

Classical and Modern Optics

Daniel A. Steck

Oregon Center for Optics and Department of Physics, University of Oregon

Classical and Modern Optics

Daniel A. Steck

Oregon Center for Optics and Department of Physics, University of Oregon

Copyright © 2006, by Daniel Adam Steck. All rights reserved.

This material may be distributed only subject to the terms and conditions set forth in the Open Publication License, v1.0 or later (the latest version is presently available at <http://www.opencontent.org/openpub/>). Distribution of substantively modified versions of this document is prohibited without the explicit permission of the copyright holder. Distribution of the work or derivative of the work in any standard (paper) book form is prohibited unless prior permission is obtained from the copyright holder.

Original revision posted 16 June 2006.

This is revision 1.7.4, 6 July 2017.

Cite this document as:

Daniel A. Steck, *Classical and Modern Optics*, available online at <http://steck.us/teaching> (revision 1.7.4, 6 July 2017).

Author contact information:

Daniel Steck
Department of Physics
1274 University of Oregon
Eugene, Oregon 97403-1274
dsteck@uoregon.edu

ISBN 000-00-00000-00-0

Acknowledgements

A number of people have improved this work by graciously providing corrections, comments, and suggestions.

Special thanks to:

- Jonathan Mackrory (U. Oregon) has pointed out a number of corrections and has co-written a number of the exercises.
- Kirk Madison (U. British Columbia) has provided numerous corrections, both minor and major.

Also, thanks to:

- Eryn Cook (U. Oregon)
- Nadav Katz (Hebrew U. Jerusalem)
- Lennart Karssen (Utrecht U.)
- Dovid Levin (Bar Ilan U.)
- Cristian Mejía (Universidad del Atlántico)
- Jordan Neuhoff (U. Oregon)
- John Noe (Stony Brook U.)
- Maximilian Vergin (Technische Universität Hamburg)

Contents

Contents	13
1 Review of Linear Algebra	15
1.1 Definitions	15
1.2 Linear Transformations	15
1.3 Matrix Arithmetic	16
1.4 Eigenvalues and Eigenvectors	17
1.5 Exercises	18
2 Ray Optics	19
2.1 Introduction	19
2.2 Ray Optics and Fermat's Principle	19
2.3 Fermat's Principle: Examples	20
2.4 Paraxial Rays	23
2.5 Matrix Optics	24
2.6 Composite Systems	28
2.6.1 Example: Thin Lens	28
2.7 Resonator Stability	29
2.7.1 Stability Condition	30
2.7.2 Periodic Motion	31
2.7.3 Resonator Stability: Standard Form	32
2.8 Nonparaxial Ray Tracing with Interfaces	34
2.8.1 Refraction and Reflection Laws: Coordinate-Free Form	35
2.8.2 Ray Tracing: A Recipe	37
2.8.3 Example: Parabolic Interface	38
2.8.3.1 Refraction	38
2.8.3.2 Reflection	41
2.9 Exercises	43
3 Fourier Analysis	55
3.1 Periodic Functions: Fourier Series	55
3.1.1 Example: Rectified Sine Wave	57
3.2 Aperiodic Functions: Fourier Transform	58
3.2.1 Example: Fourier Transform of a Gaussian Pulse	59
3.3 The Fourier Transform in Optics	60
3.4 Delta Function	61
3.5 Exercises	64
4 Review of Electromagnetic Theory	67
4.1 Maxwell Equations in Vacuum	67
4.2 Intensity	68

4.3	Maxwell Equations in Media	69
4.4	Simple Dielectric Media	69
4.5	Monochromatic Waves and Complex Notation	71
4.5.1	Alternate Complex Notation	72
4.6	Intensity in Complex Notation	72
4.6.1	Complex Notation for Simple Dielectric Media	72
4.7	Plane Waves	73
4.8	Vector Plane Waves	75
4.8.1	Wave Impedance	76
4.9	Exercises	77
5	Interference	81
5.1	Superposition of Two Plane Waves	81
5.2	Mach-Zehnder Interferometer	82
5.3	Stokes Relations	83
5.4	Mach-Zehnder Interferometer: Applications	84
5.5	Michelson Interferometer	85
5.6	Sagnac Interferometer	86
5.7	Interference of Two Tilted Plane Waves	86
5.8	Multiple-Wave Interference	87
5.9	Exercises	89
6	Gaussian Beams	91
6.1	Paraxial Wave Equation	91
6.2	Gaussian Beams	92
6.2.1	Amplitude Factor	93
6.2.2	Longitudinal Phase Factor	95
6.2.3	Radial Phase Factor	95
6.3	Specification of Gaussian Beams	96
6.4	Vector Gaussian Beams	97
6.5	ABCD Law	97
6.5.1	Free-Space Propagation	98
6.5.2	Thin Optic	99
6.5.3	Cascaded Optics	99
6.5.4	Factorization of a General Matrix	99
6.5.5	“Deeper Meaning” of the ABCD Law	100
6.5.6	Example: Focusing of a Gaussian Beam by a Thin Lens	100
6.5.7	Example: Minimum Spot Size by Lens Focusing	101
6.6	Hermite–Gaussian Beams	102
6.6.1	Doughnut and Laguerre–Gaussian Modes	104
6.7	Exercises	106
7	Fabry–Perot Cavities	113
7.1	Resonance Condition	113
7.2	Broadening of the Resonances: Cavity Damping	114
7.2.1	Standard Form	115
7.2.2	Maximum and Minimum Intensity	116
7.2.3	Width of the Resonances	116
7.2.4	Survival Probability	117
7.2.5	Photon Lifetime	117
7.2.6	Q Factor	118
7.2.7	Example: Finesse and Q	119
7.3	Cavity Transmission	120

7.3.1	Reflected Intensity	121
7.3.2	Intracavity Buildup	121
7.4	Optical Spectrum Analyzer	122
7.5	Spherical-Mirror Cavities: Gaussian Modes	124
7.5.1	Physical Modes	125
7.5.2	Symmetric Cavities	125
7.5.3	Special Cavities	126
7.5.4	Resonance Frequencies	126
7.5.5	Algebraic Digression	127
7.6	Spherical-Mirror Cavities: Hermite–Gaussian Modes	127
7.6.1	Confocal Cavity	128
7.7	Exercises	129
8	Polarization	135
8.1	Vector Plane Waves	135
8.2	Polarization Ellipse	135
8.2.1	Simple Cases	136
8.3	Polarization States: Jones Vectors	138
8.3.1	Vector Properties	139
8.4	Polarization Devices: Jones Matrices	140
8.4.1	Linear Polarizer	140
8.4.2	Wave Retarder	140
8.4.3	Polarization Rotator	142
8.4.4	Cascaded Systems	142
8.5	Coordinate Transformations	142
8.6	Normal Modes	144
8.7	Polarization Materials	144
8.7.1	Birefringence	144
8.7.1.1	Multiple Refraction	145
8.7.2	Optical Activity	147
8.8	Exercises	149
9	Fresnel Relations	151
9.1	Optical Waves at a Dielectric Interface	151
9.1.1	Phase Changes and the Brewster Angle	154
9.2	Reflectance and Transmittance	155
9.3	Internal Reflection	156
9.3.1	Phase Shifts	158
9.4	Air-Glass Interface: Sample Numbers	159
9.5	Reflection at a Dielectric-Conductor Interface	161
9.5.1	Propagation in a Conducting Medium	161
9.5.2	Inductive Heating	163
9.5.3	Fresnel Relations	165
9.6	Exercises	167
10	Thin Films	171
10.1	Reflection-Summation Model	172
10.1.1	Example: Single Glass Plate as a Fabry–Perot Etalon	173
10.2	Thin Films: Matrix Formalism	174
10.3	Optical Coating Design	178
10.3.1	Single-Layer Antireflection Coating	178
10.3.2	Two-Layer Antireflection Coating	179
10.3.3	High Reflector: Quarter-Wave Stack	181

10.4 Gradient-Index Layers	182
10.4.1 Finite Stack of Very Thin Films	183
10.4.2 Continuous Medium	184
10.4.3 Reflection Coefficient	185
10.4.4 Solution of Riccati Equations	186
10.4.5 Example: Single, Homogeneous Film	186
10.4.6 Example: Linear Permittivity Gradient	187
10.4.7 Asymptotic Forms	191
10.4.7.1 Sharp-Boundary Limit	191
10.4.7.2 Small-Reflection Limit	191
10.4.8 P-Polarization	194
10.4.8.1 Example: Linear Permittivity Gradient	195
10.5 Exercises	197
11 Fourier Analysis II: Convolution	199
11.1 Motion Blurring in Photography	199
11.2 Convolution	200
11.2.1 Example: Convolution of Box Functions	201
11.2.2 Example: Photographic Blurring	201
11.3 Convolution Theorem	202
11.3.1 Spatial Fourier Transforms	202
11.3.2 Proof	203
11.3.3 Example: Convolution of Two Gaussians	203
11.4 Application: Error Analysis	204
11.5 Application: Central Limit Theorem	205
11.5.1 Central Limit Theorem Application: Random Walk	207
11.5.2 Central Limit Theorem Application: Standard Deviation of the Mean	207
11.6 Application: Impulse Response and Green Functions	208
11.6.1 Frequency Domain	209
11.7 Application: Spectral Transmission	210
11.8 Exercises	212
12 Fourier Optics	217
12.1 Fourier Transforms in Multiple Dimensions	217
12.2 Wave Propagation in Homogeneous Media	217
12.2.1 Fingerprints of Propagation	217
12.2.2 Decomposition	218
12.2.3 Reverse Waves	219
12.2.4 Fourier-Transform Recipe	219
12.2.5 Paraxial Propagation	220
12.2.5.1 Solution of the Paraxial Wave Equation	220
12.2.6 Nonparaxial Propagation and the Diffraction Limit	221
12.3 Fraunhofer Diffraction	222
12.3.1 Far-field Propagation	222
12.3.2 Thin Lens as a Fourier Transform Computer	224
12.3.3 Example: Diffraction from a Double Slit	225
12.3.4 Example: Diffraction from a Sinusoidal Intensity-Mask Grating	226
12.3.5 Example: Diffraction from an Arbitrary Grating	226
12.4 Fresnel Diffraction	227
12.4.1 Convolution Revisited	227
12.4.2 Paraxial Impulse Response	228
12.4.3 Far-Field (Fraunhofer) Limit	228
12.4.4 Example: Fresnel Diffraction from a Slit	229

