] \cdot \cos \left[ 2\pi \left( \frac{\frac{1}{50} - \frac{1}{60}}{2} \right) t \right] \\
\approx 2 \cos \left[ 2\pi \left( \frac{t}{54.545} \right) \right] \cdot \cos \left[ 2\pi \left( \frac{t}{600} \right) \right] \\
\nu_{avg} \approx \frac{1}{54.545} \\
\nu_{mod} = \frac{1}{600}
\]

The “slowly” varying term with period 600 is generally more visible.

The product of these two sinusoids may be written as the scaled sum of sinusoids at the sum and difference frequencies, where the former oscillates at a rapid rate and
the latter oscillates more slowly:

\[
\cos [2\pi \nu_1 t] \cdot \cos [2\pi \nu_2 t] = \frac{1}{2} \cos [2\pi (\nu_1 + \nu_2) t] + \frac{1}{2} \cos [2\pi (\nu_1 - \nu_2) t]
\]

\[
= \frac{1}{2} \cos [2\pi (\nu_1 + \nu_2) t] + \frac{1}{2} \cos [2\pi (\nu_1 - \nu_2) t]
\]

\[
= \frac{1}{2} \left( \cos \left[ \frac{2\pi t}{27.27} \right] + \cos \left[ \frac{2\pi t}{300} \right] \right)
\]

\[
\nu_1 + \nu_2 = \frac{1}{27.27} \text{ Hz}
\]

\[
\nu_1 - \nu_2 = \frac{1}{300} \text{ Hz}
\]

### 3.13 Introduction to Fourier Analysis

The motion resulting from the sum of two oscillations of different frequency is complex (i.e., anharmonic) though still periodic since it repeats after a time defined by:

\[
T_{\text{mod}} = \frac{1}{\nu_{\text{mod}}} = \frac{2\pi}{\omega_{\text{mod}}} = \frac{4\pi}{\omega_1 - \omega_2}
\]

As \( \omega_1 \to \omega_2 \), \( T_{\text{mod}} \) lengthens. In the limit, \( T_{\text{mod}} \to \infty \) and \( T_{\text{avg}} \to T_1 = T_2 \).

The addition of more oscillations of different frequencies produces more and more complex motion (less like harmonic motion). For example, consider this sum of harmonic oscillators:

\[
y[t] = \sum_{n=1,3,5,...}^{\infty} \left( \pm \frac{1}{n} \cos [n\omega_0 t] \right)
\]

For each succeeding term, the amplitude decreases and the frequency increases. The first term (the fundamental frequency component) is:

\[
f_1[t] = \cos \left[ 2\pi \frac{t}{8} \right]
\]

\[
f_2[t] = 0
\]

\[
f_3[t] = -\frac{1}{3} \cos \left[ 2\pi \frac{t}{(\frac{8}{3})} \right]
\]

\[
f_4[t] = 0
\]

\[
f_5[t] = +\frac{1}{5} \cos \left[ 2\pi \frac{t}{(\frac{8}{5})} \right]
\]

Obviously, \( y[t] \) is becoming less and less harmonic as more terms are added, and in fact is starting to look like a completely different function – a square wave. Especially note that as higher frequency components are added (i.e., larger values of \( n \)), the verticals become “steeper” and the edges become “sharper.” Note also that the summation overshoots the final result where the function transitions from horizontal.
to vertical and vice versa. This overshooting is the Gibbs phenomenon, and its visibility diminishes as more terms are added. As $N \to \infty$, the function $y[t]$ becomes a periodic square wave, which is quite dissimilar from the component functions.

This result illustrates the principle of Fourier Analysis, where we determine the set of sinusoidal constituents that sum to create the function $f[t]$. The complementary operation of Fourier synthesis sums up a set of sinusoids to find the resultant.

\[ \text{Sum of sinusoids with specific different magnitudes and frequencies to produce a square wave.} \]

\( (\text{Virtually}) \) every periodic function may be decomposed into a sum of sines and cosines with definite amplitudes, frequencies, and phases. The decomposition is unique, and is called the Fourier series representation, or the spectrum of the periodic function.

The spectrum is a representation of the amplitudes, frequencies, and phases of the sinusoidal components that superpose to create the function. Often, the term spectrum is used when power spectrum would be more accurate – it is the power (or energy, the squared magnitude of the component) that is plotted rather than the amplitude.

This concept should be quite familiar to you – the spectrum of white light is analogous. White light is a periodic function – it looks the same at all times. Spherical rain droplets act as prisms to disperse white light into its constituent components – the colors of the spectrum. The brightnesses of each color correspond to the energy of the component – brighter $\implies$ more energy. The droplet prisms act as Fourier transformers since they derive the spectrum of the function As Newton showed, the spectrum can be transformed back to white light with another prism.

3.13.1 Fourier Transform

The mathematical operation that determines the amplitudes and phases of the constituent sinusoids of an arbitrary temporal function $f[t]$ (or spatial function $f[x]$) is
3.13 INTRODUCTION TO FOURIER ANALYSIS

the Fourier transform, which “projects” the input function onto each of the possible (complex-valued) sinusoidal functions \( \exp [+2\pi ivt] \) (or \( \exp [+2\pi i\xi x] \)). The process of “projection” one function onto another is exactly analogous to the operation of projecting one vector onto another via the scalar (or dot) product. Recall that the scalar product of two N-component vectors \( \mathbf{a} \) and \( \mathbf{b} \) is:

\[
\mathbf{a} \cdot \mathbf{b} = \sum_{n=0}^{N-1} (a)_n (b)_n = \sum_{n=0}^{N-1} a_n b_n = |\mathbf{a}| \ |\mathbf{b}| \ \cos [\theta]
\]

where \( \theta \) is the angle between the vectors. One of these vectors is the “reference” vector (the vector being “projected onto” – call it the first vector \( \mathbf{a} \)) and one is the “input” vector (call it the second vector \( \mathbf{b} \)). The result is a scalar, and that scalar is the projection of \( \mathbf{b} \) onto \( \mathbf{a} \) if \( \mathbf{a} \) has unit length. This can be seen from the second expression for the scalar product:

\[
\frac{\mathbf{a}}{|\mathbf{a}|} \cdot \mathbf{b} \equiv \hat{\mathbf{a}} \cdot \mathbf{b} = |\mathbf{b}| \ \cos [\theta]
\]

where \( \hat{\mathbf{a}} \) is the unit vector (vector with unit length) that points in the direction of \( \mathbf{a} \). In short, the scalar product is equal to the projection if the reference vector \( \mathbf{a} \) has unit length.

If the vectors have complex-valued components, then the scalar product is modified (to become the “inner product”) to produce the real-valued length if a vector is projected onto itself. To do this, the components of one of the vectors must be complex-conjugated:

\[
\mathbf{a} \cdot \mathbf{a} = \sum_{n=0}^{N-1} a^*_n a_n = |\mathbf{a}| \ |\mathbf{a}| \ \cos [\theta - \theta] = |\mathbf{a}|^2 \geq 0
\]

Thus the inner product of two vectors is:

\[
\mathbf{a} \cdot \mathbf{b} = \sum_{n=0}^{N-1} a^*_n b_n = c
\]

where \( c \) is generally complex-valued, but is guaranteed to be real valued if \( \mathbf{b} = \mathbf{a} \). Thus the projection of “input vector” \( \mathbf{b} \) onto “reference vector” \( \mathbf{a} \) is:

\[
\hat{\mathbf{a}} \cdot \mathbf{b} = \sum_{n=0}^{N-1} (\hat{\mathbf{a}})_n^* (\mathbf{b})_n = \sum_{n=0}^{N-1} \hat{a}^*_n b_n = c
\]

Now generalize this expression from vectors to continuous functions, say \( f[t] \) and \( a[t] \); the projection of the complex-valued “input function” \( f[t] \) onto the complex-valued “reference function” \( a[t] \) is obtained by computing the complex conjugate of the reference, multiplying it by the input function “point by point”, and then
summing the result in an integral:

\[
\int_{-\infty}^{+\infty} (a [t])^* f [t] \ dt = \mathcal{O} \{ f [t] \} = g
\]

where \( g \) is a complex number that defines the “projection” of \( f [t] \) onto \( a [t] \).

Now repeat this operation for a family of reference functions that differ by some parameter (call it “\( \nu \)”: the output values will differ depending on the value of \( \nu \) and thus produces an output function of \( \nu \):

\[
\int_{-\infty}^{+\infty} (a [t; \nu])^* f [t] \ dt = \mathcal{O} \{ f [t] \} = g [\nu]
\]

In the Fourier transform, the reference functions are the complex-valued sinusoidal functions \( \exp [+2\pi i \nu t] \):

\[
\mathcal{F} \{ f [t] \} = \int_{-\infty}^{+\infty} (\exp [+2\pi i \nu t])^* f [t] \ dt
\]

\[
= \int_{-\infty}^{+\infty} f [t] \exp [-2\pi i \nu t] \ dt \equiv F [\nu]
\]

In words, the Fourier transform evaluates the projection of the input function \( f [t] \) onto each of the infinite set of complex sinusoidal functions \( \exp [+2\pi i \nu t] = \cos [2\pi \nu t] + i \sin [2\pi \nu t] \).
Chapter 4

Traveling Waves

W §2, F §1

To date, we have considered oscillations, i.e., periodic and sometimes harmonic, variations of a physical characteristic of a system. The oscillatory system at one time is indistinguishable if observed at a later time where the time difference is an integer number of temporal periods. To maintain oscillatory behavior, the energy of the oscillator must remain within the system and cannot be lost to the outside world. We now extend this picture to oscillations that travel from the source and thus transport energy to a different location. Energy must be continually added to the system to maintain the oscillation and the transported energy can do work on other systems at a distance.

We can define a traveling wave as “a self sustaining disturbance of the medium through which it propagates,” though (as we shall see) sometimes the entity that can be called the “medium” is not so obvious. At this point, we will ignore this problem.

Our first task is to mathematically describe a traveling harmonic wave, i.e., denote a $y [t]$ that travels through space. A harmonic oscillation $y(t) = A_0 \cos[\omega_0 t]$, can be converted into a traveling wave by making the phase a function of both $x$ and $t$ in a very particular way. Consider the general case of an oscillatory function of space and time:

$$y [z, t] = A_0 \cos[\Phi [z, t]].$$

We want this oscillation to move through space, e.g., toward positive $z$. In other words, if a point of constant phase on the wave (e.g., a peak of the cosine created at a particular time $\tau$) is at a point $x_0$ in space at a time $t_0$, the same point of constant phase must move to $z_1 > z_0$ at time $t_1 > t_0$. 
“Snapshots” of sinusoidal wave at two different times $t_0$ and $t_1 > t_0$, showing motion of the peak originally at the origin at $t_0$. The wave is traveling towards $z = +\infty$ at velocity $v_0$. The phase of the first wave at the origin is $0$ radians, but that of the second is negative.

Since the wave at location $z_1$ and time $t_1$ has the same phase as the wave at location $z_0$ and time $t_0$, we can say that:

$$
\Phi[z_0, t_0] = \Phi[z_1, t_1] \implies \cos[z_0, t_0] = \cos[z_1, t_1] \implies y[z_0, t_0] = y[z_1, t_1].
$$

In addition, for the wave to maintain its shape, the phase $\Phi[z, t]$ must be a linear function of $x$ and $t$; otherwise the wave would compress or stretch out at different locations in space or time. Therefore:

$$
\Phi[z, t] = \alpha z + \beta t
\implies \alpha z_0 + \beta t_0 = \alpha z_1 + \beta t_1.
$$

As discussed, if $t_1 > t_0 \implies z_1 > z_0$ (i.e., wave moves toward $z = +\infty$), then $\alpha$ and $\beta$ must have opposite algebraic signs:

$$
\Phi[z, t] = \pm |\alpha| z \mp |\beta| t = \pm (|\alpha| z - |\beta| t)
$$

By dimensional analysis, we know that $|\alpha| z - |\beta| t$ has the “dimensionless dimensions” of angle [i.e., measured in the unitless quantity of radians]. We have already identified $\beta = \omega_0$, the angular frequency of the oscillation. Similarly, if $z$ is measured in mm, then $\alpha$ must have dimensions of radians per mm, i.e., $\alpha$ specifies the number of radians of oscillation per unit length, and thus is the angular spatial frequency of the wave, commonly denoted by $k$:

$$
y_+ [z, t] = A_0 \cos[k_0 z - \omega_0 t] \text{ traveling harmonic wave toward } z = +\infty
$$

By identical analysis, we can derive the equation for a harmonic wave moving toward $x = -\infty$. 
The waves are functions of both space and time, i.e., three dimensions \([z, y, t]\) are needed to portray them. Generally we display \(y\) either as a function of \(z\) for fixed \(t\), or as a function of \(t\) for fixed \(z\):

\[
y_{z}[z, t] = A_0 \cos[k_0 z + \omega_0 t] \rightarrow \text{traveling harmonic wave toward } z = -\infty
\]

Traveling wave as (a) function of \(z\) viewed at two different times and (b) function of \(t\) viewed at two different locations. The wave travels a distance \(\Delta z\) in the time interval \(\Delta t\), so its phase velocity is \(\frac{\Delta z}{\Delta t}\).

### 4.1 2-D Plot of 1-D Traveling Wave

The 1-D traveling wave is a function of two variables: the position \(z\) and the time \(t\), and so may be graphed on axes with these labels. An example is shown in the figure, where \(z\) is plotted on the horizontal axis and \(t\) on the vertical axis. In this case, the point at the origin at \(t = 0\) has a phase of 0 radians. That point moves in the positive \(z\) direction with increasing time, and so is a wave of the form

\[
y[z, t] = \cos[k_0 z - \omega_0 t]
\]

The points with the same phase of 0 radians at later times are positioned along the line shown. The velocity of this point of constant phase is \(\frac{\Delta z}{\Delta t}\), and thus is the reciprocal of the slope of this line.
4.2 Notation and Dimensions for Waves in a Medium

Trigonometric Notation:

$$y[z,t] - y_0 = A_0 \cos \{\Phi[z,t]\} = A_0 \cos(k_0 z \pm \omega_0 t + \phi_0)$$

Complex Notation:

$$y[z,t] = A_0 e^{+i\Phi[z,t]} = \text{Re} \left\{ A_0 e^{+i(k_0 z \pm \omega_0 t + \phi_0)} \right\}$$

1. $y$ is the coordinate (or position) of the oscillating quantity of the system, (e.g., angle, voltage, etc.)

2. $y_0$ is the equilibrium value of the quantity;

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 $y$: $[A] = [y]$;

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 $[\omega_0] = \text{radians per second}$;

6. $\nu_0$ is the temporal frequency of the oscillation, units of $[\nu_0] = \text{cycles per second} = \text{Hertz}$ (Hz), $\nu_0 = \frac{\omega_0}{2\pi}$;

7. $T_0 = \frac{1}{\nu} = \frac{2\pi}{\omega_0}$ 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, $k_0 = \frac{2\pi}{\lambda_0}$, $[k_0] = \text{radians per mm}$;

10. $\Phi$ is the phase angle of the oscillation (the argument of the sinusoid), measured in radians;

11. $\phi_0 = \text{initial phase angle}$ of the wave, i.e., phase angle at $t = 0$ and $z = 0$, $[\phi_0] = \text{radians}$. 
4.3 THE WAVE EQUATION

12. \( \sigma_0 = \frac{1}{\lambda_0} \) is the wavenumber, the number of wavelengths per unit length, \( [\sigma_0] = \text{mm}^{-1} \).

Relations between the phase and the temporal frequencies

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

4.3 The Wave Equation

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

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

The motion of a traveling wave also satisfies a differential equation, but now the position is a function of both spatial and temporal coordinates. To avoid confusion between the displacement and the \( y \)-coordinate, we will rename the coordinate \( \psi \), so in the case of a single spatial dimension, the position is \( \psi [z, t] \) and in the 3-D case, the position is \( \psi [x, y, z, t] = \psi [r, t] \). For simplicity, we will consider the 1-D spatial case in this section.

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

\[
\psi [z_0, t_0] = \psi [z_1, t_1]
\]

In other words, the traveling wave must have the same “shape” at a different location and at specific different times. This means that the derivative of the function \( \psi [z, t] \) must be zero if evaluated at the proper values of \( [z, t] \):

\[
\begin{align*}
\frac{d\psi}{dt} &= 0 \\
&= \frac{\partial \psi}{\partial z} \frac{dz}{dt} + \frac{\partial \psi}{\partial t} \\
&= \frac{\partial \psi}{\partial (z - v_\phi t)} \frac{\partial (z - v_\phi t)}{dt} \frac{dz}{dt} + \frac{\partial \psi}{\partial (z - v_\phi t)} \frac{\partial (z - v_\phi t)}{dt} \\
&= \frac{\partial \psi}{\partial (z - v_\phi t)} \left( \frac{\partial (z - v_\phi t)}{dt} \frac{dz}{dt} + \frac{\partial (z - v_\phi t)}{dt} \right) \\
&= \frac{\partial \psi}{\partial (z - v_\phi t)} \left( 1 \cdot \frac{dz}{dt} + (-v_\phi) \right) \\
&= \frac{\partial \psi}{\partial (z - v_\phi t)} \cdot \left( 1 \cdot \frac{dz}{dt} - v_\phi \right) = 0 \implies \frac{dz}{dt} = v_\phi
\end{align*}
\]

Therefore, the amplitude \( \psi \) is the the same at coordinate pairs \( [z_1, t_1] \) and \( [z_2, t_2] \),
where \( z_1 = z_0 - v_\phi (t_1 - t_0) \), if \( v_\phi \) is the velocity of the wave. Thus we can generate a traveling wave from a sinusoidal function of time by substituting \( t - \frac{z}{v_\phi} \) for the time coordinate. A sinusoidal wave traveling toward \( z = +\infty \) is obtained from the cosine function:

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

where the quantity \( \frac{\omega_0}{v_\phi} \) has dimensions of \( \text{radians} \cdot \frac{z}{m} = \text{radians} \cdot \frac{\text{m}}{\text{s}} \), which is appropriate for the angular spatial frequency that we designate by \( k_0 \):

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

Now evaluate the partial derivatives with respect to the two coordinates:

\[
\frac{\partial \psi}{\partial z} = -k_0 A_0 \sin [k_0 z - \omega_0 t]
\]

\[
= -k_0 A_0 \cos \left( (k_0 z - \omega_0 t) - \frac{\pi}{2} \right)
\]

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

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

\[
\Rightarrow \psi [z, t] = -\frac{1}{k_0^2} \frac{\partial^2 \psi}{\partial z^2}
\]

\[
\frac{\partial \psi}{\partial t} = -(-\omega_0) A_0 \sin [k_0 z - \omega_0 t] = +\omega_0 A_0 \sin [k_0 z - \omega_0 t]
\]

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

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

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

By combining these two equations, we see that:

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

\[
\Rightarrow \frac{\partial^2 \psi}{\partial z^2} = \left( \frac{k_0}{\omega_0} \right)^2 \cdot \frac{\partial^2 \psi}{\partial t^2} = \frac{1}{\left( \frac{\omega_0}{k_0} \right)^2} \cdot \frac{\partial^2 \psi}{\partial t^2}
\]
The dimensions of the scale factor are:

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

which is the velocity of the same point of constant phase on the traveling wave, what we have called \( v_\phi \). Thus the 1-D wave equation is:

\[ \frac{\partial^2 \psi}{\partial z^2} = \frac{1}{v_\phi^2} \frac{\partial^2 \psi}{\partial t^2} \]

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

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

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

The sum of the second-order partial derivatives in the 2-D and 3-D cases are both called the Laplacian, commonly denoted \( \nabla^2 \):

2-D:
\[ \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = \nabla^2 \psi \]

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

### 4.4 Velocity of Traveling Waves

As just mentioned, the phase velocity \( v_\phi \) of a wave is the speed of travel of a point of constant phase. A definition for phase velocity can be derived by dimensional analysis: 

\[ [v_\phi] = \text{mm per s}; [\omega_0] = \text{radians per s}; [k] = \text{radians per mm} \]

\[ \implies \left[ \frac{\omega_0}{k_0} \right] = \frac{\text{radians per second}}{\text{radians per mm}} = \frac{\text{radian-mm}}{\text{radian-s}} = \frac{\text{mm}}{s} \]
Slightly more rigorously, we can find the phase velocity of a wave by taking derivatives of the equation for the wave:

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

\[ \frac{\partial y}{\partial t} = -(-\omega_0)A_0 \cdot \sin \left[ k_0 z - \omega_0 t + \phi_0 \right] = +A_0 \omega_0 \cdot \sin \left[ k_0 z - \omega_0 t + \phi_0 \right], \]

\[ \frac{\partial y}{\partial z} = -(k_0)A_0 \cdot \sin \left[ k_0 z - \omega_0 t + \phi_0 \right] = -A_0 k_0 \cdot \sin \left[ k_0 z - \omega_0 t + \phi_0 \right] \]

\[ v_\phi = \left| \frac{\partial z}{\partial t} \right| = \left( \frac{\partial z}{\partial t} \right) = \left| \frac{-\omega_0}{k_0} \right| = \frac{\omega_0}{k_0}, \]

or by considering the point of constant phase \( b \) radians:

\[ k_0 z - \omega_0 t = b \implies z = \left( \frac{b}{k_0} \right) + \frac{\omega_0}{k_0} t = b' + \left( \frac{\omega_0}{k_0} \right) t \quad b' \equiv \frac{b}{k_0} \text{ is a new constant} \]

Consider the positions \( z_1 \) and \( z_2 \) of the same point of constant phase at different times \( t_1 \) and \( t_2 \):

\[ z_1 = b' + \left( \frac{\omega_0}{k_0} \right) t_1 \]

\[ z_2 = b' + \left( \frac{\omega_0}{k_0} \right) t_2 \]

\[ \implies z_1 - z_2 = \Delta z = \left( \frac{\omega_0}{k_0} \right) (t_1 - t_2) = \left( \frac{\omega_0}{k_0} \right) \Delta t \]

\[ v_\phi \equiv \frac{\Delta z}{\Delta t} = \frac{\omega_0}{k_0} = v_\phi. \]

### 4.5 Superposition of Two Traveling Waves

Consider the superposition of two traveling waves with the same amplitude, different phase velocities, and different frequencies:

\[ y_1 [z, t] = A_0 \cos \left[ k_1 z - \omega_1 t \right] \]

\[ y_2 [z, t] = A_0 \cos \left[ k_2 z - \omega_2 t \right]. \]

We can use the same derivation developed for oscillations by defining a new frequency for both:

\[ \Omega_1 \equiv \frac{k_1 z}{t} - \omega_1 \]

\[ \Omega_2 \equiv \frac{k_2 z}{t} - \omega_2 \]
4.6 STANDING WAVES

\[ y[z, t] = y_1[z, t] + y_2[z, t] = A_0 \{ \cos[k_1z - \omega_1t] + \cos[k_2z - \omega_2t] \} \]
\[ = A_0 \left\{ \cos \left( \frac{k_1z}{t} - \omega_1 \right) \right\} + \cos \left( \frac{k_2z}{t} - \omega_2 \right) \]
\[ = A_0 \{ \cos[\Omega_1t] + \cos[\Omega_2t] \} \]
\[ = 2A_0 \cos \left( \frac{\Omega_1 + \Omega_2}{2} \right) \cdot \cos \left( \frac{\Omega_1 - \Omega_2}{2} \right) \]

just as before. By evaluating the sum and difference frequencies, we obtain:

\[ \left( \frac{\Omega_1 + \Omega_2}{2} \right) t = \left( \frac{k_1z}{t} - \omega_1 + \frac{k_2z}{t} - \omega_2 \right) t \]
\[ \text{where } k_{avg} = \frac{k_1 + k_2}{2}, \quad \omega_{avg} = \frac{\omega_1 + \omega_2}{2} \]

\[ \left( \frac{\Omega_1 - \Omega_2}{2} \right) t = \left( \frac{k_1z}{t} - \omega_1 - \frac{k_2z}{t} + \omega_2 \right) t \]
\[ \text{where } k_{mod} = \frac{k_1 - k_2}{2}, \quad \omega_{mod} = \frac{\omega_1 - \omega_2}{2} \]

4.6 Standing Waves

The superposition of two waves with the same amplitude \( A_0 \), temporal frequency \( \nu_0 \), and wavelength \( \lambda_0 \), but traveling in opposite directions results when one traveling wave is reflected at an interface. The summation yields the product of the average and modulation waves, where the time dependence cancels in the former and the space dependence in the latter:

\[ f_1[z, t] + f_2[z, t] = A_0 \cos[k_0z - \omega_0t] + A_0 \cos[k_0z + \omega_0t] \]
\[ = 2A_0 \cos \left[ \frac{k_0z - \omega_0t}{2} + \frac{k_0z + \omega_0t}{2} \right] \cdot \cos \left[ \frac{k_0z - \omega_0t}{2} - \frac{k_0z + \omega_0t}{2} \right] \]
\[ = 2A_0 \cos \left[ \frac{k_0z + k_0z}{2} + \frac{-\omega_0t + \omega_0t}{2} \right] \cdot \cos \left[ \frac{k_0z - k_0z}{2} + \frac{-\omega_0t - \omega_0t}{2} \right] \]
\[ = 2A_0 \cos [k_0z] \cdot \cos [-\omega_0t] \]
\[ = 2A_0 \cos [k_0z] \cdot \cos [\omega_0t], \quad \text{because } \cos [-\theta] = + \cos [+\theta] \]
\[ = 2A_0 \cos \left[ 2\pi \frac{z}{\lambda_0} \right] \cdot \cos [2\pi \nu_0t] \]

In words, the sum of these two waves yields the product of a spatial wave (with wavelength \( \lambda_0 \) and no time dependence) and a temporal oscillation with frequency \( \nu_0 \)
and no spatial dependence. In other words, the summation wave does not travel; its extrema do not move, but stay in the same locations while oscillating in time.

Standing waves produced by the sum of waves traveling in opposite directions, shown as functions of the spatial coordinate at five different times. The sum is a spatial wave whose amplitude oscillates.

### 4.7 Anharmonic Traveling Waves, Dispersion

Thus far the only traveling waves we have considered have been harmonic, i.e., consisting of a single sinusoidal frequency. From the principle of Fourier analysis, an anharmonic traveling wave can be decomposed into a sum of traveling harmonic wave components, i.e., waves of generally differing amplitudes over a discrete set of frequencies:

\[
y[z, t] = \sum_{\ell=1}^{\infty} y_\ell = \sum_{\ell=1}^{\infty} A_\ell \cos \left[ k_\ell z - \omega_\ell t + \phi_\ell \right],
\]

where \( A_\ell, k_\ell, \) and \( \omega_\ell \) are the amplitude, angular spatial frequency, and angular spatial frequency of the \( \ell^{th} \) wave. Therefore, we can define the phase velocity of the \( \ell^{th} \) wave as:

\[
(v_\phi)_\ell = \frac{\omega_\ell}{k_\ell}.
\]

Now suppose that a particular anharmonic oscillation is composed of two harmonic components \( y[x, t] = y_1(x, y) + y_2[x, t] \). If the two components have the same phase velocity, \( (v_\phi)_1 = (v_\phi)_2 \), then points of constant phase on the two waves move with the same speed and maintain the same relative phase. The shape of the resultant wave is invariant over time. Such a wave is called **nondispersive**, because points of constant phase on the components do not separate over time.

If the phase velocities are different so that \( (v_\phi)_1 \neq (v_\phi)_2 \), the points of constant phase on the two waves move at different velocities and therefore the distance between
specific points of constant phase on the two component waves changes as a function of position or time. This ensures that the shape of the superposition wave will change as a function of time; these waves are \textit{dispersive}.

Note that the \textit{dispersion} is a characteristic of the \textit{medium} within which the waves travel, and not of the waves themselves. It is the medium that determines the velocities and thus whether the waves travel together or if they disperse with time and space.

### 4.8 Average Velocity and Modulation (“Group”) Velocity

We added two traveling waves of different frequencies and obtained the same result we saw when adding two oscillations: the sum of two harmonic waves yields the product of two harmonic waves with modulation and average spatial and temporal frequencies. Using the new terms: $k_{avg}, k_{mod}, \omega_{avg},$ and $\omega_{mod},$ we can define the phase velocities of the average and modulation waves:

$$v_{avg} = \frac{\omega_{avg}}{k_{avg}} = \frac{\omega_1 + \omega_2}{\left(\frac{k_1 + k_2}{2}\right)} = \frac{\omega_1 + \omega_2}{k_1 + k_2}$$

$$v_{mod} = \frac{\omega_{mod}}{k_{mod}} = \frac{\omega_1 - \omega_2}{\left(\frac{k_1 - k_2}{2}\right)} = \frac{\omega_1 - \omega_2}{k_1 - k_2}$$

These two velocities have the same meaning as the phase velocity of the single wave, i.e., it is the velocity of a point of constant phase of the average traveling wave frequency or of the modulation wave frequency, or beats wave. The modulation velocity is also often called the \textit{group velocity}.