12.5 Spatial Filters	231
12.5.1 Spatial Filtering of a Gaussian Beam	233
12.5.2 Visualization of Phase Objects	236
12.5.2.1 Zernike Phase-Contrast Imaging	236
12.5.2.2 Central Dark-Ground Method	237
12.5.2.3 Schlieren Method	237
12.5.2.4 Numerical Examples	240
12.6 Holography	242
12.6.1 Example: Single-Frequency Hologram	243
12.6.2 Film Holograms	243
12.6.3 Hologram of a Plane Wave and Off-Axis Holography	244
12.6.4 Setup: Off-Axis Reflection Hologram	245
12.7 Exercises	246
13 Acousto-Optic Diffraction	253
13.1 Raman–Nath Regime	254
13.1.1 Diffraction Amplitudes: Bessel Functions	256
13.1.2 Frequency Shifts	257
13.1.3 Momentum Conservation	257
13.2 Bragg Regime	259
13.2.1 Efficiency	264
13.2.2 Example: TeO_2 Modulator (Bragg Regime)	265
13.3 Borderline	265
13.4 Exercises	267
14 Coherence	271
14.1 Wiener–Khinchin Theorem	272
14.2 Optical Wiener–Khinchin Theorem	273
14.2.1 Application: FTIR Spectroscopy and the Michelson Interferometer	274
14.2.2 Example: Monochromatic Light	275
14.2.3 Normalized One- and Two-Sided Spectra	275
14.3 Visibility	277
14.4 Coherence Time, Coherence Length, and Uncertainty Measures	278
14.5 Interference Between Two Partially Coherent Sources	280
14.6 Exercises	281
15 Laser Physics	283
15.1 Overview	283
15.1.1 Laser Pumps	283
15.1.2 Gain Media	284
15.1.2.1 Gas-Phase Atoms	284
15.1.2.2 Atoms Embedded in Transparent Solids	284
15.1.2.3 Molecules	285
15.1.2.4 Semiconductor Lasers	285
15.1.3 Optical Resonator	286
15.1.4 A Simple Model of Laser Oscillation: Threshold Behavior	286
15.1.5 A Less-Simple Model of Laser Oscillation: Steady-State Oscillation	287
15.2 Light–Atom Interactions	288
15.2.1 Quantization	288
15.2.2 Fundamental Light–Atom Interactions	289
15.2.3 Einstein Rate Equations	290
15.2.4 Relations Between the Einstein Coefficients	290
15.2.5 Line Shape and Spectral Distributions	291

15.2.5.1 Broadband Light	292
15.2.5.2 Nearly Monochromatic Light	292
15.3 Light Amplification	293
15.3.1 Gain Coefficient	293
15.3.1.1 Stimulated Emission	293
15.3.1.2 Absorption	294
15.3.1.3 Spontaneous Emission	294
15.3.1.4 Combined Effects	294
15.3.2 Threshold Behavior and Single-Mode Operation	295
15.4 Pumping Schemes	296
15.4.1 Three-Level Laser	297
15.4.2 Four-Level Laser	298
15.5 Gain Coefficient	299
15.5.1 Gain in a Medium of Finite Length	301
15.6 Laser Output: CW	302
15.6.1 Optimum Output	303
15.6.2 Quantum Efficiency	304
15.7 Laser Output: Pulsed	304
15.7.1 Laser Spiking	304
15.7.2 Q-Switching	306
15.7.3 Cavity Dumper	307
15.7.4 Mode Locking	307
15.8 Exercises	309
16 Dispersion and Wave Propagation	317
16.1 Causality and the Kramers–Kronig Relations	317
16.1.0.1 DC Component	319
16.1.1 Refractive Index	319
16.1.1.1 Example: Lorentzian Absorption	320
16.2 Pulse Propagation and Group Velocity	321
16.2.1 Phase Velocity	321
16.2.2 Group Velocity	321
16.2.3 Pulse Spreading	323
16.3 Slow and Fast Light	325
16.3.1 Quantum Coherence: Slow Light	326
16.3.2 Fast Light	327
16.4 Exercises	329
17 Classical Light–Atom Interactions	331
17.1 Polarizability	331
17.1.1 Connection to Dielectric Media	332
17.1.2 Conducting Media: Plasma Model	333
17.2 Damping: Lorentz Model	333
17.2.1 Oscillator Strength	335
17.2.2 Conductor with Damping: Drude Model	336
17.3 Atom Optics: Mechanical Effects of Light on Atoms	336
17.3.1 Dipole Force	337
17.3.1.1 Dipole Potential: Standard Form	338
17.3.2 Radiation Pressure	339
17.3.2.1 Dipole Radiation	340
17.3.2.2 Damping Coefficient	342
17.3.2.3 Photon Scattering Rate	343
17.3.2.4 Scattering Force	344

17.3.3 Laser Cooling: Optical Molasses	344
17.3.3.1 Doppler Cooling Limit	346
17.3.3.2 Magneto-Optical Trap	348
17.4 Exercises	349
Index	351

Chapter 1

Review of Linear Algebra

1.1 Definitions

Before tackling optics formalism directly, we will spend a bit of time reviewing some of the mathematical concepts that we will need, specifically the basics of linear algebra.

Obviously to get started we will need to define what we mean by a matrix, the fundamental object in linear algebra. But in order to do so in a rigorous way, we need to define some more fundamental mathematical concepts.

- A **set** is a collection of objects, or **elements**. This is a naive definition that can have problems, but it will suffice for our purposes. We write $a \in A$ to say that the element a belongs to the set A .
- The **cartesian product** $A \times B$ of two sets A and B is the set of all ordered pairs $\langle a, b \rangle$ such that $a \in A$ and $b \in B$.
- A **function** $f : A \rightarrow B$ (“ f maps A into B ”) is a subset of $A \times B$ such that if $\langle a, b \rangle \in f$ and $\langle a, c \rangle \in f$ then $b = c$. That is, f maps each element in A to exactly one element in B (different elements of A can be mapped to the same element in B).
- A bit of notation, which is slightly unconventional for set theory (but sufficient for our purposes): We will use the integer n as a shorthand for the set $\{1, \dots, n\}$.

Now we can define a **matrix**, specifically an $m \times n$ matrix, as a function mapping $m \times n \rightarrow \mathbb{R}$. Of course, this is for a real matrix; a complex-valued matrix is a function mapping $m \times n \rightarrow \mathbb{C}$. We will denote a matrix by a boldface letter, such as \mathbf{A} , and we will denote **matrix elements** (results of the function) by A_{ij} , where $i \in m$ and $j \in n$. Again, we are writing real numbers, functions of the two integer indices. Of course, we can also write out all the elements of a matrix. For example, a 2×2 matrix would look like

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad (1.1)$$

where note that by convention the first index refers to the row, while the second refers to the column.

1.2 Linear Transformations

Why are matrices so important? They define linear transformations between vector spaces.

A **vector space** V is a set of elements (vectors) with *closed* operations $+$ and \cdot (**closed operations** are functions that map vectors in V back to vectors in V) that satisfy the following axioms:

$$1. \ x + (y + z) = (x + y) + z \quad \forall_{x,y,z \in V}$$

2. $\exists_{e \in V} (x + e = x)$
3. $\forall_{x \in V} \exists_{x^{-1} \in V} (x + x^{-1} = e)$
4. $x + y = y + x \quad \forall_{x,y \in V}$
5. $a \cdot (x + y) = (a \cdot x) + (a \cdot y) \quad \forall_{a \in \mathbb{R}, x,y \in V}$
6. $(a + b) \cdot x = (a \cdot x) + (b \cdot x) \quad \forall_{a,b \in \mathbb{R}, x \in V}$
7. $(ab) \cdot x = a \cdot (b \cdot x) \quad \forall_{a,b \in \mathbb{R}, x \in V}$
8. $1 \cdot x = x \quad \forall_{x \in V}$

(Quick exercise: is the set $\{0\}$ a vector space? What about the set \mathbb{Q} of all rational numbers?) All of these properties make intuitive sense for the “usual” finite-dimensional vectors in \mathbb{R}^n , but the point is that more general objects such as functions and derivative operators can live in vector spaces as well.

A **linear transformation** L of a vector space V into another vector space W is a function $L : V \rightarrow W$ such that:

1. $L(x + y) = L(x) + L(y) \quad \forall_{x,y \in V}$
2. $L(ax) = aL(x) \quad \forall_{a \in \mathbb{R}, x \in V}$

The point is that *any* linear transformation between finite-dimensional vector spaces can be represented by a matrix (see Problem 1.2).

Linear transformations are of great importance in physics because they are *tractable*. Many simple problems that you study in physics are linear, and they can be readily solved. Even in the much more difficult nonlinear cases, linear approximations are much easier to understand and give insight into the more complicated case.

We will most commonly think of finite-dimensional **vectors** as $(n \times 1)$ -dimensional matrices (single-column matrices).

1.3 Matrix Arithmetic

We will now define a few fundamental mathematical operations with matrices.

- A matrix \mathbf{C} is the **sum** of matrices \mathbf{A} and \mathbf{B} ($\mathbf{C} = \mathbf{A} + \mathbf{B}$) if $C_{ij} = A_{ij} + B_{ij}$ for all i and j . Clearly, for this to work, \mathbf{A} and \mathbf{B} must have the same dimension.
- A matrix \mathbf{C} is the **product** of matrices \mathbf{A} and \mathbf{B} ($\mathbf{C} = \mathbf{AB}$) if

$$C_{ij} = \sum_k A_{ik}B_{kj} \tag{1.2}$$

for all i and j . Clearly, for this to work, if \mathbf{A} is $m \times n$ then \mathbf{B} must have dimension $n \times p$ for some integer p , and the product \mathbf{C} will have dimension $m \times p$.

- The **identity matrix** \mathbf{I}_n is an $n \times n$ matrix defined by $(\mathbf{I}_n)_{ij} = \delta_{ij}$, where δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise).
- The **inverse matrix** of an $n \times n$ matrix \mathbf{A} is denoted \mathbf{A}^{-1} , and satisfies $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$. The inverse matrix is not guaranteed to exist, and a matrix for which the inverse exists is said to be **invertible** or **nonsingular**.
- The **transpose matrix** of a matrix \mathbf{A} is given by $(\mathbf{A}^\top)_{ij} = A_{ji}$. Clearly if \mathbf{A} is $m \times n$ then \mathbf{A}^\top is $n \times m$.

Finally, we will define the **determinant** of a matrix, something which maps a square, real-valued matrix to a real number. Let \mathbf{A} be an $n \times n$ matrix. Then the determinant is

$$\det(\mathbf{A}) := \sum_{j_1 j_2 \dots j_n} \epsilon_{j_1 j_2 \dots j_n} A_{1j_1} A_{2j_2} \dots A_{nj_n}, \quad (1.3)$$

where the **permutation symbol** $\epsilon_{j_1 j_2 \dots j_n}$ is $+1$ if (j_1, j_2, \dots, j_n) is an even permutation of $(1, 2, \dots, n)$ (i.e., can be obtained from $(1, 2, \dots, n)$ by exchanging adjacent pairs of numbers an even number of times), it is -1 if (j_1, j_2, \dots, j_n) is an odd permutation of $(1, 2, \dots, n)$, and is 0 otherwise (i.e., if any number is repeated). While this serves as a formal definition of the determinant, it is unwieldy except for small matrices. However, we will be most concerned with 2×2 matrices in optics, where the determinant is given by

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = AD - BC. \quad (1.4)$$

The determinant has the important property that it is nonzero if and only if the matrix is nonsingular.