#### 4.8.1 Example: Nondispersive Waves $(v_\phi)_1 = (v_\phi)_2$

In a nondispersive medium, the phase velocity is constant over frequency (or wavelength), i.e.,

$$(v_\phi)_1 = \frac{\omega_1}{k_1} = (v_\phi)_2 = \frac{\omega_2}{k_2}.$$ 

Note that $\omega_1 \neq \omega_2$ and $k_1 \neq k_2$ – only the ratios are equal. Now find expressions for $v_{mod}$ and $v_{avg}$.

$$v_{avg} = \frac{\omega_{avg}}{k_{avg}} = \frac{\omega_1 + \omega_2}{\left(\frac{k_1 + k_2}{2}\right)} = \frac{\omega_1 + \omega_2}{k_1 + k_2} = \frac{\omega_1 \left(1 + \frac{\omega_2}{\omega_1}\right)}{k_1 \left(1 + \frac{k_2}{k_1}\right)}$$
Since $\frac{\omega_1}{k_1} = \frac{\omega_2}{k_2}$ for nondispersive waves $\implies \frac{\omega_2}{\omega_1} = \frac{k_2}{k_1}$ and:

$$v_{\text{avg}} = \frac{\omega_1}{k_1} \frac{1 + \frac{k_2}{k_1}}{1 + \frac{k_2}{k_1}} = \frac{\omega_1}{k_1} = \frac{v_1 = v_2 = v_{\text{avg}}}{v_{\text{avg}}}$$

Similarly for the velocity of the modulation wave:

$$v_{\text{mod}} = \frac{\omega_{\text{mod}}}{k_{\text{mod}}} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\omega_1 \left(1 - \frac{\omega_2}{\omega_1}\right)}{k_1 \left(1 - \frac{k_2}{k_1}\right)}$$

Since $\frac{\omega_1}{k_1} = \frac{\omega_2}{k_2}$ for nondispersive waves, then $\frac{\omega_2}{\omega_1} = \frac{k_2}{k_1}$ and:

$$v_{\text{mod}} = \frac{\omega_1}{k_1} \frac{1 - \frac{k_2}{k_1}}{1 - \frac{k_2}{k_1}} = \frac{\omega_1}{k_1} = \frac{v_1 = v_2 = v_{\text{mod}} = v_{\text{avg}}}{v_{\text{avg}}}$$

Note also that:

$$\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\Delta\omega}{\Delta k} \implies \frac{d\omega}{dk} = \omega_{\text{mod}}$$

In a nondispersive medium, all four waves in the summation (the individual waves and the modulation and average waves) all travel with the same velocity.

### 4.9 Dispersion Relation for Nondispersive Traveling Waves

Waves are nondispersive in some important physical cases: e.g., light propagation in a vacuum and audible sound in air. Since $\frac{\omega}{k} = v_\phi$, we can easily express the temporal angular frequency $\omega$ in terms of the angular wavenumber $k$:

$$\omega = \omega(k) = (v_\phi) \cdot k \text{ where } v_\phi \text{ is constant, so that } \omega \propto k.$$  

The expression of $\omega$ in terms of $k$ is called a dispersion relation. We can plot $\omega[k]$ vs. $k$, giving a straight line in the nondispersive case.

*Dispersion relation $\omega[k] \propto k$ for nondispersive waves.*
4.10 Dispersive Traveling Waves

The more general, more common, and more important case is that of dispersive waves, where the phase velocity \( v_\phi = \frac{\omega}{k} \) varies with angular frequency \( \omega \). This is the normal state of affairs for light traveling in a medium such as glass. The common specification of the phase velocity of light in medium is the refractive index \( n \):

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

where \( v_\phi \) is the phase velocity of light in the medium. In a dispersive medium, we can interpret group velocity in another way:

\[
\omega(k) = k \cdot v_\phi \implies v_{\text{mod}} = \frac{d\omega}{dk} = \frac{d}{dk} (k \cdot v_\phi) = \left( \frac{dk}{dk} \right) v_\phi + k \cdot \left( \frac{dv_\phi}{dk} \right) = v_\phi + k \left( \frac{dv_\phi}{dk} \right).
\]

In other words, the group velocity is the sum of the phase velocity \( v_\phi \) and a term proportional to \( \frac{dv_\phi}{dk} \), which is the change in phase velocity with wavenumber. There are two cases:

\[
\frac{dv_\phi}{dk} > 0 \implies v_{\text{mod}} > v_\phi
\]
\[
\frac{dv_\phi}{dk} < 0 \implies v_{\text{mod}} < v_\phi.
\]

As the phase velocity varies, the refractive index varies inversely (faster velocity \( \implies \) smaller index). Variation of the refractive index implies a change in the refractive angle of light entering or exiting the medium (via Snell’s law). Variation of refractive index with wavelength implies that different frequencies will refract at different angles. This is the principle of the dispersing prism.

4.10.1 Example:

Consider a medium with dispersion relation of the form of a power law:

\[
\omega(k) = \alpha k^\ell
\]

where \( \ell \) is a real number. The average and modulation velocities are:

\[
v_{\text{avg}} = \frac{\omega}{k} = \frac{\alpha(k^\ell)}{k} = \alpha k^{\ell-1}
\]
\[
v_{\text{mod}} = \frac{d\omega}{dk} = \frac{d}{dk} (\alpha k^\ell) = \ell (\alpha k^{\ell-1}) = \ell \cdot v_{\text{avg}}.
\]
So if \( \ell > 1 \), then \( v_{\text{mod}} > v_{\text{avg}} \), and if \( \ell < 1 \), \( v_{\text{mod}} < v_{\text{avg}} \). The first relation corresponds to anomalous dispersion and the second to normal dispersion. The dispersion relation for normal dispersion is nonlinear and concave down, while that for anomalous dispersion is nonlinear and concave up. Of course, for nondispersive waves the dispersion relation is linear.

*Theoretical dispersion curves for (a) normal dispersion \( (\omega \propto \sqrt{k}) \) and (b) anomalous dispersion (\( \omega \propto k^2 \)).*

In a medium with normal dispersion, the refractive index \( n \) increases with frequency \( \nu \) (or \( \omega \)) and decreases with wavelength \( \lambda \). Therefore \( n \) increases as the wavenumber \( k \) increases, i.e., \( \frac{dn}{dk} > 0 \). Thus in real media, the average waves travel faster than the modulation wave. This also means that the signal impressed on an electromagnetic wave cannot travel faster than the speed of light.

*Refractive index \( n \) vs. wavelength \( \lambda \) for several media, demonstrating the decrease in index (and thus increase in phase velocity) of light with increasing wavelength.*

### 4.10.2 Propagation of Superposition of Waves in Nondispersive and Dispersive Media

Recall that an anharmonic, though periodic, oscillation can be expressed as a sum of harmonic terms of different frequencies, i.e., as a Fourier series. We can therefore find...
4.10 DISPERSE TRAVELING WAVES

the effect of dispersion on an anharmonic traveling wave by decomposing it into its Fourier series of harmonic terms and propagating each separately at its own velocity. The resultant is found by resumming the resulting components. For example, if:

\[ f[z,t] = s_1[z,t] + s_2[z,t] + s_3[z,t] \]

\[ = A_1 \sin [k_1 z - \omega_1 t] + \frac{A_1}{3} \sin [k_2 z - 3\omega_1 t] + \frac{A_1}{5} \sin [k_3 z - 5\omega_1 t] \]

As we’ve already seen, \( f[z,t] \) is the sum of the first three terms of a square wave. The wave at the source is a “blurry square wave,” as shown in (a), where the three wavelengths of the three waves are respectively 4 units, \( \frac{4}{3} \) units, and \( \frac{4}{5} \) units. In the nondispersive case, \( k_1 = 3k_2 = 5k_3 \) and \( \lambda_1 = \frac{\lambda_3}{3} = \frac{\lambda_3}{5} \), which means in turn that \( \omega_1 = \omega_2 = \omega_3 \) and \( v_1 = v_2 = v_3 \). Since all components in the waveform propagate at the same velocity, then the relative phase difference is maintained throughout and the “shape” of the wavefront doesn’t change as it propagates as shown in (b):

\[ Propagation \text{ of a waveform in a nondispersive medium: (a) sum of three sinusoids to produce a “blurry” square wave (as though lowpass filtered); (b) resulting waveform after propagating all three sinusoids towards } z = +\infty \text{ with the same phase velocity, so that all components travel the same distance. The waveform at all times is undistorted. } \]

In dispersive media, energy conservation requires that the temporal frequencies are unchanged (\( \omega_1 = \omega_2 = \omega_3 \)). However, the phase velocities are no longer equal (\( v_1 \neq v_2 \neq v_3 \neq v_1 \)) and thus the wavelengths are no longer proportional. Put another way, the wavelengths vary with velocity and it must be the oscillation frequency and not the wavelength that determines the “color” of the light. As the waves travel through the media, their relative phases will vary, and the “shape” of the waveform will become increasingly distorted. If the dispersion is normal in the medium at these frequencies, then the lower-frequency sinusoid (e.g., \( s_1[z,t] \) in this example) travels faster than a high-frequency sinusoid, so that \( v_1 > v_2 > v_3 \) in this example. Consider
the resulting waveforms if the low-frequency component has moved 1 unit and 2 units from the case shown in (a) above:

Propagation of a waveform in a medium with normal dispersion. The same waveform above is assumed: (a) after the low-frequency term has propagated by one unit (the higher-frequency terms have moved shorter distances); (b) after low-frequency term has propagated two units, showing that the distortion in the waveform has increased.

Of course, the behaviour of the individual components in anomalous dispersion is complementary; the high-frequency sinusoidal terms move faster \( v_1 < v_2 < v_3 \); the distortion still exists, but it is in some sense “reversed.”

Propagation of a waveform in a medium with anomalous dispersion assuming same “blurry” square wave used previously: (a) after the low-frequency term has propagated by one unit and the higher-frequency terms longer distances; (b) after low-frequency term has propagated two units.
4.10.3 Energy and Information Transmission in Nondispersive and Dispersive Media

The issue of differential wave velocities also is relevant to the propagation of energy, information, and "messages." This concept is interesting in its own right, and also potentially confusing, so we’ll discuss it (albeit briefly). A good source on the subject is Chapter 6 of *Waves*, by Crawford. In amplitude modulation (e.g., AM radio), the information function (typically speech or music, call it \(s[t]\)) multiplies ("modulates") a high-frequency carrier wave (call it \(r[t]\)):

\[
f[t] = s[t] \cdot r[t] = s[t] \cdot \cos[\omega_{\text{carrier}} \cdot t]
\]

The FCC decrees that the frequency of the carrier wave (\(\nu_{\text{carrier}} = \frac{\omega_{\text{carrier}}}{2\pi}\)) lies in the range \(500 \text{ kHz} \leq \nu_0 \leq 1600 \text{ kHz}\), while the audio frequencies in \(s[t]\) are much lower (\(20 \text{ Hz} \lesssim \nu_{\text{audio}} \lesssim 20 \text{ kHz}\)). The modulated signal radiates as a traveling wave either through the nondispersive vacuum of space or a normally dispersive medium of air (though the dispersion of air is small). Because the carrier frequency is so much larger than the signal frequency, the velocities of the average and modulation waves are:

\[
\begin{align*}
\nu_{\text{avg}} &= \nu_{\text{avg}} \cdot \lambda_{\text{avg}} = \frac{\omega_{\text{carrier}}}{\nu_{\text{carrier}}} \\
\nu_{\text{mod}} &= \nu_{\text{mod}} \cdot \lambda_{\text{mod}} \approx \frac{d\omega}{dk} \bigg|_{\nu_{\text{carrier}}}
\end{align*}
\]

The information (speech or music) is carried by the modulation and travels through the medium at the modulation velocity, which we know to be less than the average velocity in a normally dispersive medium. An example is shown as "snapshots" of the product of a long-period sinusoidal modulation with frequency \(\nu_1\) and period \(\lambda_1\) (taken to be \(\frac{8}{9}\) units in this example) and a short-period carrier wave with frequency \(\nu_2\) and \(\lambda_2\) (= \(\frac{8}{9}\) units). The phase velocity of the higher-frequency is assumed to be \(\frac{12}{13} \approx 93\%\) of the velocity of the lower-frequency wave). The periods of the average and modulation waves are:

\[
\begin{align*}
\lambda_{\text{avg}} &= 2 \frac{\lambda_1 \cdot \lambda_2}{\lambda_1 + \lambda_2} = 1 \text{ unit} \\
\lambda_{\text{mod}} &= 2 \frac{\lambda_1 \cdot \lambda_2}{|\lambda_1 - \lambda_2|} = 8 \text{ units}
\end{align*}
\]

The snapshots are taken at increments of \(T = \frac{T_{\text{avg}}}{4}\), so that the average wave propagates by one-quarter period between images. A point of constant phase on the average wave is denoted in each image by the black dot, which is seen to travel faster than a point of constant phase on the modulation wave; in this case, the ratio of modulation
Figure 4.1: Illustration of normal dispersion of two waves shown at increments of $\Delta t = \frac{T_{\text{avg}}}{4}$. The black dot marks a point on the average wave with the same phase, which moves faster than the corresponding point on the modulation wave.

velocity to average velocity is approximately:

$$\frac{v_{\text{mod}}}{v_{\text{avg}}} \approx 0.69$$

Thus the “information” travels about 70% as fast as the average wave.
Chapter 5

Doppler Effect

5.1 Transition from Acoustic Waves to Electromagnetic Waves

The change in the frequency of a sound wave due to relative motion of the source and/or receiver is very familiar – the increase in pitch of an approaching or receding locomotive airhorn is a common example. This effect was described mathematically by Christian Doppler in 1842, and is naively understood by many people. However, few realize the fundamental difference between the Doppler effect due to source motion and that due to receiver motion.

5.1.1 Acoustic Doppler Effect: Stationary Source, Moving Observer

Consider a point source of sound in air which emits a frequency $\nu$. The receiver moves relative to the source at velocity $v_o$. Since the source and medium are at rest, the sound has a wavelength $\lambda = \frac{v}{\nu}$, where $v$ is the velocity of sound in air ($\approx 330$ m/s at STP). Since the source is at rest, the wavefronts expand uniformly from the source. A receiver traveling toward (away from) the source passes more (fewer) peaks of the sound wave in a given time interval than (s)he would were (s)he stationary. Therefore, the receiver hears a higher (lower) pitch.

Situation for Doppler effect with stationary source and moving observer.
This is shown in a snapshot of the source, receiver, and the emerging wavefronts (i.e., a wavefront is the locus of points of constant phase on a wave). The number of wave peaks heard per unit time is the observed frequency $\nu'$, and equals the source frequency plus (minus) the number of extra cycles heard due to observer motion:

$$\nu' = \nu \pm \Delta \nu = \nu \pm \frac{v_o}{\lambda} = \frac{v}{\lambda} \pm \frac{v_o}{\lambda} = \frac{v}{\lambda} \cdot \left(1 \pm \frac{v_o}{v}\right)$$

The $+$ sign means that the receiver approaches the source.

Example: $\nu = 1000$ Hz, $v_o = 60$ mph = 88 fps = 26.8 m/s toward source

$$\nu' = 1000 \text{ Hz} \cdot \left[1 + \frac{26.8 \text{ m/s}}{330 \text{ m/s}}\right] \approx 1000 \text{ Hz} \cdot 1.081 = 1081 \text{ Hz} > 1000 \text{ Hz}$$

5.1.2 Acoustic Doppler Effect: Stationary Observer, Moving Source

Situation for Doppler effect with moving source and stationary observer. Note that the distances between adjacent maxima depend on the direction of travel of the source relative to the observer.

This is a snapshot of the wavefronts emitted by a source moving toward the receiver with velocity $v_s$. The wavefronts emitted at later times have less distance to travel to the observer, so the distance between adjacent wavefronts in the medium is actually shorter on the side where the source approaches the observer and and longer on the other side, so that:

$$\lambda' = \lambda \mp \Delta \lambda = \lambda \mp \frac{v_s}{\nu}$$

where negative sign $\implies$ source approaching observer.

Therefore the temporal frequency heard by the ear is:

$$\nu' = \frac{v}{\lambda'} = \frac{v}{\lambda \mp \frac{v_s}{\nu}} = \frac{v}{\lambda} \cdot \frac{\lambda \mp \frac{v_s}{v}}{\nu} = \nu \left[\frac{\frac{v}{v \pm v_s}}{\nu}\right] = \nu'$$

where $v$ is the velocity of the wave in the medium. In words, the observed frequency is larger if $v_s > 0$ $\implies$ the source is approaching the observer and the frequency is
smaller if the source is receding from the observer. This is the situation that is familiar
to most of use who have heard the rising pitch of fire sirens (and formerly of train
whistles). Consider the example with the same fixed source frequency $\upsilon = 1000$ Hz
and the source velocity is $v_s = 60$ mph $\approx \frac{26.8 \text{m}}{\text{s}}$ towards the observer

$$\upsilon' = 1000 \text{ Hz} \cdot \left[ \frac{330 \text{m}}{\text{s}} - \frac{26.8 \text{m}}{\text{s}} \right] \cong 1000 \text{ Hz} \cdot \left[ \frac{330}{303.2} \right] = 1088 \text{ Hz} > 1000 \text{ Hz}$$

Note that the two cases give different answers: if the source moves in the medium,
the observed frequency is larger than if the observer moves at the same velocity
relative to the source. ($1088 \text{ Hz}$ vs. $1081 \text{ Hz}$). This difference allows determination of
whether the observer or the source is moving relative to the medium.

5.1.3 Acoustic Doppler Effect: Both Source and Receiver
Moving

If both source and receiver move relative to the medium, the frequency is a combina-
tion of the two results:

$$\upsilon' = \upsilon \cdot \left[ \frac{v \pm v_o}{v \mp v_s} \right]$$

where the upper signs are used if the source and receiver approach each other

5.2 Doppler Effect for Light

5.2.1 Difference between Light and Sound

Because the Doppler effect for sound differs if the source moves rather than the
observer, it is possible to determine which is moving relative to the medium. If the
observer moves, the wavelength $\lambda$ in the medium is invariant and the change in pitch
is due to the more-or-less frequent passages of the wavefronts by the observer. If the
source moves, the wavelength of the sound in the medium changes and the sign of
the change depends on the direction of source motion. If this new wavelength is $\lambda'$,
then the new frequency is $\upsilon' = \frac{v_o}{\lambda'}$.

For light waves (electromagnetic radiation), the mechanism of wave propagation
(and hence of the Doppler effect) is fundamentally different from propagation of sound
in air. Because of this big difference, light propagation was not successfully described
until 1864, when James Clerk Maxwell collected and interpreted the four equations
which bear his name. The true nature of light was not generally accepted until post-
1880. Why is light so different?

Recall that two forces are required to sustain oscillations or propagate waves –
(1) inertia; (2) restoring force. Waves in common everyday experience (e.g., sound
in air, surface waves in water), inertia is supplied by the source (air motion from the
diaphragm, physical displacement of the water surface). The restoring force is due to
a characteristic of the medium of transmission (e.g., air pressure, gravity plus surface tension).

By the early 1800's, some characteristics of light were already known, e.g., the phase velocity \( c \) was known to be finite. The first recorded experiment to measure \( c \) was performed by Galileo around 1600. He stationed a man with a shuttered lantern on a distant hill with instructions to open the shutter as soon as he saw the light from Galileo’s lamp. By timing the interval between unshuttering his lamp and seeing the return beam, Galileo tried to measure \( c \) via \( c = \frac{2L}{t} \), where \( L \) is the distance between lanterns. His conclusion:

“If not instantaneous, light is extraordinarily rapid.”

A surprisingly good measurement of \( c \) was made by Ole Römer in 1675. The Keplerian laws of planetary motion enabled Römer to predict the times of eclipse of Jupiter’s Galilean satellites. He found that the measured times did not agree with prediction – when Jupiter was closest to earth, the times of eclipse were early, and when Jupiter was distant the times were late. Römer ascribed the difference to a finite velocity of light, and computed a value of \( c = 2 \times 10^8 \text{ m/s} \). The largest source of error was Römer’s estimate of the earth’s orbital velocity. When corrected for this error, Römer’s method yields a very accurate value of \( 3 \times 10^8 \text{ m/s} \). Besides its velocity, light had been demonstrated to have the character of a wave by Newton’s demonstration of dispersion by a prism and by the polarization experiments of Fresnel. These characteristics led to Fresnel’s hypothesis of the “aether” – the medium of transmission for light, which is analogous to air for propagation of sound. If it exists, the aether must be present everywhere, including in vacuum. From the calculations of the Doppler effect, the frequency shift of light must depend on whether the source or the observer is moving.

The need for the aether was eliminated by Maxwell (as we shall soon see), and its existence was disproved by Michelson and Morley in 1880 when they demonstrated that the velocity of light is identical parallel to or perpendicular to the orbital motion of the earth, which would not have been true had an aether been necessary for propagation.

Einstein used Michelson’s results to derive the Special Theory of Relativity, which states:

“The velocity of light is constant, regardless of the motion of the source or the observer.

In addition, there is no preferred frame of reference.”

Therefore when considering light, the Doppler effect should yield identical results if the source is moving or if the observer is moving. In fact, it is impossible to define which moves; only the relative motion is meaningful.

Einstein’s result is:

\[

\nu' = \nu \left( \frac{1 - \frac{\nu}{c}}{\sqrt{1 - (\frac{\nu}{c})^2}} \right) = \left[ \nu \left( 1 - \frac{\nu}{c} \right) \left( 1 - \left( \frac{\nu}{c} \right)^2 \right)^{-\frac{1}{2}} \right]

\]
where the square root may be approximated via applying the well-known power series:

\[(1 + u)^n = \frac{1}{0!} + \frac{n}{1!}u + \frac{n(n - 1)}{2!}u^2 + \frac{n(n - 1)(n - 2)}{3!}u^3 + \cdots + \left(\frac{n!}{(n - r)! \cdot r!}\right)u^r + \cdots\]

In this case, the series solution is:

\[
\nu' = \left[\nu \left(1 - \frac{v}{c}\right)\right] \left(1 - \frac{1}{2} \left(\frac{v}{c}\right)^2 \right)^1 + \left(-\frac{1}{2}\right) \left(-\frac{3}{2}\right) \frac{1}{2!} \left(\frac{v}{c}\right)^2 + \cdots \\
= \left[\nu \left(1 - \frac{v}{c}\right)\right] \left(1 - \frac{1}{2} \left(\frac{v}{c}\right)^2 \right) + \frac{3}{8} \left(\frac{v}{c}\right)^4 + \cdots \\
= \nu \left[1 - \left(\frac{v}{c}\right)\right] + \left[\frac{1}{2} \left(\frac{v}{c}\right)^2 - \frac{1}{2} \left(\frac{v}{c}\right)^3\right] + \left[\frac{3}{8} \left(\frac{v}{c}\right)^4 - \frac{3}{8} \left(\frac{v}{c}\right)^5\right] + \cdots 
\]

In the case of light its velocity \(c >> v\), we can dispense with the terms with orders larger than unity:

\[
\nu' \approx \nu \left(1 - \frac{v}{c}\right) = \nu - \nu \frac{v}{c} \text{ for } v \ll c.
\]

Thus if the relative velocity of source and observer is positive (so that the distance increases), then the Doppler shift decreases the frequency (and increases the wavelength) by an amount that is proportional to \(v\). This is the famous “red shift” in astronomy. Obviously the complementary solution also applies if the source and observer approach each other to produce a “blue shift.”
Chapter 6
Maxwell’s Equations

6.1 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:

\[
\mathbf{x} \equiv \begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{bmatrix}
\]

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 \( \mathbf{x} \) and \( \mathbf{y} \) 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:

\[
\mathbf{x}^T = [x_1 \ x_2 \ x_3] = \tilde{\mathbf{x}}
\]

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:

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

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:

\[
\hat{\mathbf{x}} = \frac{\mathbf{x}}{|\mathbf{x}|} = \begin{bmatrix}
\frac{x_1}{|\mathbf{x}|} \\
\frac{x_2}{|\mathbf{x}|} \\
\vdots \\
\frac{x_N}{|\mathbf{x}|}
\end{bmatrix}
\]

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” (\(\cdot\)) between the symbols for the vectors. The process also may be written as the transpose of \(\hat{\mathbf{x}}\) multiplied from the right by \(\mathbf{x}\). Therefore, the scalar product of a vector \(\hat{\mathbf{x}}\) with itself may be written in equivalent ways.

\[
|\mathbf{x}|^2 = (\mathbf{x} \cdot \mathbf{x}) \equiv \hat{\mathbf{x}}^T \mathbf{x} = \sum_{n=1}^{N} x_n^2
\]
6.1.1 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 $N$:

$$\mathbf{a} \cdot \mathbf{x} \equiv \mathbf{a}^T \mathbf{x} = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = a_1x_1 + a_2x_2 + \cdots + a_Nx_N = \sum_{n=1}^{N} a_n x_n$$

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 \(\mathbf{f}_n\). 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:

\[
\mathbf{f}_1 \cdot \mathbf{f}_2 = x_1x_2 + y_1y_2 \\
= (|\mathbf{f}_1| \cos [\theta_1]) (|\mathbf{f}_2| \cos [\theta_2]) + (|\mathbf{f}_1| \sin [\theta_1]) (|\mathbf{f}_2| \cos [\theta_2]) \\
= |\mathbf{f}_1| |\mathbf{f}_2| (\cos [\theta_1] \cos [\theta_2] + \sin [\theta_1] \sin [\theta_2]) \\
= |\mathbf{f}_1| |\mathbf{f}_2| \cos [\theta_1 - \theta_2] = |\mathbf{f}_1| |\mathbf{f}_2| \cos [\theta_1 - \theta_2]
\]

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 \textit{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 \textit{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.

Graphical comparison of Schwarz’ and the triangle inequalities for the same pair of 2-D vectors \( \mathbf{x} \) and \( \mathbf{a} \).

The Schwarz inequality may be combined with the definition of the unit vector to obtain an expression for the included angle between two unit vectors:

\[
\frac{\mathbf{a}}{|\mathbf{a}|} \cdot \frac{\mathbf{x}}{|\mathbf{x}|} = \hat{\mathbf{a}} \cdot \hat{\mathbf{x}} = \cos \theta \leq 1
\]

### 6.1.2 Cross Product

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

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

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:

\[
\mathbf{a} \times \mathbf{b} = \det \begin{bmatrix}
\hat{x} & \hat{y} & \hat{z} \\
a_x & a_y & a_z \\
b_x & b_y & b_z
\end{bmatrix}
= \hat{x}(a_y b_z - a_z b_y) + \hat{y}(a_z b_x - a_x b_z) + \hat{z}(a_x b_y - a_y b_x)
\]

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!

### 6.1.3 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:

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

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.

### 6.2 Vector Calculus

In 1864, James Clerk Maxwell published a paper on the dynamics of electromagnetic fields, in which he collected four previously described equations which 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 decrease 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.

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 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
\( \mathbf{s} \)).

\[
\mathbf{F} \cdot \mathbf{s} = W
\]

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:

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

Note that no work is performed if the force acts at right angles to the displacement;
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.

The work can be evaluated via:
\[ W = \int \mathbf{F} \cdot d\mathbf{s} = \int (\hat{x} F_x + \hat{y} F_y + \hat{z} F_z) \cdot (\hat{x} dx + \hat{y} dy + \hat{z} dz) \]
\[ = \int F_x \, dx + \int F_y \, dy + \int F_z \, dz = T + c \]

where \( T \) is the kinetic energy and \( c \) is a constant.

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:

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

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

\[ \int \mathbf{F} \cdot d\mathbf{s} = \int (\hat{x} F_x + \hat{y} F_y + \hat{z} F_z) \cdot (\hat{x} dx + \hat{y} dy + \hat{z} dz) \]
\[ = \int \left(-\frac{\partial V}{\partial x}\right) \, dx + \int \left(-\frac{\partial V}{\partial y}\right) \, dy + \int \left(-\frac{\partial V}{\partial z}\right) \, dz \]
\[ = -\int dV = -V = T + c \]
\[ \implies T + V \equiv E = \text{constant} \]

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:

\[ \mathbf{F} = \hat{x} F_x + \hat{y} F_y + \hat{z} F_z \]
\[ = 0 \hat{x} + 0 \hat{y} + \hat{z} (-mg) \]
so that:
\[- \frac{\partial V}{\partial x} = 0 \implies V = c_1\]
\[- \frac{\partial V}{\partial y} = 0 \implies V = c_2\]
\[- \frac{\partial V}{\partial z} = -mg \implies V = mg \int dz = mgz + c_3\]
\[\implies V [x, y, z] = mgz + (c_1 + c_2 + c_3) = mgz + \text{constant}\]
\[E = mgz + \frac{1}{2}mv^2\]

Under the conditions of a conservative force, we can write differentiate the first two expressions with respect to the “other” variable and equate them:
\[\frac{\partial}{\partial y} F_x = \frac{\partial}{\partial y} \left( -\frac{\partial V}{\partial x} \right) = -\frac{\partial^2 V}{\partial x \partial y}\]
\[\frac{\partial}{\partial x} F_y = \frac{\partial}{\partial x} \left( -\frac{\partial V}{\partial y} \right) = -\frac{\partial^2 V}{\partial y \partial x} = \frac{\partial}{\partial y} F_x\]

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

The same pattern of operations leads to two other relations:
\[\frac{\partial}{\partial z} F_x = \frac{\partial}{\partial x} F_z\]
\[\frac{\partial}{\partial z} F_y = \frac{\partial}{\partial y} F_z\]

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

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

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

where \( \mathbf{\hat{x}}, \mathbf{\hat{y}}, \) and \( \mathbf{\hat{z}} \) are the unit vectors along the \( x, y, \) and \( z \) axes respectively. Thus
we can write:
\[ F = -\nabla V \]

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 \( \nabla \cdot \mathbf{g} [x, y, z] \); 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.