1.4 Eigenvalues and Eigenvectors

Let \mathbf{A} be an $n \times n$ matrix. We want to consider cases where

$$\mathbf{A} \cdot \mathbf{x} = \lambda \mathbf{x}, \quad (1.5)$$

where \mathbf{x} is an n -dimensional vector and $\lambda \in \mathbb{R}$. If there are λ and \mathbf{x} that satisfy this relation then λ is said to be an **eigenvalue** of \mathbf{A} with corresponding **eigenvector** \mathbf{x} .

Why should we consider eigenproblems? Eigenvalues typically represent physically important values of physical quantities, and eigenvectors typically represent physically important elements of vector spaces (typically providing a physically significant basis for a vector space). As a simple example, consider a system of coupled oscillators (mechanical or otherwise). Each uncoupled oscillator would satisfy an equation of the form $\ddot{x} = -\omega^2 x$. When coupled together, the system of oscillators would more generally satisfy a matrix equation of the form

$$\ddot{\mathbf{x}} = \mathbf{A} \cdot \mathbf{x}. \quad (1.6)$$

Upon making the *ansatz* $\mathbf{x}(t) = \mathbf{x}(0) e^{-i\omega t}$, we obtain an eigenvalue equation,

$$\mathbf{A} \cdot \mathbf{x} = -\omega^2 \mathbf{x}. \quad (1.7)$$

Thus, the eigenvalues of \mathbf{A} represent the distinct frequencies of oscillation for the coupled system, while the eigenvectors represent how different oscillators move together to make each distinct “mode” of oscillation corresponding to each frequency.

How do we find the eigenvalues and eigenvectors? First, note that the eigenvalue condition above implies that $(\mathbf{A} - \lambda \mathbf{I}_n) \cdot \mathbf{x} = 0$, so that $\mathbf{A} - \lambda \mathbf{I}_n$ is a singular matrix. Thus, we have the condition that

$$\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0, \quad (1.8)$$

which yields the **characteristic polynomial** in λ . The eigenvalues are the roots of the characteristic polynomial. For a 2×2 matrix \mathbf{A} , the characteristic polynomial is simple (it is handy to know this):

$$\lambda^2 - \text{Tr}(\mathbf{A})\lambda + \det(\mathbf{A}) = 0. \quad (1.9)$$

Here, $\text{Tr}(\mathbf{A})$ is the **trace** of the matrix, defined as the sum over the diagonal elements. The eigenvector corresponding to an eigenvalue λ can then be found by solving the homogenous linear system $(\mathbf{A} - \lambda \mathbf{I}_n) \cdot \mathbf{x} = 0$.

If the eigenvectors are linearly independent (i.e., it is not possible to write any one of them as a linear combination of the others), then they form a “nice basis” for the vector space in the sense that in this basis, the linear transformation represented by \mathbf{A} is now represented by a diagonal matrix. Mathematically, let \mathbf{P} be a matrix such that the columns are eigenvectors of \mathbf{A} . Then $\mathbf{P}^{-1} \mathbf{A} \mathbf{P}$ is a diagonal matrix. In fact the diagonal elements are the eigenvalues. If the eigenvectors are furthermore mutually orthogonal (as is the case for real, symmetric matrices with distinct eigenvalues) and normalized, then we have $\mathbf{P}^\top = \mathbf{P}^{-1}$ (i.e., \mathbf{P} is an **orthogonal matrix**), and thus the diagonal matrix is $\mathbf{P}^\top \mathbf{A} \mathbf{P}$.

1.5 Exercises

Problem 1.1

- (a) The **trace** of a square matrix \mathbf{A} is defined by

$$\text{Tr}[\mathbf{A}] := \sum_j A_{jj}. \quad (1.10)$$

Let \mathbf{A} and \mathbf{B} be $n \times n$ matrices. Prove that the trace of the product is order-invariant,

$$\text{Tr}[\mathbf{AB}] = \text{Tr}[\mathbf{BA}]. \quad (1.11)$$

- (b) Prove that the trace of the product of $n \times n$ matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(N)}$ is invariant under cyclic permutation of the product.

Problem 1.2

Let \mathbb{M} and \mathbb{N} be vector spaces of dimension M and N , respectively, and let $\phi : \mathbb{M} \rightarrow \mathbb{N}$ be a linear transformation. Show that ϕ can be represented by an $N \times M$ matrix; that is, show that there is a matrix \mathbf{A} such that $\mathbf{A} \cdot \mathbf{x} = \phi(\mathbf{x})$ for every vector $\mathbf{x} \in \mathbb{M}$.

Problem 1.3

Let \mathbf{A} be an $n \times n$ matrix. Show that $\det(\mathbf{A}) = (-1)^n \det(-\mathbf{A})$.

Problem 1.4

Show that if \mathbf{A} is a 2×2 matrix, the characteristic polynomial is given by

$$\lambda^2 - \text{Tr}(\mathbf{A})\lambda + \det(\mathbf{A}) = 0, \quad (1.12)$$

and thus that the eigenvalues are given by

$$\lambda_{1,2} = \frac{\text{Tr}(\mathbf{A})}{2} \pm \sqrt{\left(\frac{\text{Tr}(\mathbf{A})}{2}\right)^2 - \det(\mathbf{A})}. \quad (1.13)$$

Further, show explicitly that if $\det(\mathbf{A}) = 1$, then $\lambda_2 = 1/\lambda_1$.

Problem 1.5

Let \mathbf{A} be a diagonalizable $n \times n$ matrix (i.e., there exists a matrix \mathbf{P} such that $\mathbf{A} = \mathbf{PDP}^{-1}$, where \mathbf{D} is a diagonal matrix with the eigenvalues of \mathbf{A} along the diagonal).

Show that

$$\log[\det(\mathbf{A})] = \text{Tr}[\log(\mathbf{A})]. \quad (1.14)$$

You may use the property $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ of the determinant for $n \times n$ matrices \mathbf{A} and \mathbf{B} .

Problem 1.6

- (a) Is the trace of a matrix (viewing the trace operation as a function) a linear transformation? If yes, prove it; if no, give a counterexample.
(b) Repeat (a), but for the determinant of a matrix.

Chapter 2

Ray Optics

2.1 Introduction

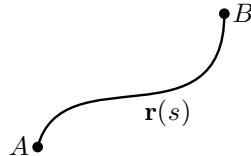
Ray optics, or **geometrical optics**, is the simplest theory of optics. The “mathematical” definition is simply that the ray is a path. The more intuitive “definition” is that a ray represents the center of a thin, slowly diverging beam of light. While this theory gets a lot of things right, it also misses many phenomena. But even when we get to wave optics, geometrical optics will still provide a lot of basic intuition that we need to understand the more complex behavior of optical waves. For now, we will ignore any wave effects, and any width or spreading of the ray (though by examining many rays, it is possible to model the imaging behavior of an optical system).

2.2 Ray Optics and Fermat’s Principle

Ray optics boils down to a single statement, which we will get to very shortly. But first, we will start by noting that the fundamental assumption in ray optics is that light travels in the form of rays. Again, as mentioned above, a single ray could represent a beam of light, but typically you would use many rays to model light propagation (e.g., to model the performance of an imaging system).

The optical rays propagate in optical media. To keep things simple, we will assume that the media are lossless, and thus we can characterize them completely by their **index of refraction**, which we will denote by n . Usually, $n \geq 1$, with $n = 1$ corresponding to vacuum. And while many media are uniform, meaning that the refractive index is uniform throughout the medium, many media also have refractive indices $n(\mathbf{r})$ that vary spatially. The only effect that we require of the refractive index is that it changes the speed of light. The speed of light in a medium of refractive index n is simply c_0/n , where c_0 is the vacuum speed of light (defined to be exactly $2.997\ 924\ 58 \times 10^8$ m/s).

Now to the fundamental principle of ray optics, called **Fermat’s Principle**. Consider a path $\mathbf{r}(s)$ inside an optical medium between points A and B , parameterized by the variable s , so that A corresponds to $\mathbf{r}(0)$, and B corresponds to $\mathbf{r}(d)$, where both s and d have the dimensions of length.



Then the **optical path length** for this path is the length of the path, but weighted by the local refractive index. Mathematically, we can define the **optical path length functional** as

$$\ell[\mathbf{r}] := \int_0^d n(\mathbf{r}) ds. \quad (2.1)$$

(optical path length)

This quantity is proportional to the time light takes to traverse the path, $\Delta t = \ell/c_0$, and is just the ordinary length of the path in the case $n = 1$. Then Fermat's Principle states that optical rays traverse paths that satisfy

$$\delta\ell = 0. \quad (2.2)$$

(Fermat's Principle)

The “ δ ” here is like a derivative, but for functions, and “ $\delta\ell$ ” here can be read as “the variation of ℓ ” or the “the first variation of ℓ .” What the statement $\delta\ell = 0$ means is that nearby paths have the same path length. Formally, this means that

$$\ell[\mathbf{r} + \epsilon\delta\mathbf{r}] - \ell[\mathbf{r}] = 0, \quad (2.3)$$

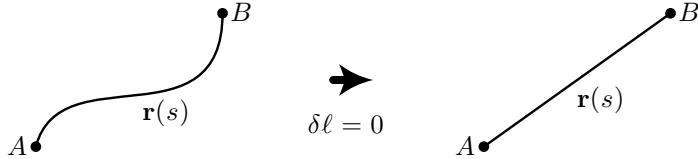
at least to first order in ϵ , where $\delta\mathbf{r}$ is an arbitrary path satisfying $\delta\mathbf{r}(0) = \delta\mathbf{r}(d) = 0$. The proper mathematical framework for all this is the **calculus of variations**, but we won't really go much more into this here. However, it is for this reason that Fermat's Principle is called a **variational principle**. It turns out that variational principles are extremely important in many areas of physics. Variational principles are conceptually odd, however. Normally you think of light starting someplace, and then asking, where does it go? When applying Fermat's Principle, the idea is to stipulate the *endpoints*, or where the light begins and ends up, and then ask, what does it do in the meantime?

Often, as happens in standard calculus, the condition $\delta\ell = 0$ yields a minimum for ℓ . Hence Fermat's Principle is often referred to as a “principle of least time.” However, this isn't necessarily the case. Often, the stationary condition turns out to be an inflection point or more commonly a saddle point, where ℓ is a minimum along one direction but a maximum along another. Obviously $\delta\ell = 0$ can never yield a true global maximum. Given a stationary path, there is always a nearby path that is slightly longer.

2.3 Fermat's Principle: Examples

We will now consider some applications of Fermat's Principle.