6.3 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.

\[
\nabla f [x, y, z] \equiv \left[ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right] = \mathbf{\hat{x}} \frac{\partial f}{\partial x} + \mathbf{\hat{y}} \frac{\partial f}{\partial y} + \mathbf{\hat{z}} \frac{\partial f}{\partial z} = \text{a vector}
\]

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.\( x:\)
6.4 Divergence of a Vector field

Derives a Scalar Field $\nabla \cdot \mathbf{g}$ from a Vector Field $\mathbf{g}$:

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

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 ($flux = \text{net outward flow}$), and thus is equal to:

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

For a vector field $\mathbf{g} [x, y, z]$ and an infinitesimal surface “element” described by its normal differential vector element $d\mathbf{a}$ directed outward from the volume, the differential element of the “flux” $F$ (a scalar) of $\mathbf{g}$ through the surface element $d\mathbf{a}$ is the scalar (“dot”) product of the vector that describes the field with the vector normal to the
6.5 CURL OF A VECTOR FIELD

surface. Thus the total flux is the integral over the surface:

\[ d\vec{F} = \mathbf{g} \cdot d\mathbf{a} \]

\[ \Rightarrow \quad \int_{\text{surface area } S} \mathbf{g} \cdot d\mathbf{a} \]

The divergence of the vector field describes the total flux through the macroscopic surface area \( A \) built up from all of the differential surface elements \( d\mathbf{a} \) 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:

\[ \nabla \cdot \mathbf{g} = \int_{\text{surface area } S} \mathbf{g} \cdot d\mathbf{a} \]

\[ = 0 \text{ if no “source” or “sink” of vector field within volume enclosed by } S \]

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

The divergence of the scalar field \( \mathbf{f}[x, y] \) is \( \nabla \cdot \mathbf{f}[x, y] \) 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)

6.5 Curl of a Vector Field

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

The curl of a vector field describes a 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:

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

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:

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

To visualize curl, imagine a vector field that describes motion of a fluid (e.g., water or wind). If a paddle-wheel placed in the fluid does not revolve, 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 will rotate only if the vector field is spatially nonuniform. Note that 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.

### 6.5.1 Example of Function with Large Curl

Consider the 3-D field composed of vectors defined by the equation:

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

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.
6.5 CURL OF A VECTOR FIELD

The vector field \( \mathbf{g}(x, y, z) = [-y, x, 0] \), for which the curl is the vector \( \nabla \times \mathbf{g} = [0, 0, +2] \). The curl vector points out of the plane of the paper towards the reader.

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 opposite of the well-known “bathtub drain vortex,” where 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:

\[
\frac{\partial g_x}{\partial x} = 0, \quad \frac{\partial g_x}{\partial y} = -1, \quad \frac{\partial g_x}{\partial z} = 0
\]

\[
\frac{\partial g_y}{\partial x} = +1, \quad \frac{\partial g_y}{\partial y} = 0, \quad \frac{\partial g_y}{\partial z} = 0
\]

\[
\frac{\partial g_z}{\partial x} = \frac{\partial g_z}{\partial y} = \frac{\partial g_z}{\partial z} = 0
\]

The curl of the field is obtained by direct substitution:

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

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

\[
= 0 \cdot \hat{x} + 0 \cdot \hat{y} + 2 \cdot \hat{z}
\]

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.

### 6.6 Laplacian of a Scalar Field

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” of the function \(f[x, y, z]\)

\[
\nabla \cdot \nabla f = \nabla^2 f
\]

\[
= \left( \frac{\hat{x}}{\partial x} \frac{\partial f}{\partial x} + \frac{\hat{y}}{\partial y} \frac{\partial f}{\partial y} + \frac{\hat{z}}{\partial z} \frac{\partial f}{\partial z} \right) \cdot \left( \frac{\hat{x}}{\partial x} \frac{\partial f}{\partial x} + \frac{\hat{y}}{\partial y} \frac{\partial f}{\partial y} + \frac{\hat{z}}{\partial z} \frac{\partial f}{\partial z} \right)
\]

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

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

\[
= \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \equiv \nabla^2 f
\]

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

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

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

The Laplacian is the spatial derivative in the 3-D wave equation, which will be considered in more detail shortly:

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

\[
\implies \nabla^2 f = \frac{1}{\epsilon \mu} \frac{\partial^2 f}{\partial t^2}
\]
6.6.1 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:

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

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

6.7 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 (such as from a moving magnet) can generate an electric field, and Ampere 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 vectors 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:

\[ \frac{\mathbf{F}}{Q} \propto \mathbf{E} \]

where the force is measured in newtons \( \left[ \frac{kg \cdot m}{s^2} \right] \) as the product of the charge \( Q \) and the electric field \( \mathbf{E} \); it has dimensions of volts per meter (equivalent to joules per coulomb).

6.7.1 A Note on Units

If you consult other books, you will likely see many differences in the equations due to the different systems of units used in electromagnetics (and thus in optics); many students (including the author!) find it difficult 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), the latter uses CGS units (centimeter, gram, second), which 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 \( k \):

\[
\mathbf{F} = k \frac{Q_1 Q_2}{r_{12}^2} \hat{r}_{12}
\]

If the charges are measured in electrostatic units (esu) (also called statcoulombs), the distance in centimeters, and the force in dynes \( (g - \text{cm}^2 - s^{-2}) \), 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 [C], the distance in meters [m], and the force in newtons \( (1 \text{ N} = 1 \text{ kg} - \text{m} - \text{s}^{-2}) \)? The value of \( k \) is determined from the knowledge that there are \( 10^5 \text{ dyn per N} \), \( 2.998 \times 10^9 \text{ esu per C} \), and \( 10^{+2} \text{ cm per m} \):

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

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

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

\[
k \equiv \frac{1}{4\pi \epsilon_0} \implies \mathbf{F} = \frac{1}{4\pi \epsilon_0} \frac{Q_1 Q_2}{r_{12}^2} \hat{r}_{12}
\]

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

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

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

(one coulomb per volt), so that 1 volt is equivalent to:

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

This “new” normalization constant 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 \( \mu_0 \):

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

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

$$\mathbf{B} = \mu_0 \mathbf{H}$$

### 6.7.2 Magnetic Fields

The concept of a magnetic field is seemingly somewhat less intuitive, 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” ($1 \text{ Wb} = 10^8 \text{ Mx}$), which was defined as the amount of 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*, labeled $\mathbf{B}$ and measured in *gauss* (CGS) or *tesla* (MKS). One gauss is one line (maxwell) through an area of 1 cm$^2$ and one tesla is 1Wb per m$^2$:

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

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

Two other vector fields are required when describing propagation of electromagnetism through matter (rather than through vacuum): the electric displacement $\mathbf{D}$ and the magnetic field intensity $\mathbf{H}$ (also called the “magnetizing force” or the “auxiliary field”). 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 due to an external field $\mathbf{E}$. It is the sum of $\mathbf{E}$ and any local field $\mathbf{P}$ generated within the matter due to the changes in positions of electric charges within the material due to that field; this induced field $\mathbf{P}$ is called the “polarization” of the material (not to be confused with the “polarization” of the electric field vector that we will mention later). $\mathbf{H}$ is a similar construct for magnetic fields. $\mathbf{E}$ 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:

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

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

where $\varepsilon$ and $\mu$ are the *electric permittivity* and *magnetic permeability* of the material, respectively. These are measures of the ability of the electric and magnetic fields to
“permeate” the medium; if \( \epsilon \) is increased, then a larger electric field exists within the material, if \( \mu \) Since we will consider propagation of light only in vacuum, \( \mathbf{D} = \mathbf{E} \) and \( \mathbf{B} = \mathbf{H} \). In vacuum, \( \mu \) and \( \epsilon \) are denoted \( \mu_0 \) and \( \epsilon_0 \) and both are set to unity in CGS units. In MKS units, the quantities are:

\[
\begin{align*}
\mu_0 &= 4\pi \cdot 10^7 \frac{N}{A^2} \quad (\text{newton per square ampere}) \\
\epsilon_0 &= 8.85 \cdot 10^{-12} \frac{F}{m} \quad (\text{farads per meter}).
\end{align*}
\]

As is true for the refractive index \( n \), the permittivity and permeability in matter are larger than in vacuum, \( \epsilon > \epsilon_0 \), and \( \mu > \mu_0 \). In fact (though we won’t discuss it in detail), \( \epsilon \) and \( \mu \) determine the phase velocity \( v \) and the refractive index \( n \) via:

\[
\begin{align*}
\frac{1}{\sqrt{\mu \epsilon}} & \quad v = \\
\frac{c}{\sqrt{\mu_0 \epsilon_0}} & \quad n = \frac{c}{v} = \frac{\sqrt{\mu_0 \epsilon}}{\sqrt{\mu_0 \epsilon_0}}
\end{align*}
\]

### 6.8 Maxwell’s Equations

Maxwell collected the four differential equations relating the electric vector field \( \mathbf{E} \) and the magnetic vector field \( \mathbf{B} \) listed below and solved them to derive the character of electromagnetic waves. The equations may be written in equivalent differential and integral forms.

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

Gauss’ law relates the flux of the electric field over a closed surface to the total charge enclosed by the surface. In its simplest terms, Gauss’ law states that the existence of electrical charges within a volume produces electric fields that pass through the surface of the volume. The flux 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. Also note that if there is no charge within the volume, there still can be flux through the enclosing surface, but the ingoing and outgoing parts of the flux cancel out.

Consider an element of the closed surface defined by its normal vector \( da \). The flux of the electric field through this surface element is:

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

where the symbol “\( \cdot \)” denotes the scalar product of the two vector quantities. According to Gauss’ law, the integral of this quantity over the entire closed surface
is:

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

where \( \rho[x, y, z] \) is the volume density of charges (measured coulombs per unit volume). If the surface encloses no charges, then this integral evaluates to zero. states that the divergence of the vector electric field is proportional to density of electric charges.

\[
\int \int_{\text{surface}} \mathbf{E} \cdot d\mathbf{a} = \int \int_{\text{volume}} \left( \frac{\partial}{\partial x} \mathbf{E}[x, y, z, t] + \frac{\partial}{\partial y} \mathbf{E}[x, y, z, t] + \frac{\partial}{\partial z} \mathbf{E}[x, y, z, t] \right) \, dV \\
= \int \int_{\text{volume}} \left( \nabla \cdot \mathbf{E}[x, y, z, t] \right) \, dV
\]

\[ \int \int_{\text{volume}} \left( \nabla \cdot \mathbf{E}[x, y, z, t] \right) \, dV = \frac{1}{\varepsilon} \int \int_{\text{volume}} \rho[x, y, z] \, dV \]

\[ \implies \nabla \cdot \mathbf{E}[x, y, z, t] = \frac{\rho[x, y, z]}{\varepsilon} \]

### 6.8.2 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 forms for Gauss’s law for the magnetic flux density:

\[ \int \int_{\text{surface}} \mathbf{B} \cdot d\mathbf{a} = 0 \]

\[ \nabla \cdot \mathbf{B}[x, y, z, t] = 0 \]

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.”

### 6.8.3 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:

$$\frac{\partial B}{\partial t} = -\nabla \times E$$

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

6.8.4 4. Ampere’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 surface. The corrected form of Ampere’s law is:

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

where the additional source term $J$ 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 Ampere’s law.

We have already seen that:

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

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 $J$. These can only be nonzero within media (such as copper wire) and thus vanish in vacuum. If we consider the propagation of light only in a vacuum, neither electric charges nor conductors are present and both source terms vanish.

6.8.5 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 are:
\[ \nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon} \quad \text{Gauss’ Law for Electric Fields, Coulomb’s Law} \]

\[ \nabla \cdot \mathbf{B} = 0 \quad \text{Gauss’ Law for Magnetic Fields} \]

\[ -\frac{\partial \mathbf{B}}{\partial t} = \nabla \times \mathbf{E} \quad \text{Faraday’s Law of Magnetic Induction} \]

\[ +\varepsilon \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \left( \frac{\mathbf{B}}{\mu} - \mathbf{J} \right) \quad \text{Ampere’s Law (with correction)} \]

The definition of curl may be used to rewrite the four vector equations of Maxwell as eight scalar equations:

\[ \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \frac{\rho}{\varepsilon} \quad \text{Gauss’ Law for } \mathbf{E} \quad (1) \]

\[ \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0 \quad \text{Gauss’ Law for } \mathbf{B} \quad (2) \]

\[ -\frac{\partial B_x}{\partial t} = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \quad \text{Faraday’s Law} \quad (3) \]

\[ -\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \quad (4) \]

\[ -\frac{\partial B_z}{\partial t} = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \quad (5) \]

\[ +\mu \varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} - J_x \quad \text{Ampere’s Law} \quad (6) \]

\[ +\mu \varepsilon \frac{\partial E_y}{\partial t} = \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} - J_y \quad (7) \]

\[ +\mu \varepsilon \frac{\partial E_z}{\partial t} = \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} - J_z \quad (8) \]

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 infinite distance away!), then the charge density \( \rho = 0 \) and the current density \( \mathbf{J} = 0 \) \( \implies \) \( J_x = J_y = J_z = 0 \).

### 6.9 Wave Equation for EM Waves

Take the curl of both sides of Faraday’s law. We can use the expression for the “curl of the curl” previously mentioned (though not derived) to evaluate the curl of the
curl of the electric field:

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

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

\[ = \mathbf{0} - \nabla^2 \mathbf{E} \]

where Gauss’ law for the electric field in a region with no sources has been used in the last step. The right side of the equation may be rewritten by applying Ampere’s law:

\[ \nabla \times \left( -\frac{\partial \mathbf{B}}{\partial t} \right) = -\frac{\partial}{\partial t} (\nabla \times \mathbf{B}) \]

\[ = -\varepsilon \mu \frac{\partial}{\partial t} \left( \frac{\partial \mathbf{E}}{\partial t} \right) \]

\[ = -\frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} \]

After equating the two sides of the equation:

\[ \nabla^2 \mathbf{E} = \frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} \]

which relates the spatial and temporal second derivatives of the electric field; this again is the wave equation, but for 3-D functions. The wave equation was first considered by d’Alembert in 1747. It assumes that no energy of the wave is lost, such as 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 \( v \) is the velocity of a point of constant phase: this is our old friend, the phase velocity. The wave equation may be rigorously derived for a transverse wave on a string – you probably saw this in the mechanics section of a physics course.

The wave equation for electric fields confirms our earlier observation:

\[ \varepsilon \mu = \frac{1}{c^2} \quad \Rightarrow \quad c = \sqrt{\frac{1}{\varepsilon \mu}} \]

Think of this result for a second; the phase velocity of the wave in a medium is related to two measurable properties of the medium; the permittivity and the permeability.

The 1-D equation may be written in the form of a “second-order homogeneous” differential equation:

\[ \left( \frac{\partial^2}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) \psi [z, t] = 0 \]

Any differential equation is linear, so that if \( \psi_1[z,t] \) and \( \psi_2[z,t] \) are solutions to the
6.9 WAVE EQUATION FOR EM WAVES

The wave equation has the simple solution:

$$\psi [z, t] = f [z \pm \nu \phi t]$$

where \( f [u] \) is any function that may be differentiated twice.

**Proof.**

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

Apply chain rule:

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial z}$$

Apply chain rule:

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial t}$$

Substitute into wave equation:

$$\frac{\partial^2 f}{\partial z^2} = \frac{1}{\nu^2} \frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial u^2} \cdot \nu^2 = \frac{\partial^2 f}{\partial u^2}$$

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

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

$$\frac{1}{\nu^2} \frac{\partial^2}{\partial t^2} (\hat{x} E_0 \cos [kz - \omega t]) = \frac{1}{\nu^2} \hat{x} E_0 (-\omega^2 \cos [kz - \omega t])$$

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

$$\implies \nu^2 = \frac{\omega^2}{k^2}$$

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

$$\psi [z, t] = f [z - \nu \phi t]$$

where the form of the function \( f \) is arbitrary, then the argument of the function \([z - \nu \phi t]\) (the “phase”) remains constant if \( x \) increases with increasing time. The “shape” \( f \) moves towards \( z = +\infty \) with increasing time without changing its shape (i.e., without “dispersion”). A second solution to this equation is:

$$\psi [z, t] = g [z + \nu \phi t]$$

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 their corresponding “wavenumber” \(k_x, k_y, k_z\) that may be written as a “wavevector” \(\mathbf{k}_0:\)

\[
\psi [x, y, z, t] = A \cos [\Phi [x, y, z, t]]
\]

\[= A \cos [k_x x + k_y y + k_z z \pm \omega_0 t - \phi_0] \]

\[= A \cos [\mathbf{k}_0 \cdot \mathbf{r} \pm \omega_0 t - \phi_0] \]

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

### 6.9.1 Electromagnetic Waves from Maxwell’s Equations

In the general case, the electric field and magnetic fields can have the form:

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

\[
\mathbf{B} [x, y, z, t] = \hat{x} B_x [x, y, z, t] + \hat{y} B_y [x, y, z, t] + \hat{z} B_z [x, y, z, t]
\]

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:

\[
\frac{\partial E_x}{\partial x} = \frac{\partial E_x}{\partial y} = \frac{\partial E_y}{\partial x} = \frac{\partial E_y}{\partial y} = \frac{\partial E_z}{\partial x} = \frac{\partial E_z}{\partial y} = 0
\]

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

\[
\mathbf{E} [x, y, z, t] = \mathbf{E} [z, t] = \hat{x} E_x [z, t] + \hat{y} E_y [z, t] + \hat{z} E_z [z, t]
\]
6.9 WAVE EQUATION FOR EM WAVES

From (9) and Gauss’ law for electric fields (1), we find that:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0 \implies \frac{\partial E_z}{\partial z} = 0 \quad (11)$$

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:

$$E_z [x, y, z] = \text{constant} \rightarrow 0 \quad (12)$$

Therefore, the electric field is now expressable in a much simpler form:

$$\mathbf{E} [x, y, z] = E_x [z, t] + E_y [z, t] \quad (13)$$

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

$$E_y [z, t] = 0 \text{ by assumption} \quad (14)$$

The expression for the electric field is:

$$\mathbf{E} [x, y, z] = \hat{x} E_x [z, t] \quad (15)$$

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:

$$-\frac{\partial B_x}{\partial t} = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = 0 - \frac{\partial E_y}{\partial z} \implies \frac{\partial B_x}{\partial t} = \frac{\partial E_y}{\partial z} = 0 \implies B_x [t] \text{ is constant} \quad (3)$$

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = \frac{\partial E_x}{\partial z} \implies \frac{\partial B_y}{\partial t} = -\frac{\partial E_x}{\partial z} \quad (4)$$

$$-\frac{\partial B_z}{\partial t} = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = 0 \implies B_z [t] \text{ is constant} \quad (5)$$

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

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} \quad (4)$$

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 \( z \):

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

\[
\Rightarrow -\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} = -k E_0 \sin [kz - \omega t]
\]

\[
\Rightarrow -B_y[z, t] = \int \frac{\partial E_x}{\partial z} dt = -(k E_0) \int \sin [kz - \omega t] dt
\]

\[
-B_y[z, t] = +k E_0 \left[ \frac{-\cos [kz - \omega t]}{-\omega} \right]
\]

\[
= \frac{k}{\omega} E_0 \cos [kz - \omega t]
\]

\[
= \frac{E_0}{(\frac{\omega}{k})} \cos [kz - \omega t]
\]

\[
= \frac{E_0}{v_\phi} \cos [kz - \omega t]
\]

\[
\Rightarrow \mathbf{B}[z, t] = \hat{y} \left( -\frac{E_0}{v_\phi} \cos [kz - \omega t] \right)
\]

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

\[
B_y = \frac{E_0}{v_\phi} \cos [kz - \omega t] = \frac{E_x}{v_\phi} \Rightarrow E_x = v_\phi B_y
\]

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.

### 6.9.2 Phase Velocity of Electromagnetic Waves

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

\[
\begin{align*}
+\mu \epsilon \frac{\partial E_x}{\partial t} &= \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} & (6) \\
+\mu \epsilon \frac{\partial E_y}{\partial t} &= \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} & (7) \\
+\mu \epsilon \frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} & (8)
\end{align*}
\]

Because \( \mathbf{E} = \hat{x} E_0 \), only (6) does not vanish:
\[
\mu \varepsilon \frac{\partial}{\partial t} (E_0 \cos [kz - \omega t]) = -\frac{\partial}{\partial z} \left( \frac{E_0}{v_\phi} \cos [kz - \omega t] \right)
\]

\[\implies \mu \varepsilon E_0 (\omega \sin [kz - \omega t]) = -\frac{E_0}{v_\phi} (-k \sin [kz - \omega t])\]

\[\implies \mu \varepsilon \omega E_0 = \frac{E_0 k}{v_\phi}\]

\[\implies \mu \varepsilon = \frac{k}{\omega v_\phi} = \frac{1}{v_\phi^2} \implies v_\phi^2 = \left(\frac{\omega}{k}\right)^2 = \frac{1}{\mu \varepsilon} \]

\[
v_\phi = \left(\sqrt{\mu \varepsilon}\right)^{-1}
\]

which we already knew from the wave equation. In vacuum, \(\mu \equiv \mu_0, \ v \equiv v_0, \ v_\phi \equiv c,\) and \(c = \sqrt{\frac{1}{\mu_0 \varepsilon_0}}.\) 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 in vacuum is:

\[
\mu \geq \mu_0 \equiv 4\pi \times 10^{-7} \text{ newton/ampere}^2 \approx 1.26 \times 10^{-6} \frac{\text{N}}{\text{A}^2}
\]

and the permittivity is:

\[
\varepsilon_0 \approx 8.85 \times 10^{-12} \frac{\text{F}}{\text{m}}
\]

These values produce the result:

\[
\mu_0 \varepsilon_0 = \left(8.85 \cdot 10^{-12} \frac{\text{C}}{\text{V} \cdot \text{m}}\right) \left(1.26 \cdot 10^{-6} \frac{\text{J}}{\text{A}^2 \cdot \text{m} \cdot \text{m}}\right)
\]

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

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

\[= 1.11 \cdot 10^{-17} \frac{\text{m}^2}{\text{J} \cdot \text{m}}
\]

\[\implies c = \sqrt{\frac{1}{\mu_0 \varepsilon_0}} = 2.99 \cdot 10^8 \frac{\text{m}}{\text{s}}, \ which \ agrees \ with \ experiment.
\]

In media (i.e., if not in vacuum), the phase velocity is different (we know it to be slower). 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, while the permittivity \(\varepsilon\) of optical materials is larger than in vacuum:

\[
\varepsilon > \varepsilon_0 \approx 8.85 \times 10^{-12} \frac{\text{F}}{\text{m}}
\]
So therefore
\[ v = \sqrt{\frac{1}{\mu \epsilon}} \cong \sqrt{\frac{1}{\epsilon}} \leq \sqrt{\frac{1}{\epsilon_0}} \cong \sqrt{\frac{1}{\mu_0 \epsilon_0}} = c \]

**Index of Refraction**

Of course, the dimensionless ratio of the velocity of light in vacuum to that in the medium is the *index of refraction* \( n \):

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

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

\[
\Rightarrow n^2 = \frac{\epsilon}{\epsilon_0}
\]

in dielectric materials. For metals and absorptive materials, and at wavelengths near absorptions, the index of refraction is complex valued and the permeabilities in the different media generally are different. The complex refractive index often is denoted by \( \tilde{n} \) and its imaginary part by \( \kappa \):

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

so that

\[
|k| = k = \frac{n \omega_0}{c} = (n + i\kappa) \frac{\omega_0}{c}
\]

which implies that the wavevector \( k \) is complex-valued as well. In this situation, the propagating electric field is written:

\[
\mathbf{E}[x, y, z, t] = E_0 \exp\left[+i \left( k \cdot \mathbf{r} - \omega_0 t \right) \right]
\]

\[
= E_0 \exp\left[+i \left( \tilde{k} \cdot \mathbf{r} \right) - \omega_0 t \right]
\]

\[
= E_0 \exp\left[+i \left( n + i\kappa \right) \left( \frac{\omega_0}{c} \right) \left( \tilde{s} \cdot \mathbf{r} \right) - \omega_0 t \right]
\]

\[
= E_0 \exp\left[+i \omega_0 \left( \frac{n}{c} \tilde{s} \cdot \mathbf{r} - t \right) \right] \exp \left[ -\kappa \frac{\omega_0}{c} \left( \tilde{s} \cdot \mathbf{r} \right) \right]
\]

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

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

so that the electric field may be simplified to:

\[
E_0 \exp\left[ i \omega_0 \left( \frac{n}{c} r - t \right) \right] \exp \left[ -\kappa \frac{\omega_0}{c} r \right] \equiv E_0 \exp\left[ i \omega_0 \left( \frac{n}{c} r - t \right) \right] \exp \left[ -\frac{r}{\delta} \right]
\]
where:

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

where \( \lambda_0 \) is the wavelength measured in vacuum. The quantity \( \delta \) has dimensions of \( \frac{m}{s} \) 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 decrease in the amplitude with distance means that the electric field is attenuated as it propagates. If \( r = 2\delta \), the amplitude is reduced by \( e^{-2} \approx 0.135 \) of the incident value.

If the imaginary part \( \kappa \) of the refractive index is zero, so that there is no attenuation, then the electric field is

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

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

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

In the case of complex-valued refractive index, the magnetic field is obtained from the electric field via:

\[ \mathbf{B} = \frac{k \times \mathbf{E}}{\omega} = \frac{n}{c} (\hat{s} \times \mathbf{E}) \]

where \( \hat{s} \) is the unit vector in the direction of the Poynting vector. Note that some authors write \( \hat{n} = n (1 + i\kappa) \), where \( \kappa \) is the attenuation index.

Values of the refractive index for common materials include:

<table>
<thead>
<tr>
<th>Medium</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>1.0 (by definition)</td>
</tr>
<tr>
<td>air</td>
<td>( \approx 1.00027 ) at S.T.P.</td>
</tr>
<tr>
<td>water</td>
<td>( \approx 1.33 )</td>
</tr>
<tr>
<td>“crown” glass</td>
<td>( \approx 1.5 )</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 ) (only transparent for IR, ( \lambda \gtrsim 2\mu m ))</td>
</tr>
</tbody>
</table>

### 6.10 Consequences of Maxwell’s Equations

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

\[ \mathbf{S} = \left( c^2 \epsilon \right) \cdot (\mathbf{E} \times \mathbf{B}) - \text{the Poynting vector} \]
In other words, the wave requires both electric and magnetic fields to propagate, and they 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.

2. The electric and magnetic fields of an electromagnetic wave are mutually perpendicular.

3. 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).

4. Both \( E \) and \( B \) travel at \( c \), the phase velocity of the wave.

5. Energy is carried by both the electric and magnetic fields, and the magnitude of the energy \( \mathcal{E} \propto E_0^2 \).

6. 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 \([\infty \geq \lambda \geq 0]\).

7. The average power of the light wave per unit area is the “irradiance,” and is determined from the Poynting vector

\[
I \left[ x, y, z, t \right] = \langle S \left[ x, y, z, t \right] \rangle = \frac{1}{\Delta T} \int_{t-\Delta T}^{t+\Delta T} S \left[ x, y, z, t' \right] dt'
\]

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

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

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

\[
= \hat{z} \left( c \varepsilon E_0^2 \right) \cdot \frac{1}{2}
\]

\[
\text{Relationship between } E \text{ and } B \text{ for a linearly polarized wave traveling from left to right; the fields are in phase.}
\]
6.11 Dispersion Redux

Earlier we considered the effects of dispersion on traveling waves from a simple point of view where we just assumed that waves with different temporal frequencies might travel at different speeds. We called the dispersion “normal” if the velocity of waves with longer wavelengths exceeds that of waves with shorter wavelengths, which also means that the index of refraction locally decreases with increasing wavelength. At this point, we will try to develop a physical picture of the process to help understand why this occurs. This will also assist in the understanding of conditions necessary for anomalous dispersion (where the index of refraction locally increases with increasing wavelength). In other words, we will see how forces generated by the electric field upon electrons in matter in turn induces the electrons to oscillate and radiate their own propagating electric fields; the process is called “scattering.”

The first theory of dispersion, based on the understanding of elastic solids, was put forth by Cauchy in 1836. He observed a relationship between the phase velocity of light in a medium and the elasticity $\varepsilon$ of the solid (the restoring force exerted upon a displaced particle by a neighboring particle) and the density $\rho$ of the medium:

$$v_\phi \propto \frac{\sqrt{\varepsilon}}{\rho}$$

if measured at wavelengths much longer than the scale of the vibrating particles in the medium. Cauchy deduced a dispersion formula that bears his name:

$$n \approx A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4}$$

where $A$, $B$, and $C$ are constants determined from measurements of $n$ at three wavelengths.

The phenomenon of anomalous dispersion may have been observed (but not pursued) by William Fox Talbot. The first significant study was performed by Le Roux in 1862, who discovered that the index of refraction of a prism containing iodine vapor was $n = 1.020$ for red light and $n = 1.019$ for blue light, so that $n_{\text{blue}} < n_{\text{red}}$, contrary to the behavior of other materials. Within 10 years, Christiansen noted similar behavior in an aniline dye that exhibited a strong absorption of green light, normal refraction for red, orange, and yellow light, but a smaller refraction angle for blue light. The refractive index of a medium that exhibited strong absorption was seen to increase rapidly as the wavelength decreased to approach the absorption band. It took some time to produce a theory of matter and light that explained this effect.

The reason for dispersion is due to the interaction of light with matter. Light can be absorbed by matter, where the energy of the light wave is converted to energy of some form in the matter (e.g., it may increase the thermal, and thus mechanical, energy in the matter). Light also can be scattered, where the electric charges in the matter (protons in the nuclei or, more usually, electrons in the atomic shells) absorb and then re-emit electromagnetic waves. If the scattering is elastic, then no energy is transferred to the medium, all of it stays in the electromagnetic wave. If inelastic,
some energy is transferred to the medium. Scattering generally occurs when the waves encounter a structure or obstacle whose dimension is smaller than a wavelength, and the wave is re-emitted into a new direction. Electromagnetic waves are scattered by electric charges that may be bound within an atom or may be “free” (unbound). A transparent material (e.g., glass) contains many charges, and the macroscopic effect is the sum of the effects from each individual charge.

The interactions of light with matter are characterized by two numerical factors: the “absorption coefficient” $\alpha$ and the (possibly complex valued) “refractive index” $n$, which both may be functions of the frequency of the incident light. The absorption is due to transfer of energy from the light to the medium; at frequencies where the absorption coefficient is small, the light can penetrate the matter to a significant depth, and thus the matter is “transparent.”

Light impinging on a medium causes the charged particles (protons in atomic nuclei or electrons in atomic shells) to vibrate at the oscillation frequency of the light. These accelerated charges emit light of that same oscillation frequency in turn (this is the scattered light). The relative phase of the incident light and the re-emitted light determines much of the effect of the medium on the incident light. For example, if the incident and scattered waves are out of phase by $\pm 180^\circ$ in some direction, then the light beam propagating in that direction will be attenuated.

The oscillations damp out because the electrons are influenced by other forces, including the electric forces induced by neighboring charges. Thus the interaction of light with the medium acts like a damped harmonic oscillator that is driven by the sinusoidal force induced by the light wave. We can think of the interaction of light with charged particles in matter as a so-called “driven” or “forced” harmonic oscillator whose amplitude decreases with time due to the damping. The charges are “bound” to fixed equilibrium locations and can oscillate with one or more resonant frequencies determined by the internal forces due to neighboring atoms. The electric charges can absorb and re-emit the light (i.e., scatter it) in ANY direction. If the frequency of the light is close to the oscillation frequency of one of the resonant states, then some of the electromagnetic energy is retained by the charge and not scattered; it instead increases the energy of the charge. The wave then loses some amplitude when scattered. If the frequency of the light is far from a resonant frequency of the charges, then the scattered light constructively interferes along the same direction as the incident light and it can pass through the medium; in other words, the medium is transparent for light with those frequencies. Even in transparent media, the phase of the re-emitted light generally is shifted by the interaction of the medium.

The first question to consider is the reason why the light is “forward scattered” in transparent media. We’ve actually already given the answer; the light emitted in this direction by all atoms interferes constructively and light emitted in other directions interferes destructively. The charges in matter that scatter the light may be viewed as uniformly distributed and quite close together (separated by fractions of nanometers, significantly less than visible wavelengths). The phase “lag” of the forward-scattered light is equivalent to a “slowing down” of the light wave, hence the index of refraction.
6.11.1 Feynman’s Model of Optical Scattering by Matter


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.

The source point is assumed to be a large distance away (to the left) and the observation point also is a large distance to the right (thus the figure is “not to scale!”). The electric field at the measurement point is the sum of the original electric field $E_s$ plus a correction term $E_2$ that results from emission by the stimulated electric charges in the glass plate. We can assume that the only absorbing (and emitting) charges are electrons, since the protons in the atomic nucleus are much more massive. The electrons in the glass oscillate under the influence of the incoming electric field from the source and this oscillation generates another electric field from each electron – the scattered light. The observed field includes contributions from the incident field and from all oscillating electrons. The modification to the observed field occurs in a way that makes the electric field inside the glass “appear” to be moving at a different phase velocity – this is the reason why the index of refraction of glass is larger than one.

Consider one electron in the glass; it “feels” the effect of the incident field and the fields generated by all of the other electrons in the glass, just as the motions of all of these other charges are influenced by the one electron that we are observing. To simplify the problem, we assume that the influences of the other electrons are small relative to the effect of the source, 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 that travels in the same direction as the incident field ($E_2$) and a field that travels in the opposite direction ($E_1$ - the “reflected” field), but the latter is small because $n \approx 1$. 
Because its source is far away, the incident electric field is a traveling plane wave that may be written in complex notation:

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

We will not use the notation “\( \omega_0 \)” for the temporal frequency of the incident (“driving”) field because it can be varied and because we will reserve that notation for the frequency of “natural oscillation.” Assume that \( z = 0 \) at the “front” (input) side of the plate and \( z = \Delta z \) at the back side, so that the phase of the electric field at the front of the plate is only a function of the observation time:

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

If there were no glass, then the phase at the back of the plate would be:

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

where the phase increment is due to the delay during the transfer of the extra distance \( \Delta z \). If the glass “slows down” the light so that the phase velocity in the glass is \( v_\phi < c \), then the phase at the rear of the glass will include an additional factor:

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

If we substitute the index of refraction, then the distance parts of the phase are proportional

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

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

Thus the “additional” phase due to the extra time to travel through the glass is:

\[ \phi [\Delta z, t; \text{glass}] - \phi [\Delta z, t'; \text{vacuum}] = \omega \left( \frac{n \cdot \Delta z}{c} - t \right) - \omega \left( \frac{\Delta z}{c} - t \right) = \omega \frac{(n - 1) \cdot \Delta z}{c} \]

and the electric field at the back of the glass is:

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

In this way, the contribution of the action of the glass plate to the electric field may be interpreted as an additive contribution to the phase instead of an additive
6.11 DISPERSION REDUX

contribution to the amplitude. This interpretation is much easier to analyze. The second exponential term with the factor of $\Delta z$ may be expanded in a Taylor series:

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

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

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

This is inserted in the expression for the electric field at a large distance “behind” the glass plate:

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

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

$$
= E_s + E_2
$$

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

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

and the second term is identified as the contribution from the charges within the glass, what is labeled in the Figure as $E_2$:

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

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 out of phase with the original electric field by $-\frac{\pi}{2}$ radians. The vector (phasor) contributions of the two fields are shown in the figure.
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. Again, we assume that the incident field has the form of a plane wave:

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

At the “front edge” of the glass, the field is:

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

The electrons in the glass “feel” this field and are driven in the same direction as the field oscillations (assume to be vertical, though this is not essential) by the force:

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

These electrons have mass $m$ and act as though bound to the atoms by little springs that exert restoring forces proportional to the distance of the electron from its equilibrium position:

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

The “normal oscillating frequency” of the electron + spring system is found from the spring force constant and the mass

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

so that the equation of motion of the electrons is:

$$m \frac{d^2 x}{dt^2} + m \omega_0^2 x = F = e E_0 \exp [-i \omega t]$$

where the last term is the “driving force” due to the electric field. We solve this equation by standard methods of differential equations, by assuming that the position $x$ also oscillates at the same rate:

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

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

$$\frac{d^2 x}{dt^2} = (-i \omega)^2 x_0 \exp [-i \omega t] = -\omega^2 x$$
Therefore, the equation of motion simplifies to:

\[
m \left( -\omega^2 x_0 \right) \exp \left[ -i\omega t \right] + m\omega_0^2 x_0 \exp \left[ -i\omega t \right] = eE_0 \exp \left[ -i\omega t \right]
\]

\[
\implies \left( -\omega^2 + m\omega_0^2 \right) x_0 = eE_0
\]

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

Therefore the motions of ALL individual charges in the glass plate due to the incident electric field are identically described by the simple expression:

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

Note that this does not include the initial positions of the charges, 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 \( x \left[ t \right] \). 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:

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

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:

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

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

The total field at the observation point is the vector sum of the contributions from the individual electrons, which may be integrated in polar coordinates. If \( \eta \) is the number density of the electrons per unit area in the glass, then the electric field is:

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

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

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

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

\[
\implies r \, dr = \rho \, d\rho
\]
The integral in the expression for $E_{all}$ becomes:

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

Note that the term $\exp \left[ +i \cdot \infty \right]$ oscillates VERY rapidly, so that we assume its effect on a charge to be negligible. Thus the electric field due to all charges is:

\[
E_{all} = (2\pi \eta e \left( -\omega^2 x_0 \right) \exp \left[ -i \omega t \right]) \cdot \frac{c}{i\omega} \left( 0 - \exp \left[ +i \omega \frac{z}{c} \right] \right)
= 2\pi \eta e \left( -\omega^2 x_0 \right) \exp \left[ -i \omega t \right] \cdot -\frac{c}{i\omega} \exp \left[ +i \omega \frac{z}{c} \right]
= (2\pi \eta ce) \cdot (-i\omega) x_0 \exp \left[ +i \omega \frac{z}{c} \right] \exp \left[ -i \omega t \right]
\]

which shows that the field measured at the observation point due to all of the oscillating charges is out of phase by $\frac{\pi}{2}$ 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:

\[
E_2 = (2\pi \eta ce) \cdot (-i\omega) \left( \frac{eE_0}{m \left( \omega_0^2 - \omega^2 \right)} \right) \exp \left[ -i \omega t \right] \exp \left[ +i \omega \frac{z}{c} \right]
= -i \left( \frac{2\pi \eta ce^2}{m} \right) \left( \frac{\omega}{\omega_0^2 - \omega^2} \right) \exp \left[ +i \omega \left( \frac{z}{c} - t \right) \right]
\]

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 before (in the box)

\[
E_2 \approx -i\omega \left( \frac{n - 1}{c} \right) \cdot \Delta z \left( E_0 \exp \left[ +i \omega \left( \frac{z}{c} - t \right) \right] \right)
= -i\omega \left( \frac{2\pi \eta ce^2}{m \left( \omega_0^2 - \omega^2 \right)} \right) \left( \frac{\omega}{\omega_0^2 - \omega^2} \right) \exp \left[ +i \omega \left( \frac{z}{c} - t \right) \right]
\]
If we equate the common factors, we obtain an expression for the index of refraction in terms of the parameters and the driving frequency $\omega$:

$$\frac{(n-1)\Delta z}{c} = \frac{(2\pi \eta c^2 e^2)E_0}{m(\omega_0^2 - \omega^2)}$$

$$\Rightarrow (n - 1) \Delta z = \frac{(2\pi \eta c^2 e^2)E_0}{m(\omega_0^2 - \omega^2)}$$

$$\Rightarrow n = 1 + \frac{(2\pi c^2 e^2)E_0}{m(\omega_0^2 - \omega^2)} \cdot \frac{\eta}{\Delta z}$$

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

$$N = \eta \cdot \Delta z \Rightarrow \frac{\eta}{\Delta z} = N$$

to obtain the final expression:

$$n = 1 + \frac{(2\pi Nc^2 e^2)E_0}{m(\omega_0^2 - \omega^2)}$$

This is the frequency dependent index of refraction in the simple model of oscillating bound electrons. It shows that the index is largest for light with temporal frequencies in the vicinity of the frequency of the “resonant” oscillations of the electrons due to the restoring forces. However, this picture is incomplete, because the electron motion actually is “damped” out by other forces in the glass.

Illustration of the behavior of the refractive index $n = 1 + \frac{(2\pi Nc^2 e^2)E_0}{m(\omega_0^2 - \omega^2)}$ where the constants are assumed to be $\frac{(2\pi Nc^2 e^2)E_0}{m(\omega_0^2 - \omega^2)} = 0.5$ and $\omega_0 = 3$. Note the discontinuity in the index at the resonance.
6.11.2 Forces on the Charge

Consider the model of the restoring force on a spring for the scattering electron with $e$:

$$F = -kx = e |E| = eE$$

where the negative sign indicates that the force tends to return the system to equilibrium. The bound electrons have no preferred direction; the medium is isotropic. If the electron is displaced by the action of the incident light by a distance $x$, then the restoring force is:

$$F = -ex$$

If there are $N$ such electrons, then the total force $P$ is:

$$P = -Nex$$

where $P$ is the polarization of the medium and may be related to the electric field if the field is static (invariant with time):

$$P = -Nex = -Ne \left( -\frac{eE}{k} \right) = \frac{Ne^2}{k} E$$

If there are no damping forces, the equation of motion (including the acceleration and restoring force) is:

$$m \frac{d^2x}{dt^2} + kx = -e |E| = -eE$$

We know from physics that the motion is an oscillation. If damping is present, the equation of motion must be modified to include a term that decays in proportion to the velocity (akin to friction):

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

The incident field $E_0$ oscillates sinusoidally; call it $E_0 \exp [-i\omega t] \implies x = x_0 \exp [-i\omega t]$. Substitute in and evaluate the derivatives:

$$m \frac{d^2}{dt^2} \exp [-i\omega t] + \alpha \frac{d}{dt} \exp [-i\omega_0 t] + k \exp [-i\omega_0 t] = \left( -m\omega^2 - i\omega + k \right) \exp [-i\omega t] = -eE_0 \exp [-i\omega t] \implies \left( -m\omega^2 - i\omega + k \right) = -eE_0$$

Substitute this into the polarization:

$$P = -Nex = \frac{Ne^2}{k} E = \frac{Ne^2}{-m\omega^2 - i\omega + k} E$$
If $\omega = 0$ (no incident light), the polarization reverts to the static value. Now substitute the constant of the system, the resonant frequency:

$$\omega_0^2 = \frac{k}{m}$$

$$\implies P = \left( \frac{Ne^2}{\omega_0^2 - \omega^2 - i\omega \alpha / m} \right) \frac{E}{m}$$

This is the formula for the frequency response of the driven damped harmonic oscillator.

### 6.11.3 Complete 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.*

In addition, the slope of the dispersion curve decreases with increasing $\lambda$:
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.

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 $\lambda = 0$, the refractive index $n = 1$. In words, the wavelength is so short and the frequency so large that the light propagates through the material as though it were not there. Put yet another way, the energies of the photons is so large that they pass through the substance without interacting.

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 phenomenon also is the reason why X rays can totally reflect at grazing incidence, and thus provides the focusing mechanism in X-ray telescopes, such as Chandra. As the wavelength continues to increase, but still lies within the X-ray region, the light incident on the material will be heavily absorbed. This is the “K-absorption edge” where the energies of the X rays is just sufficient to ionize an electron in the innermost atomic “shell” – the “K shell.” The wavelength of this absorption is $\lambda \approx 0.67$ 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 will exhibit several such absorptions.

At longer wavelengths, into the “far ultraviolet” region of the spectrum, the refractive index will decrease to a value much less than one within a wide band of anomalous dispersion (and thus of absorption). The fact that $n < 1$ in this region seems confusing, because it appears that the velocity of light exceeds $c$, but light with this wavelength cannot propagate in the material due to the strong absorption. 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 $\lambda$ as shown above. For example, the index of air is about 1.000279 at $\lambda = 486.1$ nm (Fraunhofer’s “F” line) and 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 that the nearest ultraviolet absorption is 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.

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 $\lambda \cong 8.5 \mu$m, but the absorption already is quite strong for wavelengths as short as $\lambda \cong 4 \mu$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, which varies over the interval $2.40 \leq n \leq 2.14$ for $51 \mu$m $\leq \lambda \leq 63 \mu$m. This means 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 $\sqrt{\varepsilon}$, where $\varepsilon$ is the permittivity (or dielectric constant) of the material.

**Dispersion curve of a material over wavelengths from very short out to very long.**

The index of refraction in for visible wavelengths is shown in bold face.

**Empirical Expressions for $n[\lambda]$**

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 $n \approx a$ as $\lambda \to +\infty$. The observation that the index decreases with increasing $\lambda$ demonstrates that $b > 0$.

**Cauchy’s Equation**  Cauchy came up with a better empirical relation for the refractivity with more free parameters:

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

where the parameters $A$, $B$, and $C$ are determined by measurements of $n$ for the corresponding material. Again, $A$ is the refractive index for long wavelengths ($\lambda \to +\infty$). It obviously is necessary to measure $n$ at three wavelengths to determine the three empirical constants $A$, $B$, and $C$. Again, the behavior of normal dispersion ensures that $A$ and $B$ are both positive. The dispersion may be expressed as the derivative of $n$ with respect to wavelength $\lambda$:

$$\frac{dn}{d\lambda} \approx -2 \frac{B}{\lambda^3} - 4 \frac{C}{\lambda^5} + \cdots \approx -2 \frac{B}{\lambda^3} \text{ if } C \leq B$$

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

**Sellmeier’s Equation**  These two approximations of $n [\lambda]$ 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 $\lambda_0 = \frac{2\pi\nu_0}{\omega_0}$ 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.
6.11 DISPERSION REDUX

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 in their wavelengths:

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

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

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

**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 $abcdef$, where $n_D = 1.abc$, to three decimal places, and $\nu = def$. 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: \( n_D - 1 = 0.51666 \cong 0.517 \)
- Abbé number: \( \nu = \frac{1.51666 - 1}{1.52225 - 1.51418} \cong 64.0 \)

\[ \text{Glass number} = 517640 \]

For the flint glass:
- refractivity: \( n_D - 1 = 0.70100 \cong 0.701 \)
- Abbé number: \( \nu = \frac{0.70100 - 1}{1.71748 - 1.69427} \cong 30.2 \)

\[ \text{Glass number} = 701302 \]

### 6.11.4 Optical Path Length

Because the phase velocity of light in a medium is less than that in vacuum, light takes longer to travel through a given thickness of material than through the same
“thickness” of vacuum. For a fixed distance $d$, we know that:

\[ d = v \cdot t \quad (\text{distance} = velocity \times time) \]
\[ = c \cdot t_1 \quad (\text{in vacuum}) \]
\[ = \frac{c}{n} \cdot t_2 \quad (\text{in medium of index } n) \]

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

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 $nd = ct_2$ in vacuum. The distance $nd$ traveled in vacuum in the equivalent time is the \textit{optical path length} in the medium.

### 6.11.5 Dispersion via Convolution (Optional)

We have seen how the forces generated by the electric field upon electrons in matter and makes the electrons oscillate and radiate another electric wave. The stimulation provided by the incident field may be interpreted as a temporal “driving force” applied to a system consisting of an bound electron. If we assume that the response of the electron doubles if the amplitude of the field is doubled, then the system satisfies the condition for \textit{linearity}. If the temporal response of the system depends only on the time since the stimulus and not on any “absolute” measurement of time, then the system satisfies the condition for \textit{shift invariance}. Clearly the electron cannot respond until the stimulus occurs, and therefore the system is \textit{causal}. The action of the incident light on matter may be specified by the response to a single “pulse” of light, which may be scaled in amplitude and delayed in time to construct the response of the system to any incident electric field. In other words, the action of the electron may be specified by a the response of the system to an impulse of light; this is the \textit{impulse response} $h[t]$ of the system and must be \textit{causal} so that $h[t] = 0$ for $t < 0$. This interpretation allows the application of the principles of linear and shift-invariant systems to optical dispersion.

The response $g[t]$ of any linear shift-invariant system with impulse response $h[t]$ to the input $f[t]$ may be expressed as a \textit{convolution integral}, which has the mathematical form:

\[ g[t] = \int_{-\infty}^{+\infty} f[\alpha] \cdot h[t - \alpha] \, d\alpha \]

The system is completely characterized by the form of the impulse response $h[t]$, which we know to be zero for $t < 0$ so that the system cannot “respond” until stimulated by $f[t]$. An appropriate impulse response for a damped system rises from 0 at $t = 0$, oscillates sinusoidally with some temporal frequency $\nu_0$, and decays exponentially with time constant $T_0$, (i.e., the time required for the output response to decrease by $e^{-1} \approx 0.368$):

\[ h[t] = A_0 \exp \left[ -\frac{t}{T_0} \right] \cdot STEP[t] \cdot \sin [2\pi \nu_0 t] \]
Figure 6.2: Impulse response of damped harmonic oscillator: \( h[t] = \exp[-\gamma t] \cdot \text{STEP}[t] \cdot \sin[2\pi\nu_0 t] \) where \( A_0 = 1, \nu_0 = 2, \) and \( \gamma_0 = \frac{1}{2}. \)

It often is more convenient to substitute a damping coefficient \( \gamma_0 = T_0^{-1} \) and express the sinusoidal response in terms of the even cosine function:

\[
h[t] = A_0 \exp[-\gamma_0 t] \cdot \text{STEP}[t] \cdot \cos \left[ 2\pi\nu_0 t - \frac{\pi}{2} \right]
\]

where \( \gamma_0 \) is the damping coefficient and \( \nu_0 \) is the natural oscillating frequency of the charged particle (the electron). The \text{STEP} function ensures the causality of the system. In words, when stimulated by a Dirac delta function \( \delta[t] \), the output of the system “grows” from zero with a sinusoidal shape, but the envelope of the sine wave is constrained by the decaying exponential \( \exp[-\gamma_0 t] \), as shown:

The 1-D temporal Fourier transform of this expression yields the frequency response of this system. It may be evaluated rather easily by applying known Fourier transforms:

\[
\mathcal{F}_1 \left\{ e^{-t} \cdot \text{STEP}[t] \right\} = \frac{1}{1 + 2\pi i\nu} = \frac{1 + i(-2\pi\nu)}{1 + (2\pi\nu)^2}
\]

\[
\mathcal{F}_1 \left\{ \sin[2\pi\nu_0 t] \right\} = i \cdot \frac{1}{\nu_0} \left( \delta[\nu + \nu_0] - \delta[\nu - \nu_0] \right)
\]

and the scaling theorem:

\[
\mathcal{F}_1 \left\{ f[t] \right\} = F[\nu] \iff \mathcal{F}_1 \left\{ f[\gamma_0 t] \right\} = \frac{1}{|\gamma_0|} F \left[ \frac{\nu}{\gamma_0} \right]
\]
6.11 DISPERSION REDUX

The transfer function of this system is:

\[ H[\nu] = A_0 F_1 \{ \sin [2\pi \nu_0 t] \cdot (\exp [-\gamma_0 t] \cdot \text{STEP}[t]) \} \]
\[ = A_0 F_1 \{ \sin [2\pi \nu_0 t] \cdot (\exp [-\gamma_0 t] \cdot \text{STEP}[\gamma_0 t]) \} \]
\[ = i \left( \frac{A_0}{2 |\gamma_0|} \right) (\delta[\nu + \nu_0] - \delta[\nu - \nu_0]) \cdot \frac{1}{1 + 2\pi i \left( \frac{\nu}{\gamma_0} \right)} \]
\[ = i \left( \frac{A_0}{2 |\gamma_0|} \right) (\delta[\nu + \nu_0] - \delta[\nu - \nu_0]) \cdot \frac{1 - 2\pi i \left( \frac{\nu}{\gamma_0} \right)}{1 + \left( \frac{2\pi \nu}{\gamma_0} \right)^2} \]

The real and imaginary parts of the frequency response are:

\[ \text{Re}\{H[\nu]\} = \frac{A_0 \pi}{\gamma_0^2} \left( \frac{\nu + \nu_0}{1 + \left( \frac{2\pi \nu + \nu_0}{\gamma_0} \right)^2} - \frac{\nu - \nu_0}{1 + \left( \frac{2\pi \nu - \nu_0}{\gamma_0} \right)^2} \right) \]
\[ = \frac{A_0}{2 \gamma_0^2} \left( \frac{\omega + \omega_0}{\gamma_0^2 + (\omega + \omega_0)^2} - \frac{\omega - \omega_0}{\gamma_0^2 + (\omega - \omega_0)^2} \right) \]
\[ \text{Im}\{H[\nu]\} = \frac{A_0}{2 \gamma_0} \left( \frac{1}{\gamma_0^2 + (\omega + \omega_0)^2} - \frac{1}{\gamma_0^2 + (\omega - \omega_0)^2} \right) \]

The graphs of the real part, imaginary part, magnitude, and phase are shown below, where the domain is assumed to include negative temporal frequencies. In this example, \( \nu_0 = 2 \) and \( \gamma_0 = 0.5 \).
In words, the transfer function measures the “response” of the system in that it evaluates the complex amplitude (magnitude and phase) of the oscillation of the charged particle, which we called \( x(t) \) previously. The resulting amplitude again is a function of the temporal frequency \( \nu \) of the incident light, i.e., the frequency at which the system is “driven.” Note that the phase of the transfer function is approximately 0 radians for \( |\nu| < |\nu_0| \), approximately \( \pm \pi \) radians for \( |\nu| > |\nu_0| \), and \( \pm \frac{\pi}{2} \) radians if the frequency of the incident light is \( \pm \nu_0 \). This means that the system response (the oscillation of the charged particle) is “in phase” if the frequency of the incident light is less than the “resonant frequency” of the oscillating charge. The oscillation of the charged particle is “out of phase” if the frequency of the incident light is larger than the resonant frequency. Also note from the magnitude that the system response is quite large near resonance, which means that the oscillation of the charged particle is large. Of course the meaning of light with a negative temporal frequency is not very clear in this context, and can be rectified by recognizing that the real and imaginary part of the response of the system are related due to causality, a fact reflected in the Kramers-Kronig equations, which are beyond the scope of this discussion.

As mentioned previously, the index of refraction decreases with increasing wavelength (increases with increasing temporal frequency), thus the phase velocity in a medium of light with longer wavelengths is larger and the dispersion is normal. The phase velocity of the modulation wave (the group velocity) is less than the phase velocity of the average wave, and messages travel more slowly than the carrier wave that conveys the message.

In the vicinity of an absorption due to the resonance of charged particles in the
medium, the index of refraction increases with increasing wavelength over a small range, which means that shorter wavelengths travel faster and the dispersion is anomalous. In this region, the phase velocity of the modulation wave is larger than the phase velocity of the average wave. This implies that messages can travel faster than the velocity of light. HOWEVER, since this only happens where light is absorbed, the message cannot propagate.

(a) Real part and (b) imaginary part of refractive index in vicinity of a “weak” absorption. The x-axis is the temporal frequency $\nu$ of the incident light increasing from left to right. This is (of course) proportional to the reciprocal of the wavelength $\lambda$, which therefore increases from right to left.

6.12 Dual Nature of Light: Photons

In many contexts, the “particle” picture of light is more appropriate. In imaging, for example, consider images created with different exposure times. Photographs taken in with shorter exposures generally look grainier:

Images created from increasing numbers of photons, showing increase in signal-to-noise ratio.
A “particle of light” is the photon, whose energy is proportional to the temporal frequency:

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

where \( h \) is Planck’s constant, which is often normalized by a factor of \( 2\pi \), called “h-bar.”

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

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

If \( \lambda = 550 \text{ nm} \), the energy per photon is only:

\[ E = (6.625 \times 10^{-34} \text{ J s}) \cdot \frac{3 \times 10^8 \text{ m/s}}{550 \text{ nm}} \cong 3.6 \times 10^{-19} \text{ J} \]

The “photon flux” is the number of photons per second in a light beam:

\[ \Phi = \frac{P}{h\nu}, \text{ where } P \text{ is the power} \]

Typical fluxes per unit area for some sources are shown in the table:

<table>
<thead>
<tr>
<th>Light Source</th>
<th>( \Phi / A ) [photons/sec-mm²]</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 tells something about the source. Random (incoherent) light sources (such as light bulbs) emit photons with random arrival times and a Bose-Einstein distribution. Coherent light sources, on the other hand, emit photons with a Poisson distribution, which is more uniform but still random.
6.12.1 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.