1. **Homogeneous Medium.** In a homogeneous medium, the refractive index n is constant. Thus, the minimum *optical* path length occurs for the path with the minimum length, which is a straight line.



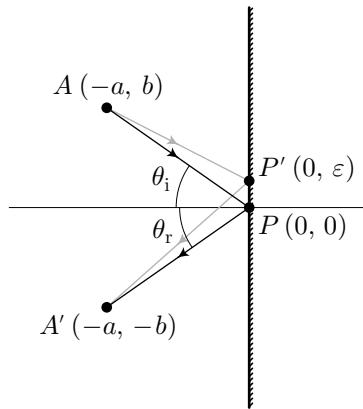
Then the optical path length is simply

$$\ell = n \int_A^B ds = nd, \quad (2.4)$$

where d is the regular length of the path. This quantity is thus minimized for a straight line connecting A and B . We know experimentally that light travels in straight lines, and this is one of the motivations for Fermat's Principle.

More formally, recall that we must have $\delta\ell = 0$, which is satisfied for a path that minimizes ℓ . But to see that the variation vanishes for the straight line, consider that any small perturbation $\pm\epsilon\delta\mathbf{r}$ to the path $\mathbf{r}(s)$ will change the path length by the same amount for either sign, just due to the symmetry of the straight line. Thus, $\delta\ell = 0$, at least to lowest order in ϵ . Note that for any path that *isn't* straight, we couldn't come to the same conclusion: for almost any perturbation to the path, one of the perturbations $\pm\epsilon\delta\mathbf{r}$ will tend to increase the path length, while the other will decrease it.

2. **Plane Mirror.** Again, we will start by fixing the endpoints of the ray, in the diagram below we label these by A and A' .



We will assume a homogeneous medium to the left of the mirror with refractive index n , and we have already shown the ray to be straight while in this medium. We can guess from symmetry that APA' is the minimum-length path (of the paths that bounce off the mirror). But let's prove it. Consider a nearby point P' . The length of $AP'A'$ is

$$\frac{\ell}{n} = \sqrt{a^2 + (b - \varepsilon)^2} + \sqrt{a^2 + (b + \varepsilon)^2}. \quad (2.5)$$

Differentiating with respect to the perturbation ε ,

$$\frac{\partial}{\partial \varepsilon} \left(\frac{\ell}{n} \right) = \frac{\varepsilon - b}{\sqrt{a^2 + (b - \varepsilon)^2}} + \frac{\varepsilon + b}{\sqrt{a^2 + (b + \varepsilon)^2}}. \quad (2.6)$$

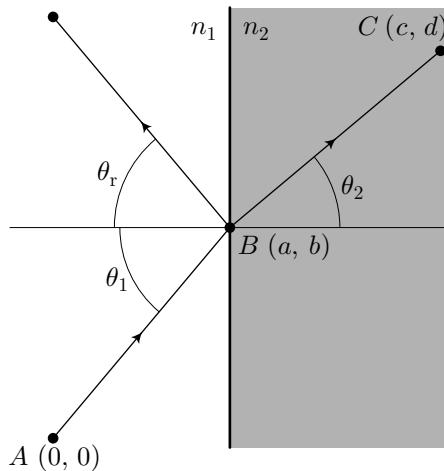
This quantity vanishes for $\varepsilon = 0$, and so the point P represents the extremal path. Thus, we conclude that

$$\theta_i = \theta_r, \quad (2.7)$$

(Law of Reflection)

thus arriving at the Reflection Law for optical rays.

3. Refractive Interface.



At a planar interface between two optical media of different refractive index, the ray splits. We have to assume this as an experimental fact at this point, since we really need full electromagnetism to show this from first principles. The reflected ray behaves according to the law of reflection that we just derived. The other, **refracted** ray is a little different. Assume that the refracted ray begins at point A and ends at point C (i.e., we take both A and C to be fixed). We will assume it crosses the

refractive interface at point B , and now we will compute where the point B must be according to Fermat's Principle, $\delta(ABC) = 0$. The path length is

$$\ell = n_1 \sqrt{a^2 + b^2} + n_2 \sqrt{(c-a)^2 + (d-b)^2}. \quad (2.8)$$

Differentiating with respect to the moveable coordinate b of B , we arrive at the extremal condition

$$\frac{\partial \ell}{\partial b} = \frac{n_1 b}{\sqrt{a^2 + b^2}} - \frac{n_2(d-b)}{\sqrt{(c-a)^2 + (d-b)^2}} = 0. \quad (2.9)$$

Using the angles marked in the diagram, we can rewrite this condition as

$$n_1 \sin \theta_1 - n_2 \sin \theta_2 = 0, \quad (2.10)$$

or

$$n_1 \sin \theta_1 = n_2 \sin \theta_2, \quad (2.11)$$

(Snell's Law)

which is simply Snell's Law. Note that

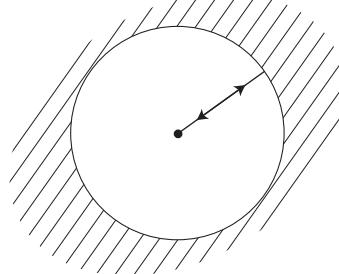
$$\sin \theta_2 = \frac{n_1}{n_2} \sin \theta_1 \leq 1. \quad (2.12)$$

If $n_1 > n_2$ then there is a critical angle θ_c given by

$$\frac{n_1}{n_2} \sin \theta_c = 1, \quad (2.13)$$

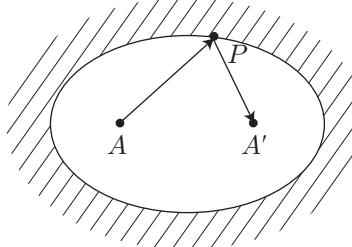
such that if $\theta_1 > \theta_c$, then there is no possible transmitted ray. All the light is instead reflected, and this phenomenon is called **total internal reflection**.

4. Spherical Mirror.



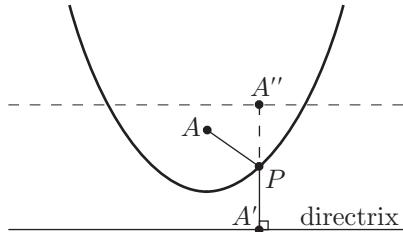
All rays from the center to the outer edge and back have the same (minimum) optical path length ℓ (i.e., the minimum ℓ for reflected rays). Thus, a spherical mirror focuses rays from an object at the center point back onto itself.

5. Elliptical Mirror.



We can define an ellipse as the set of all points $\{P : APA' = d\}$ for some constant distance d . The points A and A' are the **foci** of the ellipse. Fermat's Principle tells us immediately that since for any point P on the ellipse, we know that APA' is constant (and a minimal length for reflecting paths), and thus rays starting at the point A will end at the point A' . Thus an elliptical mirror images an object at A to the other focus A' .

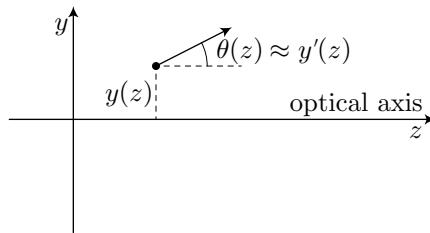
6. Parabolic Mirror.



We can define a parabola as the set of all points $\{P : AP = PA' \text{ where } PA' \perp \text{directrix}\}$. Adding PA'' to $AP = PA'$, we find $AP + PA'' = A'A''$, which is constant. So APA'' is constant for all rays. Thus, a parabolic mirror collimates all rays starting at A, which is the **focus** of the parabola. Reversing the sense of propagation, we can also conclude that all incoming parallel rays orthogonal to the directrix are concentrated to the focus A.

2.4 Paraxial Rays

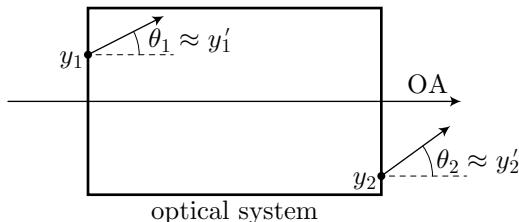
Now we will set up a formalism for keeping track of optical rays more precisely. We will represent a ray by a vector, which keeps track of the position and direction of the ray with respect to the **optical axis**—the reference axis for optical propagation. The displacement y from the optical axis and the direction θ are sufficient to *completely* specify the ray at a particular longitudinal position z .



In the paraxial approximation that we will get to below, we will see that the angle θ is approximately equivalent to the slope y' of the ray, so we will write vectors interchangeably with the angle and slope of the ray:

$$\begin{bmatrix} y \\ \theta \end{bmatrix} \approx \begin{bmatrix} y \\ y' \end{bmatrix}. \quad (2.14)$$

We want to be able to compute the change in the ray vector for any optical system. In this sense, we will model an optical system as a *transformation* of ray vectors.



In the most general case, we can write

$$y_2 = f_1(y_1, y'_1); \quad y'_2 = f_2(y_1, y'_1), \quad (2.15)$$

or in vector form,

$$\begin{bmatrix} y_2 \\ y'_2 \end{bmatrix} = \mathbf{f} \begin{bmatrix} y_1 \\ y'_1 \end{bmatrix}, \quad (2.16)$$

where the vector function \mathbf{f} models the optical system.

Assume that y_α and y'_α are small. Then we can Taylor-expand the function to lowest order in y_1 and y'_1 :

$$\begin{aligned} y_2 &= f_1(y_1, y'_1) = \left. \frac{\partial f_1}{\partial y_1} \right|_{y_1=y'_1=0} y_1 + \left. \frac{\partial f_1}{\partial y'_1} \right|_{y_1=y'_1=0} y'_1 + \text{higher-order terms in } y_1, y'_1 \\ y'_2 &= f_2(y_1, y'_1) = \left. \frac{\partial f_2}{\partial y_1} \right|_{y_1=y'_1=0} y_1 + \left. \frac{\partial f_2}{\partial y'_1} \right|_{y_1=y'_1=0} y'_1 + \text{higher-order terms in } y_1, y'_1. \end{aligned} \quad (2.17)$$

Note that we are assuming $f_1(0,0) = f_2(0,0) = 0$, as we have not written down the zeroth-order terms in the expansion. This assumes that any ray following the optical axis on the input side will continue to do so on the output side, and therefore optics like spherical lenses, interfaces, and so on are *centered* on the optical axis. It also excludes certain optics that deflect all rays, such as tilted mirrors and prisms. However, by defining the optical axis carefully, we can still accommodate these optics (see the example of the plane mirror on p. 26; tilted mirrors and prisms can be “unwrapped” in the same way). Now we can rewrite the above expansion in matrix form:

$$\begin{bmatrix} y_2 \\ y'_2 \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y'_1} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y'_1} \end{bmatrix}_{y_1=y'_1=0} \begin{bmatrix} y_1 \\ y'_1 \end{bmatrix} + \text{higher-order terms in } y_1, y'_1 \quad (2.18)$$

In the **paraxial approximation**, we will ignore the quadratic terms and model the optical system using only linear transformations. This approximation is valid for small y_α and y'_α (or equivalently, θ_α , so that $\theta_\alpha \approx \sin \theta_\alpha \approx \tan \theta_\alpha = y'_\alpha$, which justifies our interchangeable use of θ and y'). The matrix of derivatives is the **transfer matrix** that represents the optical system in the paraxial approximation. Note that for this approximation to be valid, the ray must always stay close to the optical axis. While the choice of the optical axis is in principle arbitrary, it is best chosen such that the paraxial approximation is good for optical rays of interest.

The higher order corrections that we are neglecting are treated in **aberration theory**, which we will not treat here.

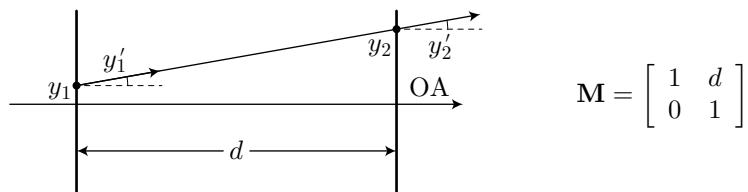
2.5 Matrix Optics

Recall from before that the most general *linear* transformation of a two-dimensional vector is a 2×2 matrix. We have written a matrix above as an expansion of a more general transformation, but for the general paraxial case we will use the notation

$$\begin{bmatrix} y_2 \\ y'_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ y'_1 \end{bmatrix}. \quad (\text{notation for } ABCD\text{-matrix propagation of rays}) \quad (2.19)$$

The matrix representing the optical system is referred to as an “*ABCD* matrix,” “ray matrix,” or “ray-transfer matrix.” We will now derive the fundamental matrices.

1. Free-Space Propagation.



$$\mathbf{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

One of the simplest cases is propagation in free space over a distance d . The ray travels in a straight line, so the angle does not change

$$y'_2 = y'_1. \quad (2.20)$$

By comparison to the matrix equation $y'_2 = Cy_1 + Dy'_1$, we can conclude that $C = 0$ and $D = 1$. Since the slope is y'_1 , the position changes according to

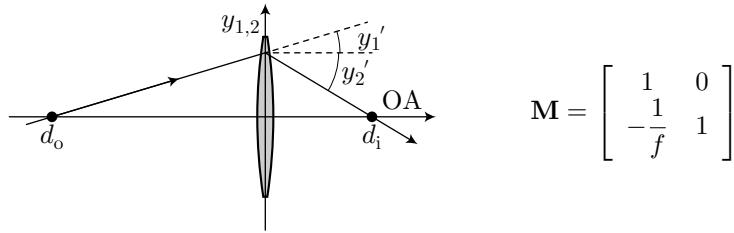
$$y_2 = y_1 + y'_1 d. \quad (2.21)$$

By comparison to $y_2 = Ay_1 + By'_1$, we can conclude that $A = 1$ and $B = d$. Thus, the free-space matrix is simply

$$\mathbf{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}. \quad (2.22) \quad (\text{free-space propagation matrix})$$

Note that this is matrix *still* valid for propagation over a distance d within a refractive medium, *independent* of n . The rays still change height according to their angle; the “compression” effect on the optical length happens because the angles themselves change when the rays *enter* and *leave* the medium, say, for a planar window of glass.

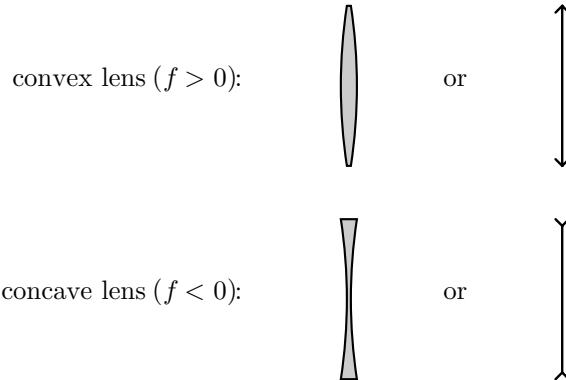
2. Thin Lens.



Because the lens is thin, the ray does not propagate over any distance Δz . The ray is continuous, so $y_2 = y_1$. Thus, $A = 1$ and $B = 0$. The ray is deflected, however, such that it satisfies the **thin lens law**:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}. \quad (2.23)$$

Here d_o is the **object distance**, d_i is the **image distance**, and f is the **focal length**, the single parameter that completely characterizes a thin lens. The sign convention for the focal length is that $f > 0$ for a convex lens, and $f < 0$ for a concave lens, as shown here (assuming a larger index for the lens than the surroundings, as for glass lenses in air):



The line drawings shown to the right in the above figure are common schematic representations of convex and concave lenses in diagrams. To arrive at the rest of the ray-matrix elements, we can

take the object and image distances to be where the ray crosses the axis before and after the lens, respectively. Thus, we can write the initial slope as

$$y'_1 = \frac{y_1}{d_o}, \quad (2.24)$$

and similarly we can write

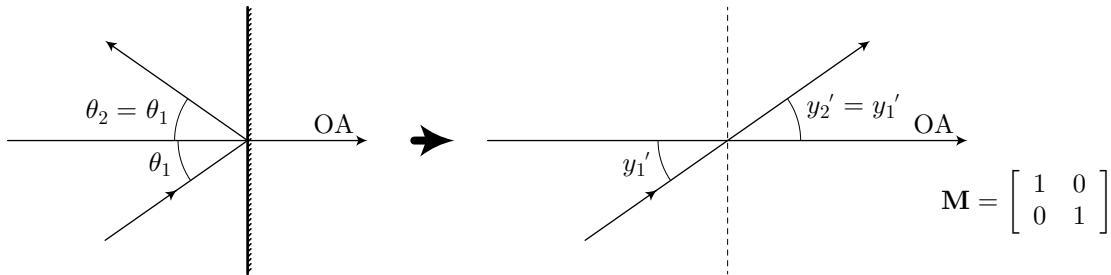
$$y'_2 = -\frac{y_2}{d_i} = -\frac{y_1}{d_i} = -y_1 \left(\frac{1}{f} - \frac{1}{d_o} \right) = -\frac{y_1}{f} + y'_1, \quad (2.25)$$

where we used $y_2 = y_1$ and the thin lens law to eliminate y_2 . Thus $C = -1/f$ and $D = 1$, and we can write the ray matrix for a thin lens as

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix}. \quad (2.26)$$

(thin-lens $ABCD$ matrix)

3. Plane Mirror.



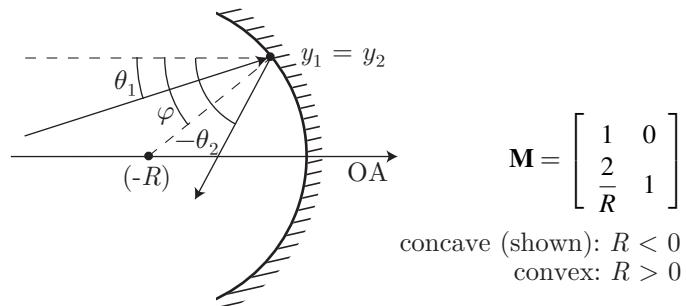
Again, this is a thin optic, so $y_2 = y_1$. The reflection law says that $\theta_2 = \theta_1$. Thus the ray matrix is simply the identity matrix:

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (2.27)$$

(planar mirror $ABCD$ matrix)

So if the ray matrix is the identity, what is the effect of a planar mirror? Really it is just to reverse the direction of propagation. If we adopt an “unfolded” convention, where the (z) optical axis always points in the general direction of the ray travel, then the mirror is really equivalent to nothing, as shown schematically in the above sketch.

4. Spherical Mirror.



For the spherical mirror, we use the sign convention that $R < 0$ for a concave mirror (as shown here) and $R > 0$ for a convex mirror. Thus we will use $(-R) > 0$ in the figure. We will also mark the angle shown as $(-\theta_2)$ because the ray, as it is drawn, points downward (compare to the plane mirror sketch). Again, $y_2 = y_1$ in the paraxial approximation (i.e., the mirror is a thin optic). We can also write the angle with the radius line as

$$\varphi = \frac{y_1}{-R}. \quad (2.28)$$

The Law of Reflection implies that the angles on either side of the radius line are equal:

$$\varphi - \theta_1 = -\theta_2 - \varphi. \quad (2.29)$$

We can rewrite this as

$$\theta_2 = \theta_1 - 2\varphi = \theta_1 + \frac{2y_1}{R}. \quad (2.30)$$

Thus, we can write the ray matrix as

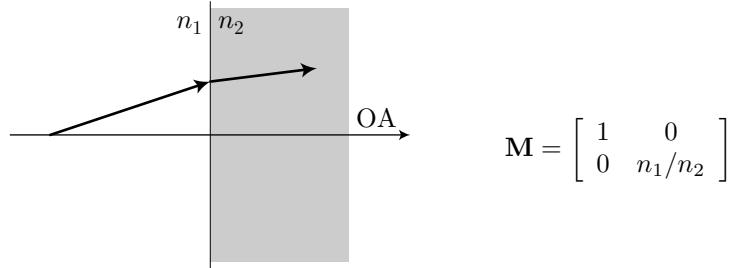
$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ \hline R & 1 \end{bmatrix}. \quad (2.31) \quad (\text{spherical-mirror } ABCD \text{ matrix})$$

Comparing this matrix to the thin-lens matrix, we see that in the paraxial approximation, a spherical mirror is equivalent to a thin lens with a focal length

$$f = -\frac{R}{2}, \quad (2.32)$$

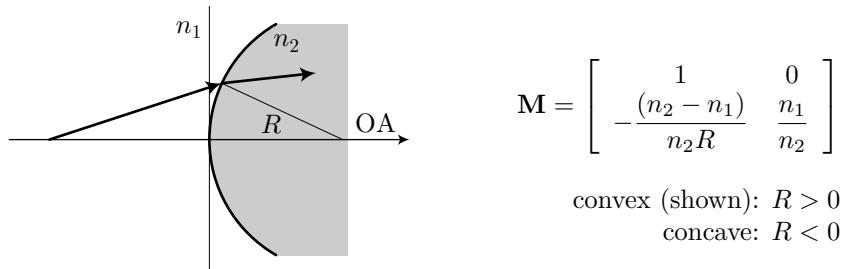
with, of course, the reversal of the optical axis.

5. **Planar Refractive Interface.** A planar refractive interface is thin, so the beam height doesn't change, but the angles change according to Snell's Law (within the paraxial approximation).



We leave the derivation of this $ABCD$ matrix as an exercise (Problem 2.7).