6.13 Optical Frequencies – Detector Response

The general equation for a traveling electromagnetic wave is:

\[ y[z, t] = A_0 \cos[kz - \omega t] = A_0 \cos \left[ \frac{2\pi}{\lambda}(z - \nu t) \right] = A_0 \text{Re} \left\{ e^{iz} \right\} \]

We see electromagnetic radiation with detectors, i.e., devices which respond in some way to incident electromagnetic radiation. The human eye is sensitive only to visible light, i.e., light with wavelengths in the range \(400\,\text{nm} \leq \lambda \leq 700\,\text{nm}\). This is not the case for all life, however. The pit viper can see radiation emitted by humans at a wavelength of about \(10\,\mu\text{m}\); it needs special receptors on the sides of its head to do this.

As shown in the plot of the electromagnetic spectrum, the frequencies of visible wavelengths are quite large: \(\nu \approx 10^{15}\,\text{Hz}\). The temporal period of an optical wave is therefore \(T = \nu^{-1} \approx 10^{-15}\,\text{s}\). Human visual receptors cannot respond fast enough to detect the periodic oscillation of the wave amplitude; we see an invariant brightness. Note that this limitation exists for all detectors of visible radiation (e.g., photographic film, light meters, etc.); they all respond to the average brightness. The same is
true for hearing; your ear cannot detect the variation of sound pressure due to the oscillation at frequencies above a few Hz. Because water waves have a much lower frequency, the amplitude and phase of the wave can be measured. Similarly, the phase can be measured of electromagnetic waves that have a much smaller temporal frequency, e.g., radio waves.

The average amplitude of a sinusoidal wave is:

\[
\langle y[z, t] \rangle = \frac{1}{T_d} \int_0^{T_d} y[z, t] dt \\
= \frac{1}{T_d} \int_0^{T_d} A_0 \cos[kz - \omega t] dt \\
= -\frac{A_0}{\omega T_d} \sin[kz - \omega t] \bigg|_{t=0}^{t=T_d}
\]

Since \(y[z, t]\) is sinusoidal, the average value of the wave will tend to zero unless \(T_d\) is smaller than the wave’s temporal period. However, the intensity (squared-magnitude) of the wave does not average to zero:

\[
E \propto \langle y^2[z, t] \rangle = \frac{1}{T_d} \int_0^{T_d} y^2[z, t] dt \\
= \frac{1}{T_d} \int_0^{T_d} A_0^2 \cos^2[kz - \omega t] dt \\
= \frac{A_0^2}{T_d} \int_0^{T_d} \cos^2[kz - \omega t] dt \\
= \frac{A_0^2}{T_d} \cdot \frac{T_d}{2} = \frac{A_0^2}{2}
\]

\[
\Rightarrow E \propto \langle y^2[z, t] \rangle \Rightarrow \frac{A_0^2}{2} \quad \text{if} \ T_d > > \nu^{-1}
\]

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

*Detectors of visible light are sensitive to time-averaged intensity, not amplitude.*
Chapter 7

Propagation of Light Waves

7.1 Plane Waves

The form of any wave (matter or electromagnetic) is determined by its source and described by the shape of its wavefront, which is defined as the locus of points in the field that have the same phase, and thus were emitted by the source at the same instant of time. If a traveling wave is emitted by a planar source, then the points of constant phase form a plane surface parallel to the face of the source. Such a wave is called a plane wave, and travels in one direction (ideally). Since energy is conserved, the total energy in the wave must equal the energy emitted by the source, and therefore the energy density (the energy passing through a unit area), is constant for a plane wave. Recall that in a wave of amplitude $A$ and frequency $\omega$, the energy $E \propto A^2 \omega^2$. Therefore, for a plane wave, the amplitude is constant; the wave does not attenuate.

Plane wave toward $z = +\infty$ at velocity $v_0 = \frac{\omega}{k}$, wavelength $\lambda = \frac{2\pi}{k}$, frequency $\nu = \frac{\omega}{2\pi}$, amplitude $A_0$:

$$f[x, y, z, t] = A_0 \cos[kz - \omega t]$$

(n.b., no variation in $y$ or $z$)

General 3-D plane wave traveling in a direction $\mathbf{k} = [k_x, k_y, k_z]$, $\mathbf{r} = [x, y, z]$ and the definition of the scalar product (dot product):

$$f[\mathbf{r}, t] = A_0 \cos[\mathbf{k}\cdot\mathbf{r} - \omega t] \implies \mathbf{k}\cdot\mathbf{r} = k_x x + k_y y + k_z z$$
7.2 Cylindrical Waves

If a wave is emitted from a line source, then the wavefronts are cylindrical. Since the wave expands to fill a cylinder of radius \( r_0 \), the wavefront crosses a cylindrical area that grows as \( \text{Area} = 2\pi rh \propto r \). Therefore, since energy is conserved, the energy per unit area must decrease as \( r \) increases:

\[
\frac{\mathcal{E}}{\text{Area}} = \text{constant} = \frac{\mathcal{E}}{2\pi rh} \propto \frac{\mathcal{E}}{r} \propto \frac{A_0^2}{r} = \text{constant}
\]

\[\Rightarrow \text{amplitude } \propto \frac{A_0}{\sqrt{r}}\]

The equation for a cylindrical wavefront emerging from (or collapsing into) a line source is:

\[
f[x, y, z, t] = A[r] \cos[kr - \omega t] = \frac{A_0}{\sqrt{r}} \cos[kr - \omega t])
\]

\[r = \sqrt{x^2 + y^2} > 0\]

“−” \( \Rightarrow \) emerging

“+” \( \Rightarrow \) collapsing

\( A_0 = \text{amplitude at } r = 0 \)
7.3 Spherical Waves

The wavefront emerging from (or collapsing into) a point is spherical. The area the wave must cross increases as \( x^2 + y^2 + z^2 = r^2 \) (area of sphere is \( 4\pi r^2 \)). Therefore the energy density drops as \( r^2 \) and the amplitude of the wave must decrease as \( \frac{1}{r} \). The equation for a spherical wave is

\[
f \left[ x, y, z, t \right] = f \left[ r, t \right] = A \left[ r \right] \cos \left[ kr \mp \omega t \right] = \frac{A_0}{r} \cos \left[ kr \mp \omega t \right], \text{ where } r > 0
\]

“−” \( \rightarrow \) emerging
“+” \( \rightarrow \) collapsing

\( A_0 = \text{amplitude at } r = 0 \)

Note the pattern for the amplitude of plane, cylindrical, and spherical waves:

\[
\begin{align*}
\text{plane wave} & \quad \Rightarrow \quad 2\text{-D source (plane)} \quad \Rightarrow \quad \text{amplitude } A \left[ r \right] \propto r^{-0} = 1 \\
\text{cylindrical wave} & \quad \Rightarrow \quad 1\text{-D source (line)} \quad \Rightarrow \quad A \left[ r \right] \propto r^{-\frac{1}{2}} \\
\text{spherical wave} & \quad \Rightarrow \quad 0\text{-D source (point)} \quad \Rightarrow \quad A \left[ r \right] \propto r^{-1}
\end{align*}
\]

Spherical waves expanding from a point source.
CHAPTER 7 PROPAGATION OF LIGHT WAVES

7.4 Huygens’ Principle

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’ model provides a very useful model for light propagation that naturally leads to expressions for “diffracted” light that will be considered later.
Chapter 8

Interaction of Light and Matter

8.1 Fermat’s Principle

Hero of Alexandria hypothesized the model of light propagation that could be called the principle of least distance:

A ray of light traveling between two arbitrary points traverses the shortest possible path in space.

This statement applies to reflection and transmission through homogeneous media (i.e., the medium is characterized by a single index of refraction). However, Fermat’s principle is not valid if the object and observation points are located in different media (i.e., the normal situation for refraction) or if multiple media are present between the points.

In 1657, Pierre Fermat modified Hero’s statement to formulate the principle of least time:

A light ray travels the path that requires the least time to traverse.

The laws of reflection and refraction may be easily derived from Fermat’s principle. A moving ray (or car, bullet, or baseball) traveling a distance $s$ at a velocity $v$ requires $t$ seconds:

$$t = \frac{s}{v}$$

If the ray travels at different velocities for different increments of distance, the travel time may be written as:

$$t = \sum_{m=1}^{M} \frac{s_m}{v_m}$$

We know that the velocity of a light ray in a medium of index $n$ is $v = \frac{c}{n}$, so that:

$$t = \sum_{m=1}^{M} \frac{s_m}{(\frac{c}{n_m})} = \frac{1}{c} \sum_{m=1}^{M} (n_m s_m) = \frac{\ell}{c}$$
Thus the time traveled to traverse the path through the medium is equal to the time required to travel a *longer path* $\ell$ in vacuum; the path is longer because $n_m \geq 1$. This longer path $\ell = ns$ is called the *optical path length*. This means that the light requires the least time to traverse the path with the shortest optical path length. The principle of least time may be reworded as:

*A ray traverses the route with the shortest optical path length.*

This result may be derived from Maxwell’s equations.

**Fermat’s Principle for Reflection**

Now consider the path traveled upon reflection that minimizes an easily evaluated optical path length:

As drawn, the angle $\theta_1$ is positive (measured from the normal to the ray) and $\theta_2$ is negative (from the normal to the ray). The ray travels in the same medium of index $n$ both before and after reflection. The components of the optical path length are:

$$s\bar{0} = \sqrt{h^2 + x^2}$$
$$\bar{0}p = \sqrt{b^2 + (a - x)^2}$$

And the expression for the total optical path length $\ell$ is:

$$\ell = n (s\bar{0} + \bar{0}p)$$
$$= n \left( \sqrt{h^2 + x^2} + \sqrt{b^2 + (a - x)^2} \right)$$

*which is a function of $x$*

By Fermat’s principle, the path length traveled is the minimum of the optical path length $\ell$, so the position of $O$ may be find by setting the derivative of $\ell$ with respect
to $x$ to zero:

$$\frac{d\ell}{dx} = n \left( \frac{2x}{2\sqrt{h^2 + x^2}} + \frac{-2(a-x)}{2\sqrt{b^2 + (a-x)^2}} \right) = 0$$

$$= \frac{x}{\sqrt{h^2 + x^2}} - \frac{a-x}{\sqrt{b^2 + (a-x)^2}}$$

$$\Rightarrow \frac{x}{\sqrt{h^2 + x^2}} = \frac{a-x}{\sqrt{b^2 + (a-x)^2}}$$

from the drawing, note that

$$\sin[\theta_1] = \frac{x}{\sqrt{h^2 + x^2}}$$
$$\sin[-\theta_2] = \frac{a-x}{\sqrt{b^2 + (a-x)^2}}$$

$$\Rightarrow \sin[\theta_1] = \sin[-\theta_2]$$
$$\Rightarrow -\theta_1 = \theta_2$$

In words, the magnitudes of the angles of incidence and reflection are equal (as already derived by evaluating Maxwell’s equations at the boundary). The negative sign is necessary because of the sign convention for the angle; the angle is measured from the normal and increases in the counterclockwise direction.

### 8.1.1 Fermat’s Principle for Refraction:

In this drawing, both $\theta_1$ and $\theta_2$ are positive (measured from the normal to the ray). The optical path length is:

$$\ell = n_1 \vec{s} \cdot \vec{o} + n_2 \vec{o} \cdot \vec{p}$$

$$= n_1 \sqrt{h^2 + x^2} + n_2 \sqrt{b^2 + (a-x)^2}$$
By Fermat’s principle, the path length traveled is the minimum of \( \ell \), so we again set the derivative of \( \ell \) with respect to \( x \) to zero:

\[
\frac{d\ell}{dx} = n_1 \cdot \frac{2x}{2\sqrt{h^2 + x^2}} + n_2 \cdot \frac{-2(a-x)}{2\sqrt{b^2 + (a-x)^2}} = 0
\]

\[
\Rightarrow n_1 \cdot \frac{x}{\sqrt{h^2 + x^2}} = n_2 \cdot \frac{a-x}{\sqrt{b^2 + (a-x)^2}} = 0
\]

\[
\sin[\theta_1] = \frac{x}{\sqrt{h^2 + x^2}} \\
\sin[\theta_2] = \frac{a-x}{\sqrt{b^2 + (a-x)^2}}
\]

\[
\Rightarrow n_1 \sin[\theta_1] = n_2 \sin[\theta_2] \\
\Rightarrow \text{Snell’s Law for refraction}
\]

Note that with this sign convention, Snell’s law may be applied to reflection by setting the refractive index of the second medium to be the negative of the first:

\[
n_1 \sin[\theta_1] = n_2 \sin[\theta_2] \\
\Rightarrow n_1 \sin[\theta_1] = -n_1 \sin[\theta_2] \\
\Rightarrow - \sin[\theta_1] = \sin[\theta_2] \\
\Rightarrow \theta_2 = -\theta_1
\]
8.2 Electromagnetic Waves at an Interface

8.2.1 Definition of Vectors for Waves at Interface

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 $\Delta \phi = \pm \pi$ 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 intensities of the three waves;
4. relative phases of the three light waves; and
5. states of polarization of the three waves.

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 the $n^{th}$ medium is:

$$|\mathbf{k}_n| = \frac{2\pi}{\lambda_n} = \frac{2\pi}{\left(\frac{\lambda_0}{n}\right)} = 2\pi \frac{n}{\lambda_0}$$
where $\lambda_0$ is the wavelength in vacuum and $\lambda_n$ 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 $\mathbf{k}_0$, the reflected vector $\mathbf{k}_r$, the transmitted (refracted) vector $\mathbf{k}_t$ and the unit vector $\hat{n}$ normal to the interface are shown:

The $\mathbf{k}$ vectors of the incident, reflected, and “transmitted” (refracted) wave at the interface between two media of index $n_1$ and $n_2$ (where $n_2 > n_1$ in the example 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 (with $n = n_1$) and so have the same wavelength and their $\mathbf{k}$ vectors have the same magnitude:

$$|\mathbf{k}_0| = |\mathbf{k}_r| = \frac{\omega_0}{v_1} = \frac{2\pi n_1}{\lambda_0}$$

$$\lambda_1 = \frac{2\pi n_1}{|\mathbf{k}_0|} = \frac{2\pi n_1}{|\mathbf{k}_r|}$$

The wavelength of the transmitted (refracted) beam is different because of the different index of refraction:

$$\lambda_2 = \frac{2\pi n_2}{|\mathbf{k}_t|}$$

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:
\[
E_{\text{incident}} = E_0 \exp \left[ +i (k_0 \cdot r - \omega_0 t) \right]
\]
where \( r = [x, y, z] \) is the position vector of the location where the phase \( k_0 \cdot r - \omega_0 t \) 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:
\[
E_{\text{reflected}} = E_r \exp \left[ +i (k_r \cdot r - \omega_r t + \phi_r) \right]
E_{\text{transmitted}} = E_t \exp \left[ +i (k_t \cdot r - \omega_t t + \phi_t) \right]
\]
where we have yet to demonstrate that \( \omega_r = \omega_t = \omega_0 \). 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 \( r = 0 \) and \( t = 0 \).

### 8.3 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 \( t \):
\[
(k_0 \cdot r - \omega_0 t)\big|_{z=0} = (k_r \cdot r - \omega_r t + \phi_r)\big|_{z=0} = (k_t \cdot r - \omega_t t + \phi_t)\big|_{z=0}
\]
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:
\[
(k_0 \cdot r)\big|_{z=0} = (k_r \cdot r + \phi_r)\big|_{z=0} = (k_t \cdot r + \phi_t)\big|_{z=0}
\]
Since the scalar products of the three wavevectors with the same position vector \( r \) 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 match at the interface to ensure that each produces the same number of waves per unit length.

From the definitions of the wavevectors we can also see that:

\[
(k_0)_x = |k_0| \cos \left( \frac{\pi}{2} - \theta_0 \right) = |k_0| \sin [\theta_0]
\]

\[
(k_r)_x = |k_r| \cos \left( \frac{\pi}{2} - \theta_r \right) = |k_r| \sin [\theta_r]
\]

where the factor of $-1$ on the reflected angle arises from the fact that the angle measured from the normal is clockwise, and hence negative. The equality of the lengths of the incident and reflected wavevectors immediately demonstrates that:

\[
(k_0)_x = (k_r)_x = |k_0| \sin [\theta_0] = |k_r| \sin [\theta_r]
\]

\[
\Rightarrow |k_0| \sin [\theta_0] = |k_r| \sin [\theta_r]
\]

\[
\Rightarrow \sin [\theta_0] = \sin [\theta_r]
\]

\[
\Rightarrow \theta_0 = -\theta_r
\]

In words, the angle of reflection is equal to the negative of the angle of incidence. We usually ignore the sign of the angle and say that the angles of incidence and reflection are equal.

Now make the same observation for the transmitted wave:

\[
(k_0)_x = |k_0| \sin [\theta_0] = \frac{2\pi n_1}{\lambda_0} \sin [\theta_0]
\]

\[
(k_t)_x = |k_t| \cos \left( \frac{\pi}{2} - \theta_t \right) = |k_t| \sin [\theta_t] = \frac{2\pi n_2}{\lambda_0} \sin [\theta_t]
\]
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.....

### 8.4 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 “polarization” of the electric field (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 (refracted) waves. Faraday's and Ampere'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:

\[
\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad \nabla \times \mathbf{B} = +\epsilon \mu \frac{\partial \mathbf{E}}{\partial t}
\]

where \( \epsilon \) and \( \mu \) are the permittivity and permeability of the medium, respectively and the phase velocity of light in the medium is:

\[
v_\phi = \sqrt{\frac{1}{\epsilon \mu}}
\]

The incident field is assumed to be a plane wave of the form already mentioned:

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

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 \( k_{0y} = 0 \):

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

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:

The boundary conditions on the electric and magnetic fields at the boundary are established from these situations.

We assume that there is no charge or current on the surface and within the cylinder that straddles the boundary. If the height \( dh \) of the cylinder is decreased towards zero, then Gauss’ laws demonstrate that the flux of the electric and magnetic fields
through the top and bottom of the cylinder (the \( z \) 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 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 \( \mathbf{k}_0 \) (the so-called “s” polarization or transverse electric (TE) waves), which also means that the electric field vector is “parallel” to the interface, and (2) the polarization is parallel to the plane of incidence defined by \( \hat{n} \) and \( \mathbf{k}_0 \) (the so-called “p” polarization or transverse magnetic (TM) waves). 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” polarization).

The electric field is parallel to the plane of incidence; this is the TRANSVERSE MAGNETIC field (TM, also called the “p” polarization).

8.5 Transverse Electric Waves, “s” or “⊥” Polarization

In the TE case in our geometry, the electric field is oriented along the $y$ direction and the wavevector has components in the $x$ and $z$ directions:

$$E_{\text{incident}}[x, y, z, t] = (\hat{x} \cdot 0 + \hat{y} \cdot |E_0| + \hat{z} \cdot 0) \exp (+i (k_{0x}x + k_{0z}z - \omega_0t))$$

$$= \hat{y}E_0 \exp (+i (k_{0x}x + k_{0z}z - \omega_0t))$$

The magnetic field is derived from the relation:

$$B = \frac{n}{c}k_0 \times E$$

$$B_{\text{incident}}[x, y, z, t]$$

$$= \left( -\cos[\theta_0] \cdot n_1 \frac{|E_0|}{c} \right) \hat{x} + 0 \hat{y} + \left[ +\sin[\theta_0] \cdot n_1 \frac{|E_0|}{c} \right] \hat{z} \exp (+i (k_{0x}x + k_{0z}z - \omega_0t))$$

The reflected fields are:

$$E_{\text{reflected}}[x, y, z, t] = \hat{y} \cdot |E_0| \exp (+i (k_{r_x}x + k_{r_z}z - \omega_0t))$$
8.5 TRANSVERSE ELECTRIC WAVES, “S” OR “\perp” POLARIZATION

\[ \mathbf{B}_{\text{reflected}}[x, y, z, t] = \left( + \cos[-\theta_0] \cdot n_1 \frac{|\mathbf{E}_0|}{c} \hat{x} + \left[ - \sin[-\theta_0] \cdot n_1 \frac{|\mathbf{E}_0|}{c} \right] \hat{z} \right) \cdot \exp \left[ +i \left( k_0 x + k_0 z - \omega t \right) \right] \]

\[ = \left( + \cos[\theta_0] \cdot n_1 \frac{|\mathbf{E}_0|}{c} \hat{x} + \left[ \sin[\theta_0] \cdot n_1 \frac{|\mathbf{E}_0|}{c} \right] \hat{z} \right) \cdot \exp \left[ +i \left( k_0 x + k_0 z - \omega t \right) \right] \]

and the transmitted (refracted) fields are:

\[ \mathbf{E}_{\text{transmitted}}[x, y, z, t] = \hat{y} \cdot \mathbf{E} \cdot \exp \left[ +i \left( k_{tx} x + k_{tz} z - \omega t \right) \right] \]

\[ \mathbf{B}_{\text{transmitted}}[x, y, z, t] = \left( - \cos[\theta_1] \cdot n_2 \frac{|\mathbf{E}_1|}{c} \hat{x} + \left[ \sin[\theta_1] \cdot n_2 \frac{|\mathbf{E}_1|}{c} \right] \hat{z} \right) \cdot \exp \left[ +i \left( k_0 x + k_0 z - \omega t \right) \right] \]

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):

\[ r_{TE} \equiv \frac{E_r}{E_0} \]
\[ t_{TE} \equiv \frac{E_t}{E_0} \]

where the subscripts denote the transverse electric polarization. The boundary condition for the normal magnetic field yields the expression:

\[ \frac{n_1}{c} \sin[\theta_0] (E_0 + E_r) = \frac{n_2}{c} \sin[\theta_1] E_t \]

while that for the tangential magnetic field:

\[ \frac{n_1}{\mu_1 c} \cos[\theta_0] (E_0 - E_r) = \frac{n_2}{\mu_2 c} \cos[\theta_1] E_t \]

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

\[ r_{TE} = \frac{E_r}{E_0} = \frac{n_1}{\mu_1} \cos[\theta_0] - \frac{n_2}{\mu_2} \cos[\theta_1] \]
\[ \frac{n_1}{\mu_1} \cos[\theta_0] + \frac{n_2}{\mu_2} \cos[\theta_1] \]
CHAPTER 8 INTERACTION OF LIGHT AND MATTER

Transmission Coefficient for TE Waves

\[
T = \frac{E_t}{E_0} = \frac{\mu_1}{\mu_1 + \mu_2} \cos \theta_0 + \frac{\mu_2}{\mu_1 + \mu_2} \cos \theta_1
\]

if \( \mu_1 = \mu_2 \) (usual case)

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:

\[
R_{TE} = \left( \frac{n_1 \cos \theta_0 - n_2 \cos \theta_1}{n_1 \cos \theta_0 + n_2 \cos \theta_1} \right)^2
\]

The transmission \( T \) 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 beam along the \( x \)-axis is larger in the medium with the larger index, and thus the area of the transmitted beam is larger:

Demonstration that the cross-sectional area of the beam is changed by refraction at
8.5 TRANSVERSE ELECTRIC WAVES, “S” OR “⊥” POLARIZATION

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:

$$T = \frac{|\mathbf{s}_2| \cdot A_2}{|\mathbf{s}_1| \cdot A_1} = \frac{\left(n_2 |E_t|^2\right) \cdot A_2}{\left(n_1 |E_0|^2\right) \cdot A_1} = \frac{n_2}{n_1} \cdot t^2 \cdot \frac{A_2}{A_1}$$

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:

$$\frac{A_2}{A_1} = \frac{w_2}{w_1} = \frac{\sin \left[\frac{\pi}{2} - \theta_t\right]}{\sin \left[\frac{\pi}{2} - \theta_0\right]} = \frac{\cos \left[\theta_t\right]}{\cos \left[\theta_0\right]}$$

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:

$$T = \frac{n_2}{n_1} \cdot t^2 \cdot \left(\frac{\cos \left[\theta_t\right]}{\cos \left[\theta_0\right]}\right)$$

Snell’s law gives a relationship between the incident and transmitted angles:

$$n_1 \sin \left[\theta_0\right] = n_2 \sin \left[\theta_t\right] \implies \sin \left[\theta_t\right] = \frac{n_1}{n_2} \sin \left[\theta_0\right]$$

$$\implies \cos \left[\theta_t\right] = \sqrt{1 - \sin^2 \left[\theta_t\right]} = \sqrt{1 - \left(\frac{n_1}{n_2} \sin \left[\theta_0\right]\right)^2}$$

Thus we can write down the transmittance $T$ in terms of the refractive indices and the incident angle:

$$T = \left(\frac{\sqrt{n_2^2 - n_1^2 \sin^2 \left[\theta_0\right]}}{n_1 \cos \left[\theta_0\right]}\right) \cdot t^2$$

For the TE case, the transmission is:

$$T_{TE} = \left(\frac{\sqrt{n_2^2 - n_1^2 \sin^2 \left[\theta_0\right]}}{n_1 \cos \left[\theta_0\right]}\right) \cdot \left(\frac{+2n_1 \cos \left[\theta_0\right]}{n_1 \cos \left[\theta_0\right] + n_2 \cos \left[\theta_t\right]}\right)^2$$
These will be plotted for some specific cases after we evaluate the coefficients for TM waves.

8.6 Transverse Magnetic Waves ("p" or "||" 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:

$$\mathbf{E}_{\text{incident}}[x, y, z, t] = (\hat{x} \cdot \mathbf{E}_0) \cos \theta_0 + \hat{z} \cdot \mathbf{E}_0 \sin \theta_0) \exp [+i (k_0 x + k_0 z - \omega_0 t)]$$

$$= (\hat{x} \cdot \mathbf{E}_0) \cos \theta_0 + \hat{z} \cdot \mathbf{E}_0 \sin \theta_0) \exp [+i (k_0 x + k_0 z - \omega_0 t)]$$

The magnetic field is in the $y$-direction:

$$\mathbf{B}_{\text{incident}}[x, y, z, t] = \left( n_1 \frac{|\mathbf{E}_0|}{c} \right) \exp [+i (k_0 x + k_0 z - \omega_0 t)]$$

The reflected fields are:

$$\mathbf{E}_{\text{reflected}}[x, y, z, t] = (\hat{x} \cdot \mathbf{E}_r) \cos \theta_0 - \hat{z} \cdot \mathbf{E}_r \sin \theta_0) \exp [+i (k_0 x + k_0 z - \omega_0 t)]$$

$$\mathbf{B}_{\text{reflected}}[x, y, z, t] = \left( n_1 \frac{|\mathbf{E}_r|}{c} \right) \exp [+i (k_0 x + k_0 z - \omega_0 t)]$$

and the transmitted (refracted) fields are:

$$\mathbf{E}_{\text{transmitted}}[x, y, z, t] = (\hat{x} \cdot \mathbf{E}_t) \cos \theta_t - \hat{z} \cdot \mathbf{E}_t \sin \theta_t) \exp [+i (k_t x + k_t z - \omega_0 t)]$$

$$\mathbf{B}_{\text{transmitted}}[x, y, z, t] = \left( n_2 \frac{|\mathbf{E}_t|}{c} \right) \exp [+i (k_t x + k_t z - \omega_0 t)]$$

In the case, the boundary condition on the normal component of $\mathbf{B}$ is trivial, but the other components are:

$$\mu_1 \sin \theta_0 (E_0 + E_r) = \mu_2 \sin \theta_2 E_t$$

$$\cos \theta_0 (E_0 - E_r) = \cos \theta_2 E_t$$

$$\frac{n_1}{\mu_1 c} (E_0 + E_r) = \frac{n_2}{\mu_2 c} E_t$$

These are solved for the reflection and transmission coefficients:

Transverse Magnetic Waves

$$r_{TM} = + \frac{n_2}{\mu_2} \cos \theta_0 - \frac{n_1}{\mu_1} \cos \theta_t$$

$$+ \frac{n_2}{\mu_2} \cos \theta_0 + \frac{n_1}{\mu_1} \cos \theta_t$$
which simplifies if the permeabilities $\mu_n$ are equal (as they usually are):

$$r_{TM} = \frac{+n_2 \cos[\theta_0] - n_1 \cos[\theta_t]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_t]} \quad \text{if} \quad \mu_1 = \mu_2$$

The corresponding reflectance is:

$$R_{TM} = \left(\frac{+n_2 \cos[\theta_0] - n_1 \cos[\theta_t]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_t]}\right)^2$$

The amplitude transmission coefficient evaluates to:

$$t_{TM} = \frac{2n_1 \cos[\theta_0]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_t]}$$

again, if the permeabilities are equal, this simplifies to:

$$t_{TM} = \frac{2n_1 \cos[\theta_0]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_t]} \quad \text{if} \quad \mu_1 = \mu_2$$

The corresponding transmittance function is:

$$T_{TM} = \left(\frac{\sqrt{n_2^2 - n_1^2 \sin^2[\theta_0]}}{n_1 \cos[\theta_0]}\right) \cdot \left(\frac{2n_1 \cos[\theta_0]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_t]}\right)^2$$

### 8.7 Comparison of Coefficients for TE and TM Waves

We should compare the coefficients for the two cases of TE and TM waves. The reflectance coefficients are:

$$r_{TE} = \frac{n_1 \cos[\theta_0] - n_2 \cos[\theta_t]}{n_1 \cos[\theta_0] + n_2 \cos[\theta_t]}$$

$$r_{TM} = \frac{+n_2 \cos[\theta_0] - n_1 \cos[\theta_t]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_t]}$$

where the angles are also determined by Snell’s law:

$$n_1 \sin[\theta_0] = n_2 \sin[\theta_t]$$

$$\implies \cos[\theta_t] = \sqrt{1 - \left(\frac{n_1}{n_2} \sin[\theta_0]\right)^2}$$

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. The same condition holds for $n_2$ and $\theta_t$. The opposite is true for the TM case: $n_1$ is applied to $\cos[\theta_t]$ and
$n_2 \text{ to } \cos[\theta_0]$. These same observations also apply to the corresponding transmission coefficients:

$$t_{TE} = \frac{+2n_1 \cos[\theta_0]}{n_1 \cos[\theta_0] + n_2 \cos[\theta_i]}$$

$$t_{TM} = \frac{+2n_1 \cos[\theta_0]}{+n_2 \cos[\theta_0] + n_1 \cos[\theta_i]}$$

### 8.7.1 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:

$$r_{TE}|_{\theta_0=0} = \frac{n_1 - n_2}{n_1 + n_2}$$

$$r_{TM}|_{\theta_0=0} = \frac{+n_2 - n_1}{+n_2 + n_1} = -r_{TE}|_{\theta_0=0}$$

$$t_{TE}|_{\theta_0=0} = \frac{+2n_1}{n_1 + n_2}$$

$$t_{TM}|_{\theta_0=0} = \frac{+2n_1}{n_1 + n_2} = t_{TE}|_{\theta_0=0}$$

cases are identical. Also, the areas of the incident and transmitted waves are identical so there is no area factor in the transmittance. The resulting formulas for reflectance and transmittance reduce to:

**normal incidence ($\theta_0 = 0$)**

$$R_{TE}(\theta_0 = 0) = R_{TM}(\theta_0 = 0) \equiv R = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2$$

$$T = \frac{4n_1n_2}{(n_1+n_2)^2}$$

### Example: “Rare-to-Dense” Reflection

If the input medium has a smaller refractive index $n$ (a rarer medium) than the second (denser) medium, so that $n_1 = 1.0 < n_2 = 1.5$, then the coefficients are:
8.7 COMPARISON OF COEFFICIENTS FOR TE AND TM WAVES

\[ r_{TE} = \frac{1.0 - 1.5}{1.0 + 1.5} = -0.2 = 0.2e^{+i\pi} \]

\[ r_{TM} = \frac{1.5 - 1.0}{1.0 + 1.5} = +0.2 \]

\[ t_{TE} = t_{TM} = \frac{2 \cdot 1.0}{1.0 + 1.5} = +0.8 \]

\[ \Rightarrow R_{TE} = R_{TM} = 0.04 \]

\[ \Rightarrow T_{TE} = T_{TM} = 0.96 \]

for “rare-to-dense” reflection

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.

Example: “Dense-to-Rare” Reflection

If the input medium is “denser” \( (n_1 = 1.5 > n_2 = 1.0) \), then these values are obtained:

\[ r_{TE} = \frac{1.5 - 1.0}{1.5 + 1.0} = +0.2 \]

\[ r_{TM} = \frac{1.0 - 1.5}{1.0 + 1.5} = -0.2 = 0.2e^{+i\pi} \]

\[ t_{TE} = t_{TM} = \frac{2 \cdot 1.5}{1.0 + 1.5} = +1.2 > 1.0? \]

\[ \Rightarrow R_{TE} = R_{TM} = 0.04 \]

\[ \Rightarrow T_{TE} = T_{TM} = 0.96 \]

There is no phase shift of the reflected amplitude in “dense-to-rare” reflection, commonly called “internal” reflection.

8.7.2 Angular Dependence of Reflection and Transmittance at “Rare-to-Dense” Interface

Consider the graphs of these coefficients for the cases of the “rare-to-dense” interface \( (n_1 = 1 < n_2 = 1.5) \). The reflection coefficients are plotted vs. incident angle measured in degrees from 0° (normal incidence) to 90° (grazing incidence).
Amplitude reflectance and transmittance coefficients for $n_1 = 1.0$ (air) and $n_2 = 1.5$ (glass) for both TE and TM waves, plotted as functions of the incident angle from $\theta_0 = 0^\circ$ (normal incidence) to $\theta_0 = 90^\circ$ (grazing incidence). The reflectance coefficient $r_{TE} < 0$ for all $\theta$, which means that there is a phase shift upon reflection, whereas $r_{TM} > 0$ for $\theta_0 < \theta_B$ (Brewster’s angle). Also note that the transmittance coefficients are very similar functions.

8.7.3 Brewster’s Angle – Angle of Complete Polarization

Note that $r_{TM} = 0$ at one particular angle ($\approx 60^\circ$) in the TM case (parallel polarization), which means that no amplitude of this wave is reflected if incident at this angle. In other words, any light reflected at this angle must be the TE wave which is completely polarized perpendicular to the plane of incidence. This is Brewster’s angle, the angle such that the reflected wave and the refracted wave are orthogonal (i.e., $\theta_0 + \theta_t = \frac{\pi}{2} \implies \theta_t = \frac{\pi}{2} - \theta_0$). 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 is sometimes called the angle of complete polarization. Note that the transmitted light contains both polarizations, though not in equal amounts.
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 \( \theta \), and the sum of the incident and refracted angle is \( 90^\circ = \frac{\pi}{2} \). Thus

\[
\theta_B + \theta_i = \frac{\pi}{2} \implies \theta_i = \frac{\pi}{2} - \theta_B.
\]

From Snell’s law, we have:

\[
n_1 \sin [\theta_i] = n_2 \sin [\theta_2]
\]

At Brewster’s angle,

\[
n_1 \sin [\theta_B] = n_2 \sin \left[ \frac{\pi}{2} - \theta_B \right]
\]

\[
= n_2 \left( \sin \left[ \frac{\pi}{2} \right] \cos [\theta_B] - \cos \left[ \frac{\pi}{2} \right] \sin [\theta_B] \right)
\]

\[
= +n_2 \cos [\theta_B]
\]

\[
n_1 \sin [\theta_B] = n_2 \cos [\theta_B]
\]

\[
\implies \frac{n_2}{n_1} = \frac{\sin [\theta_B]}{\cos [\theta_B]} = \tan [\theta_B]
\]

\[
\implies \theta_B = \tan^{-1} \left[ \frac{n_2}{n_1} \right]
\]

If \( n_1 = 1 \) (air) and \( n_2 = 1.5 \) (glass), then \( \theta_B \approx 56.3^\circ \). For incident angles larger than about \( 56^\circ \), the reflected light is plane polarized parallel to the plane of incidence. If the dense medium is water \( (n_2 = 1.33) \), then \( \theta_B \approx 52.4^\circ \). This happens at the interface with any dielectric. 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).
8.7.4 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 the zero reflectance of the TM wave at Brewster’s angle.

\[
R_{TM} = 0 \quad \text{and} \quad T_{TM} = 1 \quad \text{at “Brewster’s angle.”}
\]

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

At a “glass-to-air” interface where \( n_1 > n_2 \), 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 that satisfies:

\[
n_2 \cos \theta_0 = n_1 \cos \theta_t
\]

\[
\frac{n_1}{n_2} = \frac{\cos \theta_0}{\cos \theta_t}
\]

This corresponds to the situation where Snell’s law requires that:

\[
\sin \left[ \theta_t = \frac{\pi}{2} \right] = 1 = \frac{n_1}{n_2} \sin \theta_0 \implies \sin \theta_0 = \frac{n_2}{n_1}
\]
If \( n_1 = 1.5 \) and \( n_2 = 1.0 \), then
\[
\sin \theta_0 = \frac{2}{3} \implies \theta_0 \approx 0.73 \text{ radians} \approx 41.8^\circ \equiv \theta_c
\]

If the incident angle exceeds this value \( \theta_c \), the *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 for \( \theta_0 \geq \theta_c \) is *totally reflected*. This is the source of *total internal reflectance* ("internal" because the reflection is from glass back into glass). The phenomenon of TIR is the reason for the usefulness of optical fibers in communications.

The angular dependences of the amplitude reflection coefficients for the case \( n_1 = 1.5 \) (glass) and \( n_2 = 1.0 \) (air) are shown. Brewster’s angle in this case satisfies:
\[
\theta_B = \tan^{-1} \left( \frac{n_2}{n_1} \right) = \tan^{-1} \left( \frac{2}{3} \right) \implies \theta_B \approx 0.59 \text{ radians} \approx 33.7^\circ
\]

**Amplitude reflectance coefficients for TE and TM waves if \( n_1 = 1.5 \) (glass) and \( n_2 = 1.0 \) (air). Both coefficients 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 situation for \( \theta_0 > \theta_c \) can be interpreted as producing complex-valued \( r_{TE} \) and \( r_{TM} \).**

### 8.7.6 Practical Applications for Fresnel’s Equations

The 4% normal reflectance of one surface of glass is the reason why windows look like mirrors at night when you’re in the brightly lit room. Lasers incorporate end windows oriented at Brewster’s angle to eliminate reflective losses at the mirrors (and also thus producing polarized laser light). Optical fibers use total internal reflection. Hollow fibers use high-incidence-angle near-unity reflections.
Chapter 9

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, \( \text{CaCO}_3 \)). 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 two components of an electromagnetic wave are the electric field \( \mathbf{E} \) \([\text{V/m}]\) and the magnetic field \( \mathbf{B} \) \([\text{tesla T} = \frac{\text{webers m}}{\text{m}^2 \text{wbs}}] \).

The polarization of radiation is defined as the plane of vibration of the electric vector \( \mathbf{E} \), rather than of \( \mathbf{B} \), because the effect of the \( \mathbf{E} \) field on a free charge (an electron) is much greater than the effect of \( \mathbf{B} \). This is seen from the Lorentz equation, or the Lorentz force law:

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

where

- \( q_0 = \) charge \([\text{coulombs}]\)
- \( \mathbf{F} = \) force on the charge \([\text{newtons}, 1 \text{ N} = 1 \frac{\text{kg} \cdot \text{m}}{\text{sec}^2}]\)
- \( \mathbf{v} = \) velocity of the charge \( q_0 \), measured in \([\text{m/sec}]\)
- \( c = \) velocity of light \([3 \cdot 10^8 \frac{\text{m}}{\text{s}}]\)

The factor \( c^{-1} \) ensures that the force on the electron due to the magnetic field is usually much smaller than the electric force.

9.1 Plane Polarization = Linear Polarization

The most familiar type of polarization is linear polarization, where the \( \mathbf{E} \)-vector oscillates in the same plane at all points on the wave.

Any state of linear polarization can be expressed as a linear combination (sum) of two orthogonal states \((\text{basis states})\), e.g., the \( x \)- and \( y \)-components of the \( \mathbf{E} \)-vector for a wave traveling toward \( z = \pm \infty \):
\[
\mathbf{E} = \mathbf{E}[\mathbf{r}, t] = [\hat{x}E_x + \hat{y}E_y] \cos [kz - \omega t]
\]

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

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

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

\[
E_x = E_0 \cos [\theta]
\]

\[
E_y = E_0 \sin [\theta]
\]

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 at the same point in time and space.

---

9.2 Circular Polarization

If the \(\mathbf{E}\)-vector describes a helical (i.e., screw-like) 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 polarization is \textit{circular}:

*Electric field vector \(\mathbf{E}\) and magnetic field vector \(\mathbf{H}\) of a plane-polarized wave.*
Circular polarization occurs when the electric fields along orthogonal axes have the same amplitude by their phases differ by $\pm \frac{\pi}{2}$ radians.

If we sit 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, but with one component out-of-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:

$$motion = \mathbf{x} \cos [\omega t] + \mathbf{y} \cos \left[ \omega t \mp \frac{\pi}{2} \right] = \mathbf{x} \cos [\omega t] \pm \mathbf{y} \sin [\omega t]$$

For a traveling wave:

$$\mathbf{E} = [E_x, E_y] = \left[ E_0 \cos [kz - \omega t], E_0 \cos \left[ kz - \omega t \mp \frac{\pi}{2} \right] \right]$$

$$= [E_0 \cos [kz - \omega t], \pm E_0 \sin [kz - \omega t]]$$

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

### 9.2.1 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): Point the thumb of the hand in the direction of propagation. If the fingers point in the direction of ro-
2. Optics (also called 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.

9.2.2 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 $\pm \pi/2 = \pm 90^\circ$, then the projection of the path of the $\mathbf{E}$-vector is not a circle, but rather an ellipse. This results in elliptical polarization. Note that elliptical polarization may be either right- or left-handed, as defined above.

9.2.3 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:

\[ \text{Change in “handedness” of a circularly (or elliptically) polarized wave upon reflection by a mirror.} \]

9.2.4 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 ($\approx 10^{-8}$ s). The radiation is termed unpolarized, even though it is polarized when viewed within this short time period. Natural light is neither totally polarized nor totally unpolarized; rather, we speak of partial polarization.
9.3 Description of Polarization States

9.3.1 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.

\[
\mathbf{E} = \text{Re}\{E_0 \exp[+i(kz - \omega t)]\} = \text{Re}\{E_x \exp[+i(kz - \omega t)]\}, \text{Re}\{E_y \exp[+i(kz - \omega t - \delta)]\}
\]

\[
= \text{Re}\{[E_x, E_y e^{-i\delta}] \exp[+i(kz - \omega t)]\}
\]

\[\rightarrow \text{Jones Vector } \mathbf{E} = \begin{bmatrix} E_x \\ E_y e^{-i\delta} \end{bmatrix} \]

Examples:

1. Plane-polarized light along x-axis

\[\mathbf{E} = \begin{bmatrix} E_0 \\ 0 \end{bmatrix}\]

2. Plane-polarized light along y-axis:

\[\mathbf{E} = \begin{bmatrix} 0 \\ E_0 \end{bmatrix}\]

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

\[ \mathbf{E} = \hat{\mathbf{x}} E_0 \cos [kz - \omega t] + \hat{\mathbf{y}} E_0 \sin [kz - \omega t] \]

\[ = \hat{\mathbf{x}} E_0 \cos [kz - \omega t] + \hat{\mathbf{y}} E_0 \cos \left[ kz - \omega t - \frac{\pi}{2} \right] \]

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

\[ \implies \mathcal{E} = \text{Re} \left\{ E_0 \begin{bmatrix} 1 \\ \exp \left[ -i\frac{\pi}{2} \right] \end{bmatrix} \exp \left[ +i (kz - \omega t) \right] \right\} \]

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.

### 9.4 Generation of Polarized Light

#### 9.4.1 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.

![Diagram](Image)

*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
9.4 GENERATION OF POLARIZED LIGHT

the mechanism used in the next section.

9.4.2 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 unaffacted.

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.

![Polarization by "skein of wires" – the radiation polarized parallel to the direction of the wires in the skein is absorbed, so the radiation polarized perpendicular to the wires is transmitted. In other words, the "picket fence" model of polarization is not appropriate in this case.]

9.4.3 Generating Polarized Light by Reflection – Brewster’s Angle

H§8.6

The two polarizations of light reflected from an interface between two different dielectric media (i.e., media with different real refractive indices) see the same configuration of the interface only with normal incidence (i.e., the light is incident perpendicular to the surface). Thus the two polarizations must be identically reflected. However, if the light is incident obliquely, one polarization “sees” the bound electrons of the surface differently and therefore is reflected differently. The reflected wave is polarized to some extent; the amount of polarization depends on the angle of incidence and the index of refraction $n$. 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 $E \exp[i (k_0 z_0 \pm \omega_0 t)]$,
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 $\nu$ 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., $\theta_0 + \theta_t = 90^\circ \implies \theta_t = \frac{\pi}{2} - \theta_0$), 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 $\theta_B + \theta_t = \frac{\pi}{2} \implies \theta_t = \frac{\pi}{2} - \theta_B$.

At Brewster’s angle,

$$n_1 \sin[\theta_B] = n_2 \sin \left[ \frac{\pi}{2} - \theta_B \right]$$

$$\implies \theta_B = \tan^{-1} \left[ \frac{n_2}{n_1} \right]$$

If $n_1 = 1$ (air) and $n_2 = 1.5$ (glass), then $\theta_B \cong 56.3^\circ$. For incident angles larger than about $56^\circ$, the reflected light is plane polarized parallel to the plane of incidence. If the dense medium is water ($n_2 = 1.33$), then $\theta_B \cong 52.4^\circ$. This happens at the interface with any dielectric. The reflection at Brewster’s angle provides a handy means to determine the polarization axis of a linear polarizer – just look through the polarizer at light reflected at a steep angle.
9.4.4 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 \textit{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° 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.

\textit{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.}
Chapter 10
Birefringence – Double Refraction

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. The ordinary ray obeys Snell’s law; the extraordinary ray does not. One is called the ordinary ray, because it obeys Snell’s law for refraction. The second, or extraordinary ray, does not obey Snell. By dividing the incoming natural light into two beams in such a crystal, we can select one of the two polarizations.

10.0.5 Examples:
Refractive indices along the fast and slow axes at $\lambda = 589.3$ 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 $H_2O$)</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 $\lambda' = \lambda / n$, so light along the two polarization directions have different wavelengths:

$$\lambda'_s = \frac{\lambda}{n_s} < \lambda'_f = \frac{\lambda}{n_f}$$
10.1 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:

$$\mathbf{E}[z,t] = (\hat{x}E_x + \hat{y}E_y) \exp [+i (kz - \omega t)].$$

At the input face of the material ($z = 0$) and the output face ($z = d$), the fields are:

$$\mathbf{E}[z = 0, t] = (\hat{x}E_x + \hat{y}E_y) \exp [-i\omega t]$$

$$\mathbf{E}[z = d, t] = (\hat{x}E_x + \hat{y}E_y) \exp [+i (kd - \omega t)]$$

If $n_x = n_s > n_y = n_f$, then $\lambda_f > \lambda_s$ and:

$$k_s = k_x = \frac{2\pi n_s}{\lambda} > k_f = k_y = \frac{2\pi n_f}{\lambda}$$

The field at the output face ($z = d$) is therefore:

$$\mathbf{E}[d, t] = \left[ \hat{x}E_x \exp \left[ + i \frac{2\pi d (n_s)}{\lambda} \right] + \hat{y}E_y \exp \left[ + i \frac{2\pi d n_f}{\lambda} \right] \right] e^{-i\omega t}$$

$$= \left( \hat{x}E_x + \hat{y}E_y \exp \left[ 2\pi i d (n_f - n_s) \right] \right) \exp \left[ + i \frac{2\pi d n_f}{\lambda} \right]$$

By defining a constant phase term $\delta \equiv \frac{2\pi}{\lambda} d (n_f - n_s)$, the electric field at the output face of the birefringent material can be expressed as:

$$\mathbf{E}[d, t] = (\hat{x}E_x + \hat{y}E_y e^{i\delta}) \exp \left[ + i \frac{2\pi d n_f}{\lambda} \right]$$

On emergence from the material, the $y$-component of the polarization has a different phase than the $x$-component; the phase difference is $\delta$.

10.1.1 Example:

$\delta = +\frac{\pi}{2} \implies (n_f - n_s) d = -\frac{\lambda}{4}$, and there is a phase difference of one quarter wavelength between the polarizations of the $x$- and the $y$-components of the wave. This is a quarter-wave plate. The required thickness $d$ of the material is:

$$d = \frac{\lambda}{4 (n_s - n_f)}$$

And the emerging field is:

$$\mathbf{E}[d, t] = \left[ \hat{x}E_x + \hat{y}E_y e^{i\frac{\pi}{4}} \right] \exp [i (k_s d - \omega t)]$$
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.

### 10.1.2 Example:

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°.

### 10.2 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 \( \frac{\lambda}{4} \) plate is now linearly polarized perpendicular to the axis of the LP and no light escapes.

A circular polarizer is a sandwich of a linear polarizer and a quarter-wave plate.

### 10.3 Critical Angle – Total Internal Reflection

We also mentioned the critical angle during the discussion of the Fresnel equations. From Snell, we have the relation:

\[
\theta_1 \sin \theta_1 = \theta_2 \sin \theta_2
\]
If \( n_1 > n_2 \) then a specific angle \( \theta_1 \) exists that satisfies the condition where its sine is equal to the ratio of the two indices:

\[
\frac{n_1}{n_2} \sin[\theta_1] = 1 \implies \sin[\theta_1] = \frac{n_2}{n_1} < 1 \implies \theta_2 = \frac{\pi}{2}
\]

This means that the outgoing ray is refracted parallel to the interface (the “surface”). The incident angle \( \theta_1 \) that satisfies this condition is the critical angle \( \theta_c \):

\[
\theta_c = \sin^{-1}\left[\frac{n_2}{n_1}\right]
\]

For crown glass with \( n_d = 1.52 \), the critical angle is \( \sin^{-1}\left[\frac{1}{1.52}\right] \simeq 0.718 \text{ radians} \simeq 41^\circ \). For a common flint glass with \( n_d = 1.70 \), then \( \theta_c \simeq 0.629 \text{ radians} \simeq 36^\circ \). 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 11
Optical Interference

We now return to those thrilling days of waves to consider their effects on the performance of imaging systems. We first consider “interference” of two traveling waves that oscillate with the same frequency. “Interference” refers to the addition of waves from a small number of sources. The waves are either derived from different points on the same source wave (“division of wavefront” interference) or by dividing the light amplitude from the same source wave by using 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 imaging because it is the fundamental limit to the capability of a system to distinguish different objects.

11.1 Division of Wavefront

References: Hecht, Optics §8

The first example of optical interference involves summing two wavefronts that have traveled different paths after having been derived from a wavefront emerging from a single source. This was the first type of interference that was observed, having been noted by Thomas Young in the early 1800s.

To introduce the concept of interference, recall the identity that was derived for the sum of two oscillations with different frequencies $\omega_1$ and $\omega_2$:

\[
y_1[t] = A \cos [\omega_1 t]
\]
\[
y_2[t] = A \cos [\omega_2 t]
\]
\[
y_1[t] + y_2[t] = 2A \cos \left[ \frac{(\omega_1 + \omega_2)}{2} t \right] \cos \left[ \frac{(\omega_1 - \omega_2)}{2} t \right] = 2A \cos [\omega_{avg} t] \cdot \cos [\omega_{mod} t]
\]

In words, the sum of two oscillations of different frequency is identical to the product of two oscillations: one is the slower varying modulation (at frequency $\omega_{mod}$) and the other is the more rapidly oscillating average sinusoid (or carrier wave) with frequency $\omega_{avg}$.
A perhaps familiar example of the modulation results from the excitation of two piano strings that are mistuned. A low-frequency oscillation (the beat) is heard; as one string is tuned to the other, the frequency of the beat decreases, reaching zero when the string frequencies are equal. Acoustic beats may be thought of as interference of the summed oscillations in time.

Now consider this relationship in a broader sense where traveling waves are added, which we also derived before for two waves traveling along the \( z \)-axis:

\[
\begin{align*}
  f_1[z,t] &= A \cos [k_1 z - \omega_1 t] \\
  f_2[z,t] &= A \cos [k_2 z - \omega_2 t] \\
  f_1[z,t] + f_2[z,t] &= \{2A \cos[k_{\text{mod}} z - \omega_{\text{mod}} t]\} \cdot \cos[k_{\text{avg}} z - \omega_{\text{avg}} t]
\end{align*}
\]

\[
\begin{align*}
  k_{\text{mod}} &= \frac{k_1 - k_2}{2} \\
  \omega_{\text{mod}} &= \frac{\omega_1 - \omega_2}{2} \\
  v_{\text{mod}} &= \frac{\omega_{\text{mod}}}{k_{\text{mod}}} = \frac{\omega_1 - \omega_2}{k_1 - k_2}
\end{align*}
\]

\[
\begin{align*}
  k_{\text{avg}} &= \frac{k_1 + k_2}{2} \\
  \omega_{\text{avg}} &= \frac{\omega_1 + \omega_2}{2} \\
  v_{\text{avg}} &= \frac{\omega_{\text{avg}}}{k_{\text{avg}}} = \frac{\omega_1 + \omega_2}{k_1 + k_2}
\end{align*}
\]

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, \textit{i.e.} a traveling modulation. Note that both the average and modulation waves move along the \( z \)-axis. In this case, \( k_1, k_2, \omega_1, \text{and} \omega_2 \) 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. In this case, the modulation wave moves in the opposite direction to the average wave.

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;” it forms a standing
11.1 DIVISION OF WAVEFRONT

wave:

\[
\begin{align*}
    f_1 [z, t] &= A \cos [k_1 z - \omega_1 t] \\
    f_2 [z, t] &= A \cos [k_1 z + \omega_1 t] \\
    f_1 [z, t] + f_2 [z, t] &= \{2A \cos [k_{\text{mod}} z - \omega_{\text{mod}} t]\} \cdot \cos [k_{\text{avg}} z - \omega_{\text{avg}} t]
\end{align*}
\]

\[
\begin{align*}
    k_{\text{mod}} &= \frac{k_1 - k_1}{2} = 0 \\
    \omega_{\text{mod}} &= \frac{\omega_1 - (-\omega_1)}{2} = \omega_1 \\
    k_{\text{avg}} &= \frac{k_1 + k_1}{2} = k_1 \\
    \omega_{\text{avg}} &= \frac{\omega_1 + (-\omega_1)}{2} = 0
\end{align*}
\]

\[
\begin{align*}
    f_1 [z, t] + f_2 [z, t] &= 2A \cos [k_1 z] \cos [-\omega_1 t] \\
    &= 2A \cos [k_1 z] \cos [\omega_1 t]
\end{align*}
\]

where the symmetry of \(\cos[\theta]\) was used in the last step.

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 \(k\); a 3-D wavevector \(\mathbf{k}\) has components \([k_x, k_y, k_z]\). The vector may be written:

\[
\mathbf{k} = [k_x \hat{x} + k_y \hat{y} + k_z \hat{z}]
\]

The corresponding wave travels in the direction of the wavevector \(\mathbf{k}\) and has wavelength \(\lambda_0 = \frac{2\pi}{|\mathbf{k}|}\). In other words, the length of \(\mathbf{k}\) is the magnitude of the wavevector:

\[
|\mathbf{k}| = \sqrt{k_x^2 + k_y^2 + k_z^2} = \frac{2\pi}{\lambda_0}.
\]

The temporal oscillation frequency \(\omega\) is determined from the magnitude of the wavevector through the dispersion relation:

\[
\omega = v_\phi \cdot |\mathbf{k}| \implies \nu = \frac{v_\phi}{\lambda_0}
\]

For illustration, consider a simple 2-D analogue of the 1-D traveling plane wave. The wave travels in the direction of the 2-D wavevector \(\mathbf{k}\) which is in the \(x - z\) plane:

\[
\mathbf{k} = [k_x, 0, k_z]
\]

The points of constant phase with with phase angle \(\phi = C\) radians is the set of points
in the 2-D space \( \mathbf{r} = [x = 0, y, z] = (r, \theta) \) such that the scalar product \( \mathbf{k} \cdot \mathbf{r} = C \):

\[
\mathbf{k} \cdot \mathbf{r} = \mathbf{r} \cdot \mathbf{k} \\
= |\mathbf{k}| |\mathbf{r}| \cos \theta \\
= k_x x + k_z z = C \text{ for a point of constant phase}
\]

Therefore, the equation of a 2-D wave traveling in the direction of \( \mathbf{k} \) with linear wavefronts is:

\[
f [x, y, t] = A \cos [k_x x + k_z z - \omega t] \\
= A \cos [\mathbf{k} \cdot \mathbf{r} - \omega t]
\]

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:

\[
f [x, y, z, t] = f [\mathbf{r}, t] \\
= A \cos [k_x x + k_y y + k_z z - \omega t] \\
= A \cos [\mathbf{k} \cdot \mathbf{r} - \omega t]
\]

This plane wave could have been created by a point source at a large distance to the left and below the \( z \)-axis.