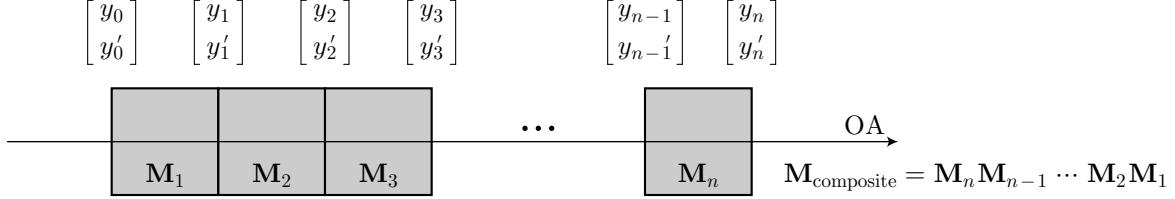
6. **Spherical Refractive Interface.** A spherical refractive interface is a bit more complicated, in the way that the angle changes through the interface, but Snell's Law still of course governs the behavior. (Within the paraxial approximation, this interface is a thin optic, so the height again doesn't change.)



We also leave the derivation of this $ABCD$ matrix as an exercise (Problem 2.7).

2.6 Composite Systems

Now we can consider more general optical systems, or *composite* optical systems made up of the more basic optical elements.



When propagating the ray through the composite system, we can start on the first component by applying the first matrix:

$$\begin{bmatrix} y_1 \\ y'_1 \end{bmatrix} = \mathbf{M}_1 \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix}. \quad (2.33)$$

We can repeat this for the second optical element:

$$\begin{bmatrix} y_2 \\ y'_2 \end{bmatrix} = \mathbf{M}_2 \begin{bmatrix} y_1 \\ y'_1 \end{bmatrix} = \mathbf{M}_2 \mathbf{M}_1 \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix}. \quad (2.34)$$

Iterating this procedure, we can arrive at the transformation for the entire system:

$$\begin{bmatrix} y_n \\ y'_n \end{bmatrix} = \mathbf{M}_n \mathbf{M}_{n-1} \cdots \mathbf{M}_2 \mathbf{M}_1 \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix} =: \mathbf{M}_{\text{composite}} \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix}. \quad (2.35)$$

So, the ray matrix of a composite systems is simply the *product* of the individual ray matrices. **Note the right-to-left ordering of the product.**

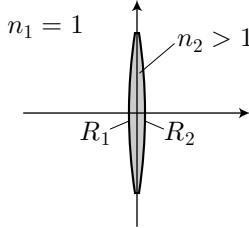
$$\mathbf{M}_{\text{composite}} = \mathbf{M}_n \mathbf{M}_{n-1} \cdots \mathbf{M}_2 \mathbf{M}_1$$

(ABCD matrix for composite optical system) (2.36)

\mathbf{M}_1 acts *first* on the input ray, so it must be the rightmost in the product.

2.6.1 Example: Thin Lens

We can regard a thin lens as a composition of two cascaded refractive interfaces. Since the lens is thin, we assume that there is no distance between the interfaces.



The composite matrix is thus

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{(n_1 - n_2)}{n_1 R_2} & \frac{n_2}{n_1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{(n_2 - n_1)}{n_2 R_1} & \frac{n_1}{n_2} \end{bmatrix} = \begin{bmatrix} \frac{(n_2 - n_1)}{n_1} \left(\frac{1}{R_2} - \frac{1}{R_1} \right) & 0 \\ (n_2 - 1) \left(\frac{1}{R_2} - \frac{1}{R_1} \right) & 1 \end{bmatrix}. \quad (2.37)$$

If we take $n_1 = 1$, this simplifies to

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ (n_2 - 1) \left(\frac{1}{R_2} - \frac{1}{R_1} \right) & 1 \end{bmatrix}. \quad (2.38)$$

If we compare this to the standard thin-lens matrix, we can equate the C matrix entries and write

$$\frac{1}{f} = -(n_2 - 1) \left(\frac{1}{R_2} - \frac{1}{R_1} \right), \quad (2.39)$$

(Lensmaker's formula)

which is known as the **Lensmaker's formula**. The sign conventions work out as follows. For a *convex* lens, we have $R_1 \geq 0$ and $R_2 \leq 0$, which means that

$$\frac{1}{f} = (n_2 - 1) \left(\frac{1}{|R_2|} + \frac{1}{|R_1|} \right) \Rightarrow f > 0, \quad (\text{convex lens}) \quad (2.40)$$

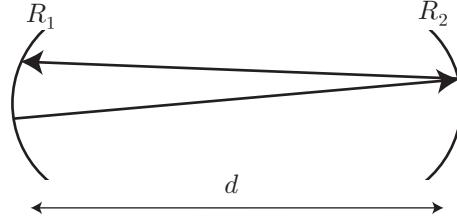
and thus a positive focal length. For a *concave* lens, we have $R_1 \leq 0$ and $R_2 \geq 0$, which means that

$$\frac{1}{f} = -(n_2 - 1) \left(\frac{1}{|R_2|} + \frac{1}{|R_1|} \right) \Rightarrow f < 0, \quad (\text{concave lens}) \quad (2.41)$$

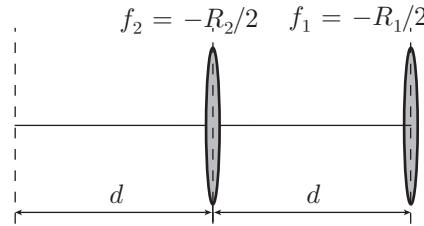
and thus a negative focal length.

2.7 Resonator Stability

We want to consider **resonators**, or optical systems that trap light rays. Such things are very important for the operation of lasers, where light often needs to pass through a gain medium many times, or for interferometry, as we'll get to later. As a basic example, let's look at a resonator composed of two spherical mirrors separated by a distance d :



It is easier to analyze this if we “unwrap” the system into an equivalent waveguide of lenses as follows:



This is just the “unit cell” of the waveguide, which repeats over and over again for each round-trip of the ray in the cavity. We have exploited the equivalence of spherical mirrors and thin lenses here. As we have drawn it, i.e., for 2 concave mirrors, $f_{1,2} = |R_{1,2}/2|$.

The matrix for one round trip (or the waveguide unit cell) is the product of two free-space propagation matrices and two thin-lens (spherical mirror) matrices:

$$\mathbf{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{bmatrix} \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}. \quad (2.42)$$

Note the proper order of multiplication, which follows the path of the ray through the cavity/waveguide: first (the rightmost matrix) is left-to-right propagation of distance d , then reflection off the right mirror,

right-to-left propagation, and reflection off the left mirror (the leftmost matrix in the product). Multiplying this all out, we get

$$\mathbf{M} = \begin{bmatrix} 1 - \frac{d}{f_2} & d \left(2 - \frac{d}{f_2} \right) \\ -\frac{1}{f_1} - \frac{1}{f_2} + \frac{d}{f_1 f_2} & \left(1 - \frac{d}{f_1} \right) \left(1 - \frac{d}{f_2} \right) - \frac{d}{f_1} \end{bmatrix} \quad (2.43)$$

for the cavity round-trip matrix.

2.7.1 Stability Condition

The question we want to ask now is, does the resonator confine the ray? In other words, is the cavity **stable**? Consider the ray after n round trips in the cavity:

$$\begin{bmatrix} y_n \\ y'_n \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^n \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix} \quad (2.44)$$

To answer this question, we can diagonalize the matrix. Recall that the characteristic polynomial for a 2×2 matrix is

$$\lambda^2 - \text{Tr}(\mathbf{M})\lambda + \det(\mathbf{M}) = 0. \quad (2.45)$$

For a ray matrix, there is a general result that states that (see Problem 2.9)

$$\det(\mathbf{M}) = \frac{n_1}{n_2}, \quad (2.46)$$

where n_1 is the refractive index at the input of the optical system and n_2 is the refractive index at the output. For the matrix describing a single pass through a resonator, the input and output are exactly the same place, thus $n_1 = n_2$ and $\det(\mathbf{M}) = 1$. Thus, the characteristic polynomial becomes

$$\lambda^2 - \text{Tr}(\mathbf{M})\lambda + 1 = 0. \quad (2.47)$$

The eigenvalues of \mathbf{M} are the roots of this polynomial, given by

$$\lambda_{\pm} = \beta \pm \sqrt{\beta^2 - 1}, \quad (2.48)$$

where $\beta := \text{Tr}(\mathbf{M})/2 = (A + D)/2$.

So how does this help? Remember that we can decompose an arbitrary vector into eigenvectors. In particular, for the initial condition vector we can write

$$\begin{bmatrix} y_0 \\ y'_0 \end{bmatrix} = \alpha_+ \begin{bmatrix} y_+ \\ y'_+ \end{bmatrix} + \alpha_- \begin{bmatrix} y_- \\ y'_- \end{bmatrix} \quad (2.49)$$

for some constants α_{\pm} , and the vectors $[y_{\pm} \ y'_{\pm}]^T$ are the eigenvectors corresponding to λ_{\pm} :

$$\mathbf{M} \begin{bmatrix} y_{\pm} \\ y'_{\pm} \end{bmatrix} = \lambda_{\pm} \begin{bmatrix} y_{\pm} \\ y'_{\pm} \end{bmatrix}. \quad (2.50)$$

Then after one round trip in the cavity,

$$\mathbf{M} \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix} = \alpha_+ \lambda_+ \begin{bmatrix} y_+ \\ y'_+ \end{bmatrix} + \alpha_- \lambda_- \begin{bmatrix} y_- \\ y'_- \end{bmatrix}, \quad (2.51)$$

and after n passes,

$$\begin{bmatrix} y_n \\ y'_n \end{bmatrix} = \mathbf{M}^n \begin{bmatrix} y_0 \\ y'_0 \end{bmatrix} = \alpha_+ \lambda_+^n \begin{bmatrix} y_+ \\ y'_+ \end{bmatrix} + \alpha_- \lambda_-^n \begin{bmatrix} y_- \\ y'_- \end{bmatrix}. \quad (2.52)$$

Let's simplify things a bit by only considering the positions y_n . Then we can write

$$y_n = (\alpha_+ y_+) \lambda_+^n + (\alpha_- y_-) \lambda_-^n =: \gamma_+ \lambda_+^n + \gamma_- \lambda_-^n, \quad (2.53)$$

where we are introducing the new constants $\gamma_{\pm} := \alpha_{\pm}y_{\pm}$ for notational convenience.

There are two possibilities that we have to consider: either $|\beta| \leq 1$ or $|\beta| > 1$. Let's consider the $|\beta| > 1$ case first. Then clearly λ_{\pm} are real, since the argument of the radical is positive: $\beta^2 - 1 > 0$. We can also see that $|\lambda_+| > 1$ and $|\lambda_-| < 1$ (in fact, $\lambda_+\lambda_- = 1$, since $\det(\mathbf{M}) = \lambda_+\lambda_- = 1$). Now let's reexamine the solution:

$$y_n = \gamma_+\lambda_+^n + \gamma_-\lambda_-^n. \quad (2.54)$$

The first term *grows exponentially* with n , while the second *damps* exponentially away. Thus, for *generic* initial conditions (i.e., $\gamma_+ \neq 0$), the solution grows exponentially as $y_n \sim \lambda_+^n$. This is the **unstable case**, since the solution runs away to infinity.

Now let's consider the other case, $|\beta| \leq 1$. Then $\beta^2 - 1 < 0$, so we can write

$$\lambda_{\pm} = \beta \pm i\sqrt{1 - \beta^2}, \quad (2.55)$$

and clearly now the eigenvalues are complex. Note also that both eigenvalues have unit modulus:

$$|\lambda_{\pm}|^2 = \lambda_{\pm}\lambda_{\pm}^* = (\beta \pm i\sqrt{1 - \beta^2})(\beta \mp i\sqrt{1 - \beta^2}) = \beta^2 + (1 - \beta^2) = 1. \quad (2.56)$$

So $|\lambda_{\pm}| = 1$, and thus it follows that $|\lambda_{\pm}^n| = 1$. We can already see that y_n will stay bounded as n increases, so this is the **stable case**. Let's see this more explicitly: define $\phi := \cos^{-1} \beta$

$$\beta = \cos \phi, \quad \sqrt{1 - \beta^2} = \sin \phi. \quad (2.57)$$

Then $\lambda_{\pm} = \exp(\pm i\phi)$, and $\lambda_{\pm}^n = \exp(\pm in\phi)$. Thus we can write the solution as

$$y_n = \gamma_+e^{in\phi} + \gamma_-e^{-in\phi} = y_{\max} \sin(n\phi + \phi_0) \quad (2.58)$$

for some constants y_{\max} and ϕ_0 , which can be obtained from γ_{\pm} noting that y_n must be real-valued. In other words, each pass through the cavity simply increments the phase of a harmonic oscillation by some fixed amount ϕ .

As a side note, in both cases we can determine the constants of the motion from the initial condition. We can do this by finding the eigenvectors and decomposing the initial ray vector, or, for example, we can equivalently use y_0 and y_1 :

$$y_0 = \gamma_+ + \gamma_-, \quad y_1 = \gamma_+\lambda_+ + \gamma_-\lambda_-. \quad (2.59)$$

Solving these two equations leads to

$$\gamma_{\pm} = \frac{y_1 - y_0\lambda_{\mp}}{\lambda_{\pm} - \lambda_{\mp}} \quad (2.60)$$

as a compact formula for the coefficients.

Thus the **stability condition** for the ray to remain bounded in the long term is simply $|\beta| \leq 1$. We can also write

$$|\text{Tr}(\mathbf{M})| \leq 2 \quad (2.61) \quad (\text{resonator stability condition})$$

or

$$|A + D| \leq 2 \quad (2.62) \quad (\text{resonator stability condition})$$

for the stability condition explicitly in terms of the matrix elements.

2.7.2 Periodic Motion

A condition more restrictive than the stability condition is the **periodic ray condition**, which states that the ray repeats itself exactly after s passes through the cavity, where s is some integer:

$$y_{m+s} = y_m \text{ for all } m. \quad (2.63)$$

If s is the smallest integer for which this is true, then s is called the **period** of the ray. Note that in the paraxial approximation, the existence of a single periodic ray implies that *all* rays for the optical system are periodic (except in trivial cases), because the matrix \mathbf{M}^s collapses to the identity. More generally, though, in nonlinear systems both periodic and nonperiodic rays are possible in the same system, just depending on the initial condition.

We can explore this a bit further mathematically. Obviously for a ray to be periodic the resonator must be stable. We can thus rewrite the periodic-ray condition (2.63) as

$$\gamma_+ e^{in\phi+is\phi} + \gamma_- e^{-in\phi-is\phi} = \gamma_+ e^{in\phi} + \gamma_- e^{-in\phi}. \quad (2.64)$$

This holds for arbitrary coefficients if the phases differ by exact multiples of 2π , i.e.,

$$s\phi = 2\pi q \quad (2.65)$$

for some integer q . Thus,

$$\beta = \cos\left(\frac{2\pi q}{s}\right) \quad (2.66)$$

(periodic-ray condition)

for integers q and s for periodic motion to occur.

2.7.3 Resonator Stability: Standard Form

For the two-mirror resonator that we started out with, we can write out the stability condition more explicitly. Starting with $|(A + D)/2| \leq 1$, we can write

$$0 \leq \frac{A + D + 2}{4} \leq 1, \quad (2.67)$$

which, after inserting the matrix elements from Eq. (2.43) and a bit of algebra, we can rewrite the stability condition as

$$0 \leq \left(1 - \frac{d}{2f_1}\right) \left(1 - \frac{d}{2f_2}\right) \leq 1. \quad (2.68)$$

(resonator stability condition)

It is conventional to define the **stability parameters**

$$g_{1,2} := \left(1 - \frac{d}{2f_{1,2}}\right) = \left(1 + \frac{d}{R_{1,2}}\right), \quad (2.69)$$

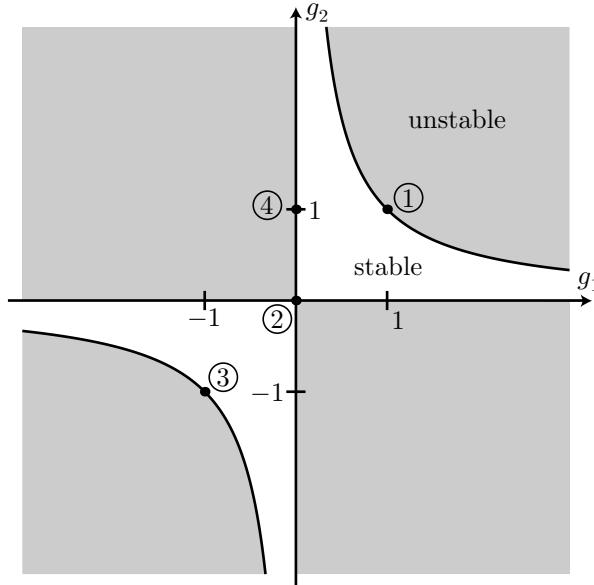
(cavity stability parameters)

where the rightmost expression applies to the original two-mirror resonator rather than the equivalent lens waveguide. In terms of these parameters, the stability condition is particularly simple:

$$0 \leq g_1 g_2 \leq 1. \quad (2.70)$$

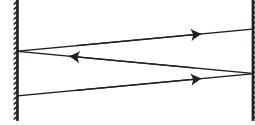
(resonator stability condition)

Then we can sketch the **stability diagram** according to this inequality.

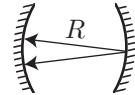


The shaded regions correspond to unstable resonators, the light regions are stable. There are four special cases marked in the diagram that we should consider. All four cases are on the borders of the stability regions, so these are all **marginally stable** cases.

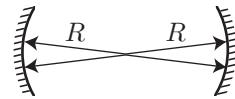
Case 1. $g_1 = g_2 = 1 \implies R_{1,2} = \infty$ (**Planar resonator**)



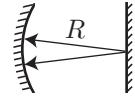
Case 2. $g_{1,2} = 0 \implies R_{1,2} = -d$ (**Confocal resonator**)



Case 3. $g_{1,2} = -1 \implies R_{1,2} = -d/2$ (**Spherical resonator/symmetrical concentric**)



Case 4. $g_1 = 0, g_2 = 1 \implies R_1 = -d, R_2 = \infty$ (**Confocal-planar resonator, or half of a spherical resonator**)



The **planar resonator** is special in the following sense. We can compute

$$\beta = 2g_1g_2 - 1 = 1. \quad (2.71)$$

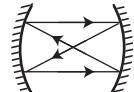
Recalling that we defined the phase angle $\phi = \cos^{-1} \beta$, we find $\phi = 0$ for this case. From Eq. (2.58), we see that $y_n = y_0$ for all n . Hence every ray is periodic with period 1 for a planar resonator. [We can also see this directly by noting that the cavity round-trip matrix (2.43) reduces to $\mathbf{M} = \mathbf{I}_2$ for $f_{1,2} = \infty$.] Actually,

this is an artifact of the eigenvalue formalism; in this marginally stable (“parabolic”) case, the solutions can increase *linearly* rather than exponentially: $y_n = y_0 + ny'_0$.

The **confocal resonator** is one of the most important resonators, as we will see when we discuss optical spectrum analyzers. For the confocal resonator, $d = -R$, so $\beta = 2g_1g_2 - 1 = -1$ and $\phi = \cos^{-1}(-1) = \pi$. From Eq. (2.58), we see that

$$y_n = (-1)y_{n-1} = (-1)^n y_0. \quad (2.72)$$

Thus, every ray is periodic with period 2, so each ray repeats itself after 2 passes through the cavity. [Again, we can also see this directly by noting that the cavity round-trip matrix (2.43) reduces to $\mathbf{M} = -\mathbf{I}_2$ for $f_{1,2} = d/2$.] On a single pass, the position reverses itself (this happens for the angle as well), leading to “figure-eight” orbits in the cavity.



2.8 Nonparaxial Ray Tracing with Interfaces

Tracing *paraxial* rays lends itself to analytic calculations and even simple numerical calculations, using the matrix formalism. Tracing *nonparaxial* rays is much more difficult in general, and analytic calculations typically depend on the existence of fairly simple solutions, symmetries, and a clever setup. Ray tracing for fairly general interfaces can be done readily on a computer however. You can spend hundreds to thousands of dollars on commercial ray-tracing software. Or, armed with Snell’s Law, the Law of Reflection, and a brain, you can get pretty far on your own for free. We will set up the basic formalism for ray tracing here, and go through a simple example to show the general idea, with code in an open-source, mathematical programming language (Octave).

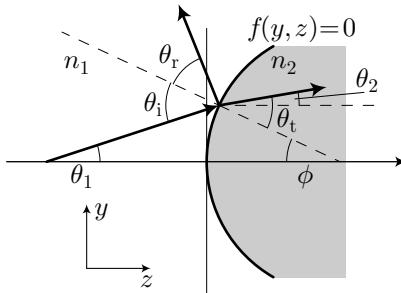
First, let’s go through the basic idea for rays confined to a plane (valid for cylindrical optics, or axially symmetric optics). This will give us some intuition for the problem, and then we’ll redo everything later in a more general form that extends to fully three-dimensional rays.