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:

$$f_1[x, y, z, t] + f_2[x, y, z, t] = A \cos [k_1 \cdot \mathbf{r} - \omega t] + A \cos [k_2 \cdot \mathbf{r} - \omega t]$$

$$= 2A \cos [k_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t] \cdot \cos [k_{\text{mod}} \cdot \mathbf{r} - \omega_{\text{mod}} t]$$

where the average and modulation wavevectors are:

$$k_{\text{avg}} = \frac{k_1 + k_2}{2} = \frac{(k_x)_1 + (k_x)_2}{2} \mathbf{\hat{x}} + \frac{(k_y)_1 + (k_y)_2}{2} \mathbf{\hat{y}} + \frac{(k_z)_1 + (k_z)_2}{2} \mathbf{\hat{z}}$$

$$k_{\text{mod}} = \frac{k_1 - k_2}{2} = \frac{(k_x)_1 - (k_x)_2}{2} \mathbf{\hat{x}} + \frac{(k_y)_1 - (k_y)_2}{2} \mathbf{\hat{y}} + \frac{(k_z)_1 - (k_z)_2}{2} \mathbf{\hat{z}}$$

and the average and modulation angular temporal frequencies are:

$$\omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2}$$

$$\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2}$$

Note that the average and modulation wavevectors $k_{\text{avg}}$ and $k_{\text{mod}}$ point in different directions, in general, 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 (the wavevector $k$ and the angular temporal frequency $\omega$), the phases of two traveling waves usually differ even if the temporal frequencies are equal. Consider the superposition of two such waves:

$$\omega_1 = \omega_2 \equiv \omega$$

The component waves travel in different directions so the components of the wavevectors differ:

$$k_1 = [(k_x)_1, (k_y)_1, (k_z)_1] \neq k_2 = [(k_x)_2, (k_y)_2, (k_z)_2]$$

Since the temporal frequencies are equal, so must be the wavelengths:

$$\lambda_1 = \lambda_2 = \lambda \rightarrow |k_1| = |k_2| \equiv |k|.$$
cies are:

\[
\omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} = \omega_0 \\
\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2} = 0
\]

The summation of the two traveling waves with identical magnitudes may be expressed as:

\[
f_1[x, y, z, t] + f_{21}[x, y, z, t] = A \cos(k_1 \cdot \mathbf{r} - \omega_0 t) + A \cos(k_2 \cdot \mathbf{r} - \omega_0 t) \\
= 2A \cos(k_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t) \cdot \cos(k_{\text{mod}} \cdot \mathbf{r} - 0 \cdot t) \\
= 2A \cos(k_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t) \cdot \cos(k_{\text{mod}} \cdot \mathbf{r})
\]

Therefore, 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 wave in space along the direction of \(k_{\text{mod}}\) that does not move. This second stationary wave is analogous to the phenomenon of beats, and is called interference in optics.

### 11.1.1 Superposition of Two Same-Frequency Plane Waves

Consider the superposition of two plane waves:

\[
f_1[x, y, z, t] = A \cos(k_1 \cdot \mathbf{r} - \omega_0 t) \\
f_2[x, y, z, t] = A \cos(k_2 \cdot \mathbf{r} - \omega_0 t)
\]

\[
k_1 = [k_x, k_y = 0, k_z] \\
k_2 = [-k_x, 0, k_z]
\]

i.e., the wavevectors differ only in the \(x\)-component, and there only by a sign. Therefore the two wavevectors have the same “length”:

\[
|k_1| = |k_2| = \frac{2\pi}{\lambda} \\
\implies \lambda_1 = \lambda_2 \equiv \lambda.
\]

Also note that:

\[
k_z = |k| \cos[\theta] = \frac{2\pi}{\lambda} \cos[\theta] \\
k_x = \frac{2\pi}{\lambda} \sin[\theta]
\]
These two waves could have been generated at point sources located above and below the $z$-axis a large distance to the left. This is the classic “Young’s double slit” experiment, where light from a single source is split into to waves (spherical waves in this case) and propagate a large distance 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 distances:

$$\theta \approx \frac{d/2}{L} = \frac{d}{2L}.$$ 

If $L \gg d$, then

$$\theta \approx \tan \theta \approx \sin \theta \approx \frac{d}{2L}.$$ 

The superposition of the two electric fields is:

$$f[x, y, z, t] = f_1[x, y, z, t] + f_2[x, y, z, t]$$

$$= 2A \cos \left( k_{avg} \cdot \mathbf{r} - \omega_{avg} t \right) \cdot \cos \left( k_{mod} \cdot \mathbf{r} \right)$$

$$= 2A \cos \left[ 2\pi \frac{z}{\lambda} \cos \theta - \omega_0 t \right] \cos \left[ 2\pi \frac{x}{\lambda} \sin \theta \right]$$

The first term (with the time dependence) is a traveling wave in the direction defined by $\mathbf{k} = [0, 0, k_z]$, while the second term (with no dependence on time) is a spatial
wave along the $x$ direction. The amplitude variation in the $x$ direction is:

$$2A \cos \left[2\pi \frac{x}{\lambda} \sin [\theta]\right] = 2A \cos \left[2\pi \frac{x}{\left(\frac{\lambda}{\sin[\theta]}\right)}\right]$$

which has a period of $\frac{\lambda}{\sin[\theta]}$. The irradiance (the measurable intensity) of the superposition is:

$$|f [x, y, z, t]|^2 = 4A^2 \cos^2 \left[2\pi \frac{z}{\lambda} \cos [\theta] - \omega_0 t\right] \cos^2 \left[\frac{2\pi x \sin [\theta]}{\lambda}\right]$$

The second cosine terms can be rewritten using:

$$\cos^2 [\theta] = \frac{1}{2} (1 + \cos [2\theta])$$

As before, the first term varies rapidly due to the angular frequency term $\omega_0 \approx 10^{14}$ Hz. Therefore, just the average value is detected:

$$\langle |f [x, y, z, t]|^2 \rangle = 4A^2 \cos^2 \left[\frac{2\pi x \sin [\theta]}{\lambda}\right] \cdot \frac{1}{2}$$

$$= 2A^2 \left[\frac{1}{2} \left(1 + \cos \left[\frac{4\pi x \sin [\theta]}{\lambda}\right]\right)\right]$$

$$= A^2 \left(1 + \cos \left[2\pi \frac{x}{\left(\frac{\lambda}{2\sin[\theta]}\right)}\right]\right)$$

This derivation may also be applied to find the irradiance of one of the individual component waves:

$$I_1 = \langle |f_1 [x, y, z, t]|^2 \rangle$$

$$I_2 = \langle |f_2 [x, y, z, t]|^2 \rangle$$

$$I_0 = \langle |f_1 [x, y, z, t]|^2 \rangle = \langle |A \cos [\mathbf{k}_1 \cdot \mathbf{r} - \omega_0 t]|^2 \rangle = A^2 \langle \cos^2 [\mathbf{k}_1 \cdot \mathbf{r} - \omega_0 t]\rangle$$

$$= A^2 \cdot \frac{1}{2}$$
So the irradiance of the sum of the two waves can be rewritten in terms of the irradiance of a single wave:

\[
\langle |f[x, y, z, t]|^2 \rangle = 4I_0 \cos^2 \left( \frac{2\pi x \sin[\theta]}{\lambda} \right)
\]

\[
= 2I_0 \left[ 1 + \cos \left( \frac{2\pi x \cdot 2 \cdot \sin[\theta]}{\lambda} \right) \right]
\]

\[
= 2I_0 \left[ 1 + \cos \left( 2\pi \left( \frac{x}{\lambda} \right) \right) \right]
\]

This is a “biased sinusoid” that is positive or zero everywhere. In words, the irradiance exhibits a nonnegative sinusoidal modulation with period \( X = \frac{\lambda}{2\sin[\theta]} \) and oscillates between 0 and \( 2I_0 \cdot (2) = 4I_0 \) 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 \textit{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 pattern; this is the principle behind \textit{optical holography}. The argument of the cosine function is the \textit{optical phase difference} of the two waves. 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 \textit{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 \textit{destructive interference}. 
11.1 DIVISION OF WAVEFRONT

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.

### 11.1.2 Superposition of Two Plane Waves with Different Frequencies

For further illustration, consider the case the two waves travel in the same directions, so that \( \mathbf{k}_1 \neq \mathbf{k}_2 \), but with different temporal frequencies \( \omega_1 \neq \omega_2 \). This means that \( |\mathbf{k}_1| \neq |\mathbf{k}_2| \). The average and modulation wavevectors are found as before, but the modulation wave now travels (it is not stationary along the \( x \)-axis) because both \( \mathbf{k}_{\text{mod}} \neq \mathbf{0} \) and \( \omega_{\text{mod}} \neq 0 \). Consider the example of two component waves: \( f_1 [r, t] \) directed at an angle \( \theta_1 = +40^\circ \approx \frac{2}{3} \) radian with \( \lambda_1 = 8 \) units and \( \omega_1 = \frac{1}{8} \) radians per second, and \( f_2 [r, t] \) directed at \( \theta_2 = -40^\circ \approx -\frac{2}{3} \) radian with \( \lambda_2 = 12 \) units and \( \omega_2 = \frac{1}{12} \) radians per second. The corresponding average and modulation frequencies are:

\[
\omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} = \frac{2\pi}{2} \left( \frac{1}{8} + \frac{1}{12} \right) = \frac{2\pi}{9.6} = \frac{5\pi}{24} \text{ radians/s}
\]

\[
\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2} = \frac{2\pi}{2} \left( \frac{1}{8} - \frac{1}{12} \right) = \frac{2\pi}{48} \text{ radians/s}
\]
The superposition may be written as the product of the average and modulation waves:

\[ f_1 [\mathbf{r}, t] + f_2 [\mathbf{r}, t] = 2 \cdot f_{\text{avg}} [\mathbf{r}, t] \cdot f_{\text{mod}} [\mathbf{r}, t] \]

where the full expressions for the average and modulation waves are:

\[ f_{\text{avg}} [\mathbf{r}, t] = \cos \left[ \mathbf{k}_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}} t \right] \]

\[ = \cos \left[ \frac{2\pi}{48} \cdot (x \sin [40^\circ] + 5z \cos [40^\circ]) - \frac{10\pi t}{48} \right] \]

\[ = \cos \left[ \frac{2\pi}{48} (x \sin [40^\circ] + 5z \cos [40^\circ]) - 5t \right] \]

\[ \approx \cos \left[ 2\pi \left( \frac{x}{74.674} + \frac{z}{12.532} \right) - 2\pi \left( \frac{t}{9.6} \right) \right] \]

\[ f_{\text{mod}}(\mathbf{r}, t) = \cos (\mathbf{k}_{\text{mod}} \cdot \mathbf{r} - \omega_{\text{mod}} t) \]

\[ = \cos \left[ \frac{2\pi}{48} \cdot (5x \sin [40^\circ] + z \cos [40^\circ]) - \frac{2\pi t}{48} \right] \]

\[ = \cos \left[ \frac{2\pi}{48} (5x \sin [40^\circ] + z \cos [40^\circ]) - 2t \right] \]

\[ \approx \cos \left[ 2\pi \left( \frac{x}{74.674} + \frac{z}{12.532} \right) - 2\pi \left( \frac{t}{24} \right) \right] \]

Note that both the average and modulation waves are traveling waves; they are headed in different directions with different frequencies and different velocities. The temporal frequencies are \( \nu_{\text{avg}} = \frac{5}{48} \) Hz and \( \nu_{\text{mod}} = \frac{2}{48} \) Hz. If the intensity (squared-magnitude) of the sum is averaged over time at an observation plane located downstream on the \( z \)-axis, both traveling waves will average out to create a uniform irradiance pattern; no stationary fringe pattern will be visible.
CHAPTER 11 OPTICAL INTERFERENCE

Sum of two sinusoidal traveling waves where the periods are related by \( \frac{\lambda_2}{\lambda_1} = \frac{1}{12} \).

The two waves travel in the directions \( \pm 40^\circ \), 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.

\[ |k_2| = 2\pi/\lambda_2 \propto 1/8 \]

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.
11.2 Fringe Visibility – Coherence

The visibility of a sinusoidal fringe pattern is a quality that corresponds quite closely to modulation, which is a term used by electrical engineers (sparkies). Given a nonnegative sinusoidal irradiance (intensity) distribution with maximum $I_{\text{max}}$ and minimum $I_{\text{min}}$ (so that $I_{\text{min}} \geq 0$), the visibility of the sinusoidal fringe pattern is:

$$V \equiv \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}$$

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 largely 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 “greene” 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:

$$f_1 [x, z; \lambda_1] + f_2 [x, z; \lambda_1] = 2A \cos \left[ \frac{2\pi x}{\lambda_1} \sin \theta \right] \cos \left[ \frac{2\pi z}{\lambda_1} \cos \theta - \omega_1 t \right]$$

$$= 2A \cos \left[ 2\pi \frac{x}{\left( \frac{\lambda_1}{\sin \theta} \right)} \right] \cos \left[ \frac{2\pi z}{\lambda_1} \cos \theta - \omega_1 t \right]$$

which is the sum of a stationary sinusoid and a traveling wave in the $+z$-direction. If we add a second pair of plane waves with different wavelength $\lambda = \lambda_2$ but the same “tilts,” the amplitude pattern can also be calculated. We have to add the amplitude of four waves, but we can still add them in pairs. The first pair produces the same amplitude pattern that we saw before. The second wave also produces a pattern that differs only in the periods of the sinusoids.
The sum of four tilted plane waves can be calculated by summing the pair due to one wavelength and that due to the other.

The second pair of wavefronts with \( \lambda = \lambda_2 \) yield a similar result, though the period of the stationary fringes and the temporal frequency of the traveling wave differ. The expression for the sum of the two pairs is:

\[
\sum_{n=1}^{4} f_n [x, z; \lambda] = f_1 [x, z; \lambda_1] + f_2 [x, z; \lambda_1] + f_1 [x, z; \lambda_2] + f_2 [x, z; \lambda_2] \\
= 2A \cos \left[ \frac{2\pi x}{(\frac{\lambda_1}{\sin[\theta]})} \right] \cos \left[ \left( \frac{2}{\lambda_1} \cos [\theta] \right) z - \omega_1 t \right] \\
+ 2A \cos \left[ \frac{2\pi x}{(\frac{\lambda_2}{\sin[\theta]})} \right] \cos \left[ \left( \frac{2\pi}{\lambda_2} \cos [\theta] \right) z - \omega_2 t \right] \\
= 2A \cos \left[ \frac{2\pi x}{(\frac{(k_{avg})_1}{\sin[\theta]})} \right] \cos \left[ (k_{avg})_1 z - \omega_1 t \right] + 2A \cos \left[ \frac{2\pi x}{(\frac{(k_{avg})_2}{\sin[\theta]})} \right] \cos \left[ (k_{avg})_2 z - \omega_2 t \right]
\]
11.2 FRINGE VISIBILITY – COHERENCE

where:

\[
\begin{align*}
|k_{avg1}| &= (k_{avg1})_1 = \frac{2\pi}{\lambda_1 \cos \theta} \\
|k_{avg2}| &= (k_{avg2})_2 = \frac{2\pi}{\lambda_2 \cos \theta} \\
(v_{avg1})_1 &= \frac{\omega_1}{(k_{avg1})_1} = \frac{2\pi \nu_1}{\lambda_1} \cos \theta = \frac{\nu_1 \lambda_1}{z} \cos \theta = \frac{c}{z} \cos \theta \\
v_2 &= \frac{\omega_2}{(k_{avg2})_2} = \frac{2\pi \nu_2}{\lambda_2 \cos \theta} = \frac{\nu_2 \lambda_2}{z} \cos \theta = \frac{c}{z} \cos \theta = v_1
\end{align*}
\]

Thus the phase velocities of the two travelling waves are equal.

\[
2A \cos \left[2\pi \frac{x}{\left(\frac{\nu_1}{\sin \theta}\right)}\right] \cos \left[(k_{avg1})_1 (z - \omega_1 t)\right] + 2A \cos \left[2\pi \frac{x}{\left(\frac{\nu_2}{\sin \theta}\right)}\right] \cos \left[(k_{avg2})_2 (z - \omega_2 t)\right]
\]

Because the phase velocities of the two “average” waves are equal, the traveling waves differ only in wavenumber and both travel down the z-axis with the same velocity:

\[
\sum_{n=1}^{4} f_n [x, z; \lambda] = 2A \cos \left[2\pi \frac{x}{\left(\frac{\nu_1}{\sin \theta}\right)}\right] \cos \left[(k_{avg1})_1 (z - \nu_1 t)\right] + 2A \cos \left[2\pi \frac{x}{\left(\frac{\nu_2}{\sin \theta}\right)}\right] \cos \left[(k_{avg2})_2 (z - \nu_2 t)\right]
\]

Here we make a simplifying assumption that is strictly incorrect: we assume that the two traveling wave parts contribute to the amplitude pattern in the same way and can be factored out:

\[
\sum_{n=1}^{4} f_n [x, z; \lambda] = 2A \cos \left[2\pi \frac{x}{\lambda_1} \cos \theta - \omega_1 t\right] \left(\cos \left[2\pi \frac{x}{\left(\frac{\lambda_1}{\sin \theta}\right)}\right] + \cos \left[2\pi \frac{x}{\left(\frac{\lambda_2}{\sin \theta}\right)}\right]\right)
\]

The squared magnitude of the amplitude is:

\[
\left|\sum_{n=1}^{4} f_n [x, z; \lambda]\right|^2 = 4A^2 \left(\cos \left[2\pi \frac{x}{\left(\frac{\lambda_1}{\sin \theta}\right)}\right] + \cos \left[2\pi \frac{x}{\left(\frac{\lambda_2}{\sin \theta}\right)}\right]\right)^2 \left(\cos \left[2\pi \frac{z}{\lambda_1} \cos \theta - \omega_1 t\right]\right)^2
\]
and the time average yields the irradiance:

\[ I_{x, z} = \left\langle \left( \sum_{n=1}^{4} f_n [x, z; \lambda] \right)^2 \right\rangle \]

\[ = 4A^2 \left( \cos \left( 2\pi \frac{x}{\lambda_1 \sin[\theta]} \right) + \cos \left( 2\pi \frac{x}{\lambda_2 \sin[\theta]} \right) \right)^2 \left\langle \cos^2 \left( \frac{2\pi z \cos[\theta] - \omega_1 t}{\lambda_1} \right) \right\rangle \]

\[ = 2A^2 \left( \cos \left( 2\pi \frac{x}{\lambda_1 \sin[\theta]} \right) + \cos \left( 2\pi \frac{x}{\lambda_2 \sin[\theta]} \right) \right)^2 \]

The sum of the two stationary cosine waves also may be recast as the product of cosines with the average and modulation frequencies:

\[ \cos \left( 2\pi \frac{x}{\lambda_1 \sin[\theta]} \right) + \cos \left( 2\pi \frac{x}{\lambda_2 \sin[\theta]} \right) \]

\[ = 2 \cos \left( 2\pi x \sin[\theta] \cdot \left( \frac{\frac{1}{\lambda_1} + \frac{1}{\lambda_2}}{2} \right) \right) \cdot \cos \left( 2\pi x \sin[\theta] \cdot \left( \frac{\frac{1}{\lambda_1} - \frac{1}{\lambda_2}}{2} \right) \right) \]

\[ = 2 \cos \left( \pi x \sin[\theta] \cdot \left( \frac{\lambda_1 + \lambda_2}{\lambda_1 \lambda_2} \right) \right) \cdot \cos \left( \pi x \sin[\theta] \cdot \left( \frac{\lambda_1 - \lambda_2}{\lambda_1 \lambda_2} \right) \right) \]

\[ = 2 \cos \left( \frac{2\pi x \sin[\theta]}{\lambda_1 \lambda_2} \lambda_{avg} \right) \cdot \cos \left( \frac{2\pi x \sin[\theta]}{\lambda_1 \lambda_2} \lambda_{mod} \right) \]

where \( \lambda_{avg} = \frac{\lambda_1 + \lambda_2}{2} \) and \( \lambda_{mod} = \frac{\lambda_1 - \lambda_2}{2} \).

The final expression for the irradiance is the product of two sinusoidal irradiance
patterns with identical maxima and zero minima:

\[ I(x, z) = 2A^2 \cdot \left| 2 \cos \left( \frac{2\pi x \sin \theta}{\lambda_1 \lambda_2} \right) \lambda_{\text{avg}} \right| \cdot \cos \left( \frac{2\pi x \sin \theta}{\lambda_1 \lambda_2} \lambda_{\text{mod}} \right) \right|^{2} = 4A^2 \cos^2 \left( \frac{2\pi x \sin \theta}{\lambda_1 \lambda_2} \lambda_{\text{avg}} \right) \cdot \cos^2 \left( \frac{2\pi x \sin \theta}{\lambda_1 \lambda_2} \lambda_{\text{mod}} \right)

\[ = 4A^2 \left[ 1 + \cos \left( \frac{2\pi x \sin \theta}{\lambda_1 \lambda_2} \lambda_{\text{avg}} \right) \right] \cdot \left( 1 + \cos \left( \frac{2\pi x}{D_{\text{avg}}} \right) \right) \cdot \left( 1 + \cos \left( \frac{2\pi x}{D_{\text{mod}}} \right) \right)

\[ = 4A^2 \cdot \left( 1 + \cos \left( \frac{2\pi x}{D_{\text{avg}}} \right) \right) \cdot \left( 1 + \cos \left( \frac{2\pi x}{D_{\text{mod}}} \right) \right)

where the respective periods \( D_{\text{avg}} \) and \( D_{\text{mod}} \) of the two spatial oscillations are defined:

\[ D_{\text{avg}} = \frac{\lambda_1 \lambda_2}{\lambda_{\text{avg}} \sin |\theta|} = \frac{1}{\sin |\theta|} \left( \frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right)^{-1} \propto |\lambda_1 + \lambda_2|^{-1} \propto (\lambda_{\text{avg}})^{-1}

\[ D_{\text{mod}} = \frac{\lambda_1 \lambda_2}{\lambda_{\text{mod}} \sin |\theta|} = \frac{1}{\sin |\theta|} \left| \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right|^{-1} \propto |\lambda_1 - \lambda_2|^{-1} \equiv (\Delta \lambda)^{-1}

In the case where the two emitted wavelengths are close together such that \( \lambda_1 \approx \lambda_2 \approx \lambda_{\text{avg}} \gg \lambda_{\text{mod}} \), the expressions for the periods of the two component oscillations may be simplified:

\[ D_{\text{avg}} \approx \frac{L \lambda_{\text{avg}}}{2d}

\[ D_{\text{mod}} \approx \frac{L (\lambda_{\text{avg}})^2}{2d \cdot \Delta \lambda}

After cancelling the common terms, the relative lengths of the spatial periods of the modulations are:

\[ D_{\text{mod}} = D_{\text{avg}} \cdot \frac{\lambda_{\text{avg}}}{\Delta \lambda} >> D_{\text{avg}} \text{ if } \Delta \lambda << \lambda_{\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. The period \( D_{\text{mod}} \) limits the range of \( x \) over which the short-period fringes can be seen. 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.
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 \textit{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 where the phase difference remains approximately constant (for \( x \approx 0 \)) over the averaged time interval. Such a source is said to be \textit{temporally incoherent}. It is difficult (though not impossible) to see fringes generated by an incoherent source.

### 11.2.1 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 \textit{bandwidth} of the source:

\[
\Delta \nu \equiv |\nu_1 - \nu_2| = \left| \frac{c}{\lambda_1} - \frac{c}{\lambda_2} \right| = c \cdot \frac{|\lambda_2 - \lambda_1|}{\lambda_1 \cdot \lambda_2} = c \cdot \Delta \lambda \frac{1}{\lambda_1 \cdot \lambda_2}
\]

If the source includes a third wavelength \( \lambda_3 \) midway between the extrema \( \lambda_1 \) and \( \lambda_2 \) (so that \( \lambda_3 = \lambda_{\text{avg}} \)), the factors \( \Delta \lambda \) and \( \lambda_{\text{avg}} \) are unchanged, but the irradiance pattern must be different in some way. The irradiance pattern generated by this three-line source is more difficult to calculate, but the result can 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):
11.2 FRINGE VISIBILITY – COHERENCE

(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 a third wavelength is added at $\lambda_{\text{avg}}$ with the same “power”. The distance between peaks of the fringe pattern has increased; (c) if all wavelengths between the maximum and minimum are used, the amplitudes of the neighboring maxima decrease.

The irradiance pattern generated from the superposition of these fringe patterns exhibits the short-period fringes due to $\lambda_{\text{avg}}$ and the long-period fringes due to $\Delta \lambda$.

If we add a third fringe pattern due to $\lambda_3 = \lambda_{\text{avg}}$, the resulting irradiance fringe pattern is shown in (b). Note that the region of visible fringes covers approximately the same extent of the $x$-axis, but the distance between such regions has increased. By extension, if all wavelengths are included in the source between $\lambda_1$ and $\lambda_2$, visible fringes will exist only in one region centered about $x = 0$.

Because the region of fringes created by the three-line source 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 first two are equally coherent, but less coherent than light emitted by the single-line source. The coherence may be quantified based on the temporal bandwidth. For a source whose range emitted wavelengths is:

$$\Delta \lambda = \lambda_{\text{max}} - \lambda_{\text{min}},$$

the corresponding temporal bandwidth is:

$$\Delta \nu = c \cdot \frac{\Delta \lambda}{\lambda_{\text{max}} \cdot \lambda_{\text{min}}},$$

Note that the dimensions of $\Delta \nu$ are (time)$^{-1}$. The time delay over which the phase difference of light emitted from one source point is predictable (and thus over which fringes may be generated) is the inverse of this bandwidth:

$$\Delta \tau = \frac{1}{\Delta \nu} = \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{c \cdot \Delta \lambda},$$

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 is the distance traveled
by an electromagnetic wave during the coherence time:
\[
c \cdot \Delta \tau \equiv \Delta \ell = \frac{c}{\Delta \nu} = \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{\Delta \lambda}
\]
and 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 fringes are visible is half the period \(D_{\text{mod}}\):

\[
D_{\text{mod}} \equiv \frac{1}{2 \cdot \sin \theta} \left| \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right|^{-1} = \frac{1}{2 \cdot \sin \theta} \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{\Delta \lambda}
\]

Thus the range of \(x\) over which fringes are visible is proportional to the coherence length; if the bandwidth is “wide”, then the coherence length is “short” and the fringes are visible over only a small region.

Lasers are the best available coherent sources; to a very good approximation, most lasers emit a single wavelength \(\lambda_0\) so that \(\Delta \lambda = 0\). The period of the modulating sinusoid \(D_{\text{mod}} = \infty\); fringes are visible at all \(x\). A coherent source should be employed when an optical interference pattern is 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.

### 11.2.2 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 vectors \(\mathbf{E}_1\) and \(\mathbf{E}_2\) to generate a field \(\mathbf{E}\). The resulting irradiance is:

\[
I = \mathbf{E} \cdot \mathbf{E} = |\mathbf{E}|^2 = |\mathbf{E}_1 + \mathbf{E}_2|^2
\]

\[
= (\mathbf{E}_1 + \mathbf{E}_2) \cdot (\mathbf{E}_1 + \mathbf{E}_2)
\]

\[
= (\mathbf{E}_1 \cdot \mathbf{E}_1) + (\mathbf{E}_2 \cdot \mathbf{E}_2) + (\mathbf{E}_1 \cdot \mathbf{E}_2) + (\mathbf{E}_2 \cdot \mathbf{E}_1)
\]

\[
= (\mathbf{E}_1 \cdot \mathbf{E}_1) + (\mathbf{E}_2 \cdot \mathbf{E}_2) + 2(\mathbf{E}_1 \cdot \mathbf{E}_2)
\]

\[
= I_1 + I_2 + 2(\mathbf{E}_1 \cdot \mathbf{E}_2)
\]

where \(I_1\) and \(I_2\) are the irradiances due to \(\mathbf{E}_1\) and \(\mathbf{E}_2\), respectively.

Consider the irradiance in the case where the incident fields are plane waves trav-
eling in directions \( \hat{e}_1 \) and \( \hat{e}_2 \), respectively:

\[
\begin{align*}
\mathbf{E}_1 &= \hat{e}_1 E_1 \cos [\mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t] \\
\mathbf{E}_2 &= \hat{e}_2 E_2 \cos [\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t] \\
I &= I_1 + I_2 + 2 \langle \hat{e}_1 E_1 \cos [\mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t] \cdot \hat{e}_2 E_2 \cos [\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t] \rangle \\
&= I_1 + I_2 + 2 E_1 E_2 (\hat{e}_1 \cdot \hat{e}_2) \langle \cos [\mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t] \cos [\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t] \rangle \\
&= I_1 + I_2 + 2 E_1 E_2 (\hat{e}_1 \cdot \hat{e}_2) \langle \cos [(\mathbf{k}_1 - \mathbf{k}_2) \cdot \mathbf{r} - (\omega_1 - \omega_2) t] \rangle \\
&= I_1 + I_2 + 2 E_1 E_2 (\hat{e}_1 \cdot \hat{e}_2) \langle \cos [2 \mathbf{k}_{\text{mod}} \cdot \mathbf{r} - 2 \omega_{\text{mod}} t] \rangle
\end{align*}
\]

In the case where the two components are polarized orthogonally so that \( \hat{e}_1 \cdot \hat{e}_2 = 0 \), then the irradiance is the sum of the component irradiances and no interference is seen.

In the case where \( \omega_1 = \omega_2 \) so that \( \omega_{\text{mod}} = 0 \), and \( I_1 = I_2 \), then the output irradiance is:

\[
I = 2 I_1 (1 + \cos [2 \mathbf{k}_{\text{mod}} \cdot \mathbf{r}])
\]

which again says that the irradiance includes a stationary sinusoidal fringe pattern with spatial period \( \lambda_{\text{mod}} = \frac{2\pi}{|\mathbf{k}_{\text{mod}}|} \).

### 11.3 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, \textit{etc.}), or of the system (path length, traveling distance, index of refraction, \textit{etc.}). Interferometers are generally divided into two classes that specify the method of separating a single wavefront into two (or more) wavefronts that may be recombined. The classes are \textit{division of wavefront} and \textit{division of amplitude}. The former type has been considered; it divides wavefronts emitted by the source into two pieces and redirects one or both of them down different paths. They are recombined in a fashion such that \( \mathbf{k}_1 \neq \mathbf{k}_2 \), even though \( |\mathbf{k}_1| = |\mathbf{k}_2| \). The interference pattern is generated from the \( \mathbf{k}_{\text{mod}} \) portion of the sum of the wavefronts. Division-of-amplitude interferometers use a partially reflecting mirror – the \textit{beamsplitter} – to divide the wavefront into two beams which 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.

#### 11.3.1 Michelson Interferometer

Division-of-Amplitude interferometers are distinguished from the just-considered division-of-wavefront interferometers by the presence of a \textit{beamsplitter}, which divides the incident radiation into two parts by partial reflection/partial transmission. The two
wavefronts are directed down different paths before recombining to create interference. For example, consider the Michelson interferometer shown below.

![Optical schematic of the Michelson interferometer where the incident wavefront is planar.](image)

Note that the beamsplitter reflects part of the wave and transmits the rest. From the definition of the amplitude reflection coefficient:

$$ r = \frac{n_1 - n_2}{n_1 + n_2} \text{ at normal incidence} $$

we see 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.

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:

$$ E_1 = \sqrt{I_1} = \sqrt{\frac{I_0}{2}} = \sqrt{\frac{E_0^2}{2}} $$

$$ = E_0 \cdot \frac{1}{\sqrt{2}} = 0.707 \cdots E_0 $$

and the amplitude of the transmitted beam $E_2 = E_1$.

Because each beam is reflected once and transmitted once before being recombined, the amplitude of each component when recombined is:

$$ (E_1)_{out} = (E_2)_{out} = E_0 \cdot \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2}} = \frac{1}{2} E_0 $$

Each beam experiences a phase delay proportional to the optical distance traveled in its arm of the interferometer:

$$ \Phi_1 = \frac{2\pi}{\lambda_1} \cdot d_1 = \frac{2\pi}{\lambda_1} \cdot n_1 d_1 = k \cdot n_1 d_1 $$
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 \( n_2 d_2 = 2n_2 L_2 \). 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 \( \frac{1}{\lambda} \); i.e., longer wavelengths (red light) experience smaller phase delays than shorter wavelengths (blue light).

The amplitude at the detector is the sum of the amplitudes:

\[
E(t) = \frac{E_0}{2} \left( \cos \left[ \frac{2\pi}{\lambda_1} \cdot n_1 \cdot 2L_1 - \omega_1 t \right] + \cos \left[ \frac{2\pi}{\lambda_2} \cdot n_2 \cdot 2L_2 - \omega_1 t \right] \right)
\]

which has the form \( \cos [A] + \cos [B] \), and thus may be rewritten as:

\[
E(t) = \frac{2E_0}{2} \cos \left[ \frac{2\pi}{\lambda_1} \left( \frac{n_1 L_1 + n_2 L_2}{2} \right) - 2\pi \nu_1 t \right] \cdot \cos \left[ \frac{2\pi}{\lambda_2} \left( \frac{n_1 L_1 - n_2 L_2}{2} \right) \right]
\]

If the indices of refraction in the two paths are equal (usually \( n_1 = n_2 \equiv 1 \)), then the expression is simplified:

\[
E(t) = E_0 \cos \left[ \frac{2\pi n (L_1 + L_2)}{\lambda_1} - 2\pi \nu_1 t \right] \cdot \cos \left[ \frac{2\pi n (L_1 - L_2)}{\lambda_1} \right]
\]

\[
= E_0 \cos \left[ \frac{2\pi (L_1 + L_2)}{\lambda_1} - 2\pi \nu_1 t \right] \cdot \cos \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right] \text{ for } n = 1
\]
One of the multiplicative terms is a rapidly oscillating function of time; the other term is stationary in time. The time average of the squared magnitude is the irradiance:

\[
I = \langle |E(t)|^2 \rangle = E_0^2 \left\langle \cos^2 \left[ \frac{2\pi (L_1 + L_2)}{\lambda_1} - 2\pi \nu_1 t \right] \cdot \cos^2 \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right] \right\rangle \\
= E_0^2 \left\langle \cos^2 \left[ \frac{2\pi (L_1 + L_2)}{\lambda_1} - 2\pi \nu_1 t \right] \right\rangle \cdot \cos^2 \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right] \\
= E_0^2 \frac{1}{2} \cos^2 \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right]
\]

The identity \( \cos^2 [\theta] = \frac{1}{2} (1 + \cos [2\theta]) \) may be used to recast the expression:

\[
I = E_0^2 \frac{1}{2} \cdot \frac{1}{2} \left( 1 + \cos \left[ \frac{2\pi}{\lambda_1} \cdot 2 (L_1 - L_2) \right] \right) \\
= \frac{E_0^2}{4} \left( 1 + \cos \left[ \frac{2\pi}{\lambda_1} \cdot 2 (L_1 - L_2) \right] \right)
\]

Note that this is not a function of time or position, but only of the lengths of the arms of the interferometer and of \( \lambda \). 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 path difference, which is directly related to the shape of the wavefronts. 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). 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 $2 \cdot (L_1 - L_2) \cdot \cos [\theta]$.

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:

$$OPD = 2 \cdot (L_1 - L_2) \cdot \cos [\theta]$$

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:

$$\Delta \Phi = k \cdot OPD = \frac{2\pi}{\lambda_1} \cdot 2 \left[ L_1 - L_2 \right] \cdot \cos [\theta] = \frac{4\pi \cdot (L_1 - L_2) \cdot \cos [\theta]}{\lambda_1} \text{ radians}$$

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:

$$\Delta \Phi = 2\pi m = \frac{4\pi \cdot (L_1 - L_2) \cdot \cos [\theta]}{\lambda_1} \rightarrow mL_1 = 2 \left( L_1 - L_2 \right) \cdot \cos [\theta]$$

The corresponding angles $\theta$ are specified by:

$$\theta = \cos^{-1} \left[ \frac{mL_1}{2 \cdot (L_1 - L_2)} \right]$$

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.
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 the Michelson Interferometer**

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:

\[
OPD = (n - 1) t = \Delta m \cdot \lambda_1
\]

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)

### 11.3.2 Mach-Zehnder Interferometer

The M-Z interferometer is very similar to the Michelson, except that a second beam-splitter 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. 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:

\[
(n_1 - 1) \cdot \delta \cdot \frac{2\pi}{\lambda_1} = m \cdot 2\pi \rightarrow n_1 = 1 + \frac{m \lambda_1}{\delta}
\]
**11.3 Divison-of-Amplitude Interferometers**

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.

11.3.3 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:

$$
\cos[\omega_1 t] + \cos[\omega_2 t] = 2 \cos[\omega_{avg} t] \cdot \cos[\omega_{mod} t]
$$

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.
11.3.4 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 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 detected. 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. 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 $r$:

$$R^2 = (R - d)^2 + r^2 \implies r^2 = 2Rd - d^2 \approx 2Rd \text{ for } d \ll R$$

$$d \approx \frac{r^2}{2R} \text{ for } d \ll R$$

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:

$$m \cdot \frac{\lambda}{2} = OPD = nd \rightarrow d = m \cdot \frac{\lambda}{2} \text{ in air}$$
Chapter 12

Optical Diffraction and Imaging

In geometrical (ray) optics, light is assumed to propagate in straight lines from the source (rectilinear propagation). However, Grimaldi observed in the 1600s that this model does not conform to reality after light interacts with an obstruction. Grimaldi observed that light deviates from straight-line propagation into the shadow region. He named this phenomenon *diffraction*. This spreading of a bundle of rays affects the sharpness of shadows cast by opaque objects; the edges become fuzzy because light propagates into the geometrical shadow region.

Diffraction really is the same phenomenon as interference. In both, 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; the effect is considered to be interference. If the apertures are large (compared to the wavelength $\lambda$), then the spherical-wave contributions from a 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. The mathematical model for diffraction is straightforward to develop, though computations may be tedious.

Recall the form of 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
CHAPTER 12 OPTICAL DIFFRACTION AND IMAGING

origin is:

$$s \left[ x_1, y_1, z_1, t \right] = \frac{E_0}{\sqrt{x_1^2 + y_1^2 + z_1^2}} \cos \left[ k_x \cdot x + k_y \cdot y + k_z \cdot z - \omega_0 t \right]$$

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:

$$s \left[ \mathbf{r}, t \right] = \frac{E_0}{|\mathbf{r}|} \cos \left[ \mathbf{k}_0 \cdot \mathbf{r} - \omega_0 t \right] \rightarrow \frac{E_0}{|\mathbf{r}|} \exp \left[ \mathbf{k}_0 \cdot \mathbf{r} - \omega_0 t \right]$$

where $$|\mathbf{r}| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}$$ and $$|\mathbf{k}_0| = \frac{2\pi}{\lambda_0}$$.

For large $$|\mathbf{r}|$$, 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.

### 12.1 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:

$$E \left[ x_1, y_1; z_1, x_0, y_0, 0 \right] = \frac{E_0}{|\mathbf{r}|} \cos \left[ \mathbf{k}_0 \cdot \mathbf{r} - \omega_0 t \right], \text{ where } |\mathbf{r}| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}$$

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
12.1 DIFFRACTION INTEGRALS

$E_0$, then the integral is simplified somewhat:

$$E_{\text{total}}[x_1, y_1; z_1] = \alpha^2 \int_{-\infty}^{+\infty} E[x_1, y_1; z_0, y_0, 0] \, dx_0 \, dy_0$$

$$= \int_{\text{aperture}} E_0 \left( \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} \right) \cos \left( \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} - \omega_0 t \right) \, dx_0 \, dy_0$$

where $\alpha$ is a constant that is evaluated from physical considerations. It can be shown that $\alpha = (\sqrt{\lambda_0})^{-1} \Rightarrow \alpha^2 = \frac{1}{i\lambda_0}$, where the factor of $i^{-1} = \exp[\frac{-i\pi}{2}]$ is a constant phase. 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:

$$E_{\text{total}}[x_1, y_1; z_1] = \frac{E_0}{i\lambda_0} \int_{-\infty}^{+\infty} f[x_0, y_0] \left( \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} \right) \cos \left( \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} - \omega_0 t \right) \, dx_0 \, dy_0$$

This expression is the *diffraction integral*. Again, this expression LOOKS 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:

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 *Fre’nel*).

2. Observation plane $z_1$ located a large distance the source plane at $z_0 = 0$ so that $z_1 \equiv \infty$. 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 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.

### 12.2 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, $\omega_0 \approx 10^{15}$ radians/second) and with distance (because $|\mathbf{k}|$ is very small, $|\mathbf{k}| \approx 10^{-7}$ 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 $|\mathbf{r}|$ 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
Now consider the approximation of the distance $|r|$. The complete expression is:

$$|r| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}$$

$$= z_1 \cdot \sqrt{1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}}$$

This is an EXACT expression, but it may be expanded into a power series by applying the binomial theorem. The general binomial expansion is:

$$(1 + \alpha)^n = 1 + n\alpha + \frac{n \cdot (n-1)}{2!} \alpha^2 + \frac{n \cdot (n-1) \cdot (n-2)}{3!} \alpha^3 + \ldots + \frac{n!}{(n-r)!r!} \alpha^r + \ldots$$

This series converges to the correct value if $\alpha^2 < 1$. For the case $n = \frac{1}{2}$ (square root), the result is:

$$(1 + \alpha)^{\frac{1}{2}} = 1 + \frac{\alpha}{2} - \frac{1}{8} \alpha^2 + \frac{1}{16} \alpha^3 - \ldots$$

which leads to an expression for the distance $|r|:

$$|r| = z_1 \left(1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right)^{\frac{1}{2}}$$

If $z_1$ is sufficiently large, terms of second and larger order may be assumed to be sufficiently close to zero that they may be ignored. In other words, we assume that:

$$(x_1 - x_0)^2 + (y_1 - y_0)^2 \ll z_1^2$$

The resulting approximation for the propagation distance is:

$$|r| \approx z_1 \left(1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}\right) = z_1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{2z_1}$$

This may be simplified further by recasting the electric field expression into complex
CHAPTER 12 OPTICAL DIFFRACTION AND IMAGING

notation:

\[ E[x_1, y_1; z_1, x_0, y_0, 0] \approx \frac{E_0}{i\lambda_0 z_1} \Re \left\{ \exp \left[ + \frac{2\pi i}{\lambda_0} \left( z_1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{2z_1} \right) - 2\pi i\nu_0 t \right] \right\} \]

\[ = \frac{E_0}{i\lambda_0 z_1} \Re \left\{ \exp \left[ + \frac{2\pi i}{\lambda_0} z_1 \right] \cdot \exp \left[ -2\pi i\nu_0 t \right] \cdot \exp \left[ + \frac{i\pi}{\lambda_0 z_1} (x_1 - x_0)^2 + (y_1 - y_0)^2 \right] \right\} \]

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 from the observation point. 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 the unreasonable part of the assumption of Fresnel diffraction; 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:

\[ E_{\text{total}}[x_1, y_1; z_1] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} E[x_1, y_1; z_1; x_0, y_0, 0] \, dx_0 \, dy_0 \]

\[ \approx \frac{1}{i\lambda_0 z_1} \exp \left[ \left( \frac{z_1}{\lambda_0} - i\nu_0 t \right) \right] \cdot \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ + \frac{i\pi}{\lambda_0 z_1} (x_1 - x_0)^2 + (y_1 - y_0)^2 \right] \, dx_0 \, dy_0 \]

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.

12.2.1 Fresnel Diffraction as a Convolution Integral

Consider the Fresnel diffraction integral:

\[ F[x_1, y_1] = \frac{1}{i\lambda_0 z_1} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ + \frac{i\pi}{\lambda_0 z_1} (x_1 - x_0)^2 + (y_1 - y_0)^2 \right] \, dx_0 \, dy_0 \]
Define the exponential to be a function $h$ that depends on the four variables in a particular way:

$$h [x_1 - x_0, y_1 - y_0] \equiv \frac{1}{i \lambda_0 z_1} \exp \left[ + \frac{i \pi}{\lambda_0 z_1} \left( (x_1 - x_0)^2 + (y_1 - y_0)^2 \right) \right]$$

In other words, the Fresnel diffraction integral may be written as:

$$F [x_1, y_1] = \int_{-\infty}^{+\infty} f [x_0, y_0] \ h [x_1 - x_0, y_1 - y_0] \ dx_0 \ dy_0$$

Integral equations of this form abound in all areas of physical science, and particularly in imaging; they are called convolution integrals. The function $h$ is the shape of the integral function and is often called the impulse response of the integral operator. In imaging, and particularly in optics, the impulse response often is called the point-spread function. In other areas of physics, it has other names (e.g., Green’s function). 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 $\alpha, \beta$:

$$F [x, y] = \int_{-\infty}^{+\infty} f [\alpha, \beta] \ h [x - \alpha, y - \beta] \ d\alpha \ d\beta$$

$$\equiv f [x, y] * h [x, y]$$

where the form of the impulse response for Fresnel diffraction is:

$$h [x, y] = \frac{1}{i \lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i \frac{\pi}{\lambda_0 z_1} \left( x^2 + y^2 \right) \right]$$

This impulse response is called a “chirp” function – the real and imaginary parts are both sinusoids with varying spatial frequency and that also differs with distance $z_1$. The parameters of the chirp often are combined into $\sqrt{\lambda_0 z_1} \equiv \alpha_0$

$$h [x, y] = \frac{1}{i \lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i \pi \frac{x^2 + y^2}{\alpha_0^2} \right]$$

so that the phase of the chirp function is $\pi$ where $\sqrt{x^2 + y^2} = \alpha_0$. 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 $\alpha_0 = 1$ and $\alpha_0 = 2$. The source distance is larger by a factor of four in the second case.
1-D profiles of the impulse response of Fresnel diffraction for (a) $\sqrt{z_1} = 1$ and (b) $\sqrt{z_1} = 2$, 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**

Below are computed simulations of the profiles of diffraction patterns 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.
12.2 FRESNEL DIFFRACTION

in Hecht.

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) $\sqrt{z_1} = 1$; (b) $\sqrt{z_1} = 2 \implies z$ 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 $\frac{1}{4}$.

Since convolution is linear and shift invariant, the “images” of rectangular apertures may be calculated at these two distances by replicating the impulse responses, reversing one, and adding the amplitudes before computing the irradiance.

1-D profiles of the irradiance (squared magnitude) of diffraction patterns from rectangle functions for different distances from the origin by replicating the impulse responses, reversing one, and adding the amplitudes before computing the irradiance:

(a) $\sqrt{z_1} = 1$; (b) $\sqrt{z_1} = 2 \implies z$ is four times larger.

12.2.2 Characteristics of Fresnel Diffraction

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 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.

### 12.3 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:

\[
\exp \left[ \frac{i\pi}{\lambda_0 z_1} \left( (x_1 - x_0)^2 + (y_1 - y_0)^2 \right) \right]
= \exp \left[ \frac{i\pi (x_0^2 + y_0^2)}{\lambda_0 z_1} \right] \cdot \exp \left[ \frac{i\pi (x_1^2 + y_1^2)}{\lambda_0 z_1} \right] \cdot \exp \left[ -\frac{i2\pi}{\lambda_0 z_1} (x_0 x_1 + y_0 y_1) \right]
\]

If the source is restricted to be near to the optical axis so that \(x_0, y_0 \approx 0\) (or, more rigorously, if \(x_0^2 + y_0^2 << \lambda_0 z_1\)), then

\[
\frac{(x_0^2 + y_0^2)}{\lambda_0 z_1} \approx 0
\]

so that:

\[
\exp \left[ \frac{i\pi (x_0^2 + y_0^2)}{\lambda_0 z_1} \right] \approx 1.
\]
This resulting diffraction integral simplifies to:

\[
F[x_1, y_1] = \frac{1}{i\lambda_0 z_1} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ -\frac{i\pi}{\lambda_0 z_1} \left( (x_1 - x_0)^2 + (y_1 - y_0)^2 \right) \right] dx_0 \, dy_0
\]

\[
\approx \frac{1}{i\lambda_0 z_1} \exp \left[ \frac{i\pi (x_1^2 + y_1^2)}{\lambda_0 z_1} \right] \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ -\frac{i2\pi}{\lambda_0 z_1} (x_0 x_1 + y_0 y_1) \right] dx_0 \, dy_0
\]

Similarly, if the observation point is near to the optic axis so that \( x_1^2 + y_1^2 < \ll \lambda_0 z_1 \), then:

\[
\exp \left[ \frac{i\pi (x_1^2 + y_1^2)}{\lambda_0 z_1} \right] \approx 1.
\]

Though \( x_0 \) and \( x_1 \) are sufficiently small for these approximations, the third exponential term is retained because their difference may be larger:

\[
\exp \left[ \frac{i\pi}{\lambda_0 z_1} \left( (x_1 - x_0)^2 + (y_1 - y_0)^2 \right) \right] \approx \exp \left[ -\frac{2\pi i (x_0 x_1 + y_0 y_1)}{\lambda_0 z_1} \right]
\]

Considered in the observation plane as functions of \( x_1 \) and \( y_1 \), the phase of the wavefront is proportional to the source variables \([x_0, y_0]\); the wavefront is a plane. The corresponding approximation for the diffraction integral is:

\[
E_{\text{total}}[x_1, y_1; z_1] = \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ -\frac{2\pi i (x_0 x_1 + y_0 y_1)}{\lambda_0 z_1} \right] dx_0 \, dy_0
\]

The diffracted light far from the source is a summation of the plane waves generated by each source point. This is called the Fraunhofer diffraction formula, and the resulting patterns are VERY different from Fresnel diffraction from the same aperture. In fact, the formula can be interpreted as a Fourier transform where the frequency coordinates are mapped back to the space domain via \( \xi = \frac{x_1}{\lambda_0 z_1}, \eta = \frac{y_1}{\lambda_0 z_1} \):

\[
E_{\text{total}}[x_1, y_1; z_1] = \left( \frac{1}{i\lambda_0 z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ -2\pi i (x_0 x_1 + y_0 y_1) \right] dx_0 \, dy_0
\]

\[
= \left( \frac{1}{z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cdot \mathcal{F}_2 \{ f[x, y] \}_{\xi = \frac{x_1}{\lambda_0 z_1}, \eta = \frac{y_1}{\lambda_0 z_1}}
\]

The resulting irradiance patterns are the time-averaged squared magnitudes of the fields.

\[
|E_{\text{total}}[x_1, y_1; z_1]|^2 \propto \left( \frac{1}{\lambda_0 z_1^2} \right) \cdot \left| \mathcal{F}_2 \{ f[x, y] \}_{\xi = \frac{x_1}{\lambda_0 z_1}, \eta = \frac{y_1}{\lambda_0 z_1}} \right|^2
\]

\[
= \frac{1}{\lambda_0^2 z_1^2} \cdot \left| F \left[ \frac{x_1}{\lambda_0 z_1}, \frac{y_1}{\lambda_0 z_1} \right] \right|^2
\]
All of the Fourier transform relationships apply to Fraunhofer diffraction patterns. The patterns (the “images”) scale in inverse proportion to the original functions; larger input functions \( f[x, y] \) produce smaller (and brighter) diffraction patterns. Movement of an object off axis produces a linear phase is the pattern that is not visible in the irradiance, but has an effect if two such patterns are added.

12.3.1 Computed Examples of Fraunhofer Diffraction

Below are shown the profiles of square apertures and computed simulations of the resulting amplitude diffraction pattern in the Fraunhofer diffraction region (the irradiance is the squared magnitude of the plotted amplitude). In both cases, the object is a point source located at \( z_1 \approx \infty \) so that the light “fills” both apertures with zero phase. For a fixed (large) distance from the object, the “images” of the diffracted light get “narrower” and “taller” as the aperture width increases.

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.

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.
12.4 Fraunhofer Diffraction and Imaging

A monochromatic point object located a long distance away from an imaging system produces a set of wavefronts that are “regularly” spaced (separated by the wavelength) and are approximately planar. The entrance pupil of the optical system (the image of the aperture stop in object space) collects a section of the plane wavefront and the optical elements convert it to a spherical wave 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 the on-axis source continues as before 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 points 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 stop: the monochromatic point object is located a large distance to the left of the stop and the diffracted light forms a Fraunhofer irradiance pattern on the observation screen a large distance from the stop; if the object consists of two point sources, the diffracted amplitude is the sum of the individual amplitudes separated by the same angle \( \theta_0 \) by which the sources are apparently separated.

Of course, real optical imaging systems consist of more than just a stop; they also include lenses and/or mirrors that change the curvature of the piece of the wavefront that passes through the stop/pupil. The curvature of section of the (approximately) spherical wave converges to a point in the image plane. Because the distance to the object is large, the image plane is approximately co-located with the focal plane, i.e., the distance to the image is approximately equal to the focal length \( f \) of the lens. We can interpret the action of the lens as “bringing infinity close;” it 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. In other words, 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 \( f_{\text{eff}} \) 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 calculation of the squared magnitude eliminates this factor. After accounting for the phase factor, the irradiance image of a single point source in monochromatic light is proportional to the squared magnitude of the Fourier transform of the entrance pupil:

\[
|E_{\text{total}}(x_1, y_1; f)|^2 = h \left[ x_1, y_1; f \right] \propto \left| F \left[ \frac{x_1}{\lambda_0 \cdot f}, \frac{y_1}{\lambda_0 \cdot f} \right] \right|^2
\]

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.

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: \( \frac{1}{z_2} \)
An optical system imaging a point source creates a scaled replica the Fraunhofer diffraction pattern of the stop on the image plane. In other words, the impulse response of the imaging system is the Fraunhofer diffraction pattern of the aperture stop.

Lord Rayleigh derived an equation for the angular resolution of an imaging system for sources with wavelength $\lambda_0$ and separated by a distance $d$. These are located a distance $L$ from the system whose entrance pupil diameter is $D$. The Rayleigh criterion for resolution is that:

$$D \propto \frac{L\lambda_0}{d}$$

Note the similarity to the formula for the separation between interference fringes in Young’s experiment. The angular radius of the image spot is a good measure of the angular resolution:

$$\Delta \theta \approx \frac{d}{L} \approx \frac{\lambda_0}{D}$$

(note the similarity between this equation and that for Young’s two-slit experiment). For systems with imaging elements that have circular cross-sections, the angular radius of the image spot is somewhat larger:

$$\Delta \theta \approx 1.22 \cdot \frac{\lambda_0}{D}.$$
so the diameter of the image spot is:

\[ 2 \cdot \Delta \theta \approx 2.44 \cdot \frac{\lambda_0}{D}. \]

The diameter of the primary mirror of the Hale telescope on Palomar Mountain is \( D = 5.08 \text{ m} \), so the theoretical minimum angular separation is \( \Delta \theta \approx 0.24 \cdot \lambda_0 \), for \( \lambda_0 \) measured in meters. In green light (\( \lambda_0 = 550 \text{ nm} \)), the angular separation is:

\[ \Delta \theta \approx 1.32 \cdot 10^{-7} \text{ radians} = 0.132 \text{ mrad} \approx 0.03 \text{ arc-seconds} \]

The diameter of the primary mirror of the Hubble space telescope is approximately \( 1/2 \) that of the Hale telescope (\( D \approx 2.4 \text{ m} \)), so the angular resolution of the optics at 550 nm is approximately twice as large (\( 0.26 \mu \text{rad} \approx 0.6 \text{ arc-seconds} \)). Of course, there is no atmosphere to mess up the Hubble images.

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:

\[ f \begin{bmatrix} x, y \end{bmatrix} = CYL \left( \frac{r}{D} \right) \]

Its Fourier transform is:

\[ F \begin{bmatrix} \xi, \eta \end{bmatrix} = \pi \left( \frac{D}{2} \right)^2 \text{SOMB} (D \rho) \]

\[ = \frac{D^2}{4} \cdot \frac{2J_1(\pi D \rho)}{\pi D \rho} = \frac{D}{2\rho} J_1 (\pi D \rho) \]

We make the substitution:

\[ \rho = \sqrt{\xi^2 + \eta^2} \rightarrow \frac{\sqrt{x_1^2 + y_1^2}}{\lambda_0 \cdot f_{eff}} = \frac{r}{\lambda_0 \cdot f_{eff}} \]

\[ \implies D \rho \rightarrow \frac{Dr}{\lambda_0 \cdot f_{eff}} \]

We know that the first zero of \( J_1 (\pi u) \) occurs at \( u_0 \approx 1.22 \), which occurs in this system at:

\[ \frac{Dr_0}{\lambda_0 \cdot f_{eff}} \approx 1.22 \]

\[ \implies r_0 \approx 1.22 \cdot \lambda_0 \left( \frac{f_{eff}}{D} \right) \]

\[ r_0 \approx 1.22 \cdot \lambda_0 \cdot f/\# \]

This is the dimension of the diffraction spot on the image plane (focal plane) and is an easily remembered equation for the spatial resolution of an optical system. That the resolution is directly proportional to the wavelength means that a system with a
fixed focal length and entrance pupil diameter can resolve smaller separations if the wavelength is reduced. Also, systems with smaller focal ratios have better resolution ("faster systems resolve better").

In real life, it is very difficult (meaning very expensive) to fabricate optical imaging systems with large entrance pupils whose performance is constrained by this diffraction limit. The performance of the system is much more often constrained by the limitations of the optical design and fabrication that results in optical aberrations (e.g., spherical aberration, astigmatism, distortion). Generally speaking, the aberrations become more difficult and expensive to correct for larger entrance pupils.

Of course, the ultimate resolution of the Hale telescope actually is limited by atmospheric turbulence, which creates random variations in the air temperature and thus in the refractive index. 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 ("tip-tilt") move the image from side to side and up-down. Quadratic phase errors ("defocus") act like additional lenses that move the image plane backwards or forwards along the optical axis. In general, the 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.