The setup is as follows. The optical axis lies along the z -direction, and a surface (reflective or refractive interface) is defined by the relation

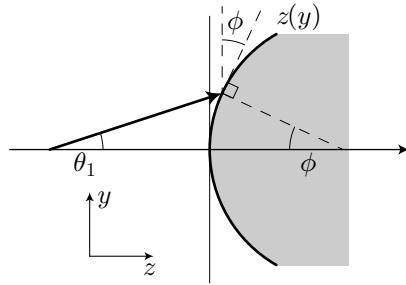
$$f(y, z) = 0. \quad (2.73)$$

(representation of interface)

The incident angle is θ_1 , and we wish to find the refracted or reflected angle θ_2 , given that the ray encounters the interface at the point (y, z) .



To apply Snell’s Law or the Law of Reflection, we will need to track the direction normal to the surface, which we represent via the angle ϕ (the angle between the optical axis and the surface-normal direction). To represent this angle, note that ϕ is also the angle between the y -axis and the tangent to the surface at the intersection point.



And given that $f(y, z) = 0$ can be solved for the alternate function representation $z(y)$ of the surface, we then have

$$\phi = \tan^{-1} \left(\frac{dz}{dy} \right). \quad (2.74)$$

(normal angle of interface)

Then for the case of refraction, we can relate the input and output angles via Snell's Law and reading off the various angles as:

$$\begin{aligned} \theta_i &= \theta_1 + \phi \\ \theta_t &= \sin^{-1} \left(\frac{n_2}{n_1} \sin \theta_i \right) \\ \theta_2 &= \theta_t - \phi. \end{aligned} \quad (2.75)$$

(refraction)

Similarly, in the case of reflection, we have

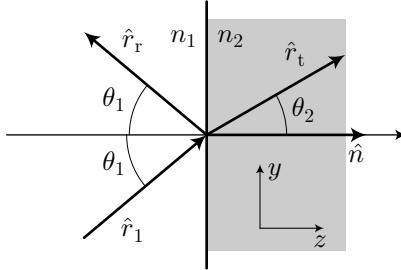
$$\begin{aligned} \theta_i &= \theta_1 + \phi \\ \theta_r &= \theta_i \\ \theta_2 &= \theta_r + \phi, \end{aligned} \quad (2.76)$$

(reflection)

where θ_2 here (not explicitly shown in this case) is still the angle of the outgoing ray with respect to the optical axis. In the reflection case, of course, the z -component of the reflected ray should have the opposite sign with respect to the incident ray. Given that we can find the point of intersection of the incident ray with the interface, then we can trace the outgoing refracted and reflected rays, and then repeat for subsequent interfaces encountered by the continuing ray. We will develop this procedure more explicitly in the more general case below.

2.8.1 Refraction and Reflection Laws: Coordinate-Free Form

The angle-based formalism that we developed above is perfectly adequate for rays confined to a plane. However, for more general ray propagation, the generalization in terms of angles becomes more cumbersome, because the direction of the ray is represented by two angles, and applying Snell's Law involves a reorientation with respect to two *other* angles specifying the surface normal at the intersection point. Therefore, in the more general case, we will take a slightly different approach, and consider refraction at an interface, where we adapt the coordinate system to the surface. The incident ray, whose direction is represented by the unit vector \hat{r}_1 , is incident on the surface as shown, and is either reflected or refracted into the final direction, which we will represent by the unit vector \hat{r}_r or \hat{r}_t , respectively.



In the diagram, we show a planar interface, but this is really just a zoomed version of a general, possibly curved surface, such that it appears flat, and \hat{n} is the unit normal vector to the surface at the point of intersection of the incident ray with the surface. In adapting the coordinate system we are not losing generality, because the idea is to eliminate all explicit references to coordinates, and only express the output direction \hat{r}_r or \hat{r}_t in terms of the incident direction \hat{r}_1 and the normal vector \hat{n} . Note that there are *two* sensible orientations for \hat{n} , corresponding to $\pm\hat{z}$ in this coordinate system, and we should be careful about this.

Starting with the reflected and refracted angles

$$\begin{aligned}\hat{r}_r &= \sin \theta_1 \hat{y} - \cos \theta_1 \hat{z} \\ \hat{r}_t &= \sin \theta_2 \hat{y} + \cos \theta_2 \hat{z},\end{aligned}\tag{2.77}$$

we can eliminate \hat{y} and \hat{z} in favor of \hat{r}_1 and \hat{n} to write (Problem 2.23)

$$\begin{aligned}\hat{r}_r &= \hat{r}_1 - 2 \operatorname{sgn}(\hat{n} \cdot \hat{r}_1) \cos \theta_1 \hat{n} \\ \hat{r}_t &= \frac{n_1}{n_2} \hat{r}_1 + \operatorname{sgn}(\hat{n} \cdot \hat{r}_1) \left(\cos \theta_2 - \frac{n_1}{n_2} \cos \theta_1 \right) \hat{n}.\end{aligned}\tag{2.78}$$

Then eliminating the remaining angles, and thus any explicit references to the coordinate system, we can obtain the coordinate-independent forms (Problem 2.24)

$$\begin{aligned}\hat{r}_r &= \hat{r}_i - 2(\hat{n} \cdot \hat{r}_i)\hat{n} \\ \hat{r}_t &= \left(\frac{n_1}{n_2} \right) \hat{r}_i + \left[\operatorname{sgn}(\hat{n} \cdot \hat{r}_i) \sqrt{1 - \left(\frac{n_1}{n_2} \right)^2 |\hat{n} \times \hat{r}_i|^2} - \frac{n_1}{n_2} (\hat{n} \cdot \hat{r}_i) \right] \hat{n}.\end{aligned}\tag{2.79}$$

(reflected and refracted directions, coordinate-independent)

for the reflected and refracted rays. Again, we have been careful to keep our results invariant under the replacement $\hat{n} \rightarrow -\hat{n}$, so that either possible normal vector of the surface is suitable for our purposes.

At this point it is worth expanding on a couple of details. In the formula for the refracted ray in Eqs. (2.79), note that the cross product $|\hat{n} \times \hat{r}_i|$ can take any value between 0 and 1. Thus, if $n_1 > n_2$, it is possible for the argument of the square root to become negative, and in this case \hat{r}_t becomes imaginary. This means that there is in fact no transmitted ray; the entire ray is reflected, and this phenomenon is called **total internal reflection**, a topic to which we will return in the context of waves (Section 9.3). Similarly, we have mentioned that at an arbitrary interface the possibilities of both reflection and refraction of a ray. The fraction reflected vs. refracted is best discussed in the context of wave optics (Chapters 9 and 10), and is complicated somewhat by the possibility of wave interference from reflections from different interfaces. Furthermore, when tracing all possible reflected and refracted rays, the number of rays proliferates exponentially with the number of interfaces involved. Typically, a ray-tracing analysis focuses on one set of rays to keep the problem tractable (e.g., tracing the refracted rays through a compound-lens system, or the reflection from a single optical surface), and therefore requires some thought and care in the setup of the analysis.

2.8.2 Ray Tracing: A Recipe

Now with the formulae (2.79) in hand to compute the direction of the ray after interacting with the interface, the only remaining tasks are to find the intersection of the incoming ray with the surface, and then to determine the normal vector.

Since we define the interface in terms of the vanishing of a function,

$$f(\mathbf{r}) = 0, \quad (2.80)$$

(representation of interface)

finding the intersection of a ray with the surface amounts to a root-finding problem. If we assume homogeneous media between interfaces, then the rays are straight lines, and this reduces our task to a one-dimensional root-finding problem, which is a fairly simple one. More explicitly, we can take the ray, described by its location \mathbf{r} , to be parameterized by the location along the optical (z) axis (at least, between encounters with interfaces, where ray may reverse its direction with respect to the optical axis), so that we may write the ray as $\mathbf{r}(z)$. Then we are finding the root of $f(z) = f[\mathbf{r}(z)]$, or solving

$$f(z) = f\left[x_0 + \frac{r_{ix}}{r_{iz}}(z - z_0), y_0 + \frac{r_{iy}}{r_{iz}}(z - z_0), z\right] = 0, \quad (2.81)$$

(interface-intersection condition)

where $\mathbf{r}_0 = (x_0, y_0, z_0)$ is some reference location for the ray (say the “source point” of the ray), and $\hat{\mathbf{r}}_i = (r_{ix}, r_{iy}, r_{iz})$ is the direction unit vector of the incident ray (the same as the incident vector in the previous section). Here we have written the x and y coordinates of the ray in terms of the initial point and the tangents r_{ix}/r_{iz} and r_{iy}/r_{iz} of the ray-propagation angles. In general, finding the intersection with a surface requires searching the function for various z to first bracket the interface. Mathematically, two points z_1 and z_2 bracket the interface [i.e., bracket the root of $f(z)$] if the function changes sign between these two points:

$$f(z_1)f(z_2) < 0. \quad (2.82)$$

(root-bracketing condition)

This can often be done through knowledge of the geometry, but in more general-purpose software, a range of z must be searched systematically in an attempt to bracket the intersection, if one exists (the ray can, of course, sail past the edge of an optic). Once bracketed, there are several simple algorithms to find the root. The bisection algorithm for example, says to try the midpoint $\bar{z} = (z_1 + z_2)/2$, and then use the same condition to see whether (z_1, \bar{z}) or (\bar{z}, z_2) bracket the root. The process can be iterated to find an arbitrarily narrow interval that brackets the root.

Once we determine the intersection of the ray with a surface, we need to find the direction of the outgoing ray. The unit-normal vector, specifying the orientation of the surface at the point of incidence, is simply

$$\hat{\mathbf{n}} = \frac{\nabla f}{|\nabla f|}, \quad (2.83)$$

(normal vector to surface $f = 0$)

evaluated at the point of incidence. Intuitively, this is because the surface is defined by the constant value $f = 0$, and so the maximum change in f occurs perpendicular to the surface. Then the outgoing-ray direction is set by Eqs. (2.79); recall that the sign of $\hat{\mathbf{n}}$ (which could vary because $f(\mathbf{r})$ is not unique— $-f(\mathbf{r})$ is as good as $f(\mathbf{r})$ for specifying the surface) doesn’t affect the results.

So to summarize the procedure:

1. Begin with the ray at some initial reference point \mathbf{r}_0 , traveling in the direction $\hat{\mathbf{r}}_i$, incident on some interface described by $f(\mathbf{r}) = 0$.
2. Find the point of intersection of the ray with the interface, by bracketing and then finding the root of Eq. (2.81).
3. Determine the normal vector $\hat{\mathbf{n}}$ of the surface at this point, using Eq. (2.83).

4. Use Eqs. (2.79) to find the orientation of the reflected and/or refracted ray.
5. Lather, rinse, repeat: set the reference point \mathbf{r}_0 to the intersection point, use the refracted or reflected direction as the new orientation vector, and repeat the whole process for the next interface (or for the outside boundary of the plot, if this was the last interface). You may need to take care if the direction along the optical axis became reversed by the prior interface.

2.8.3 Example: Parabolic Interface

As an example, let's consider a parabolic interface, where the parabola is specified by

$$z - z_c = ay^2. \quad (2.84)$$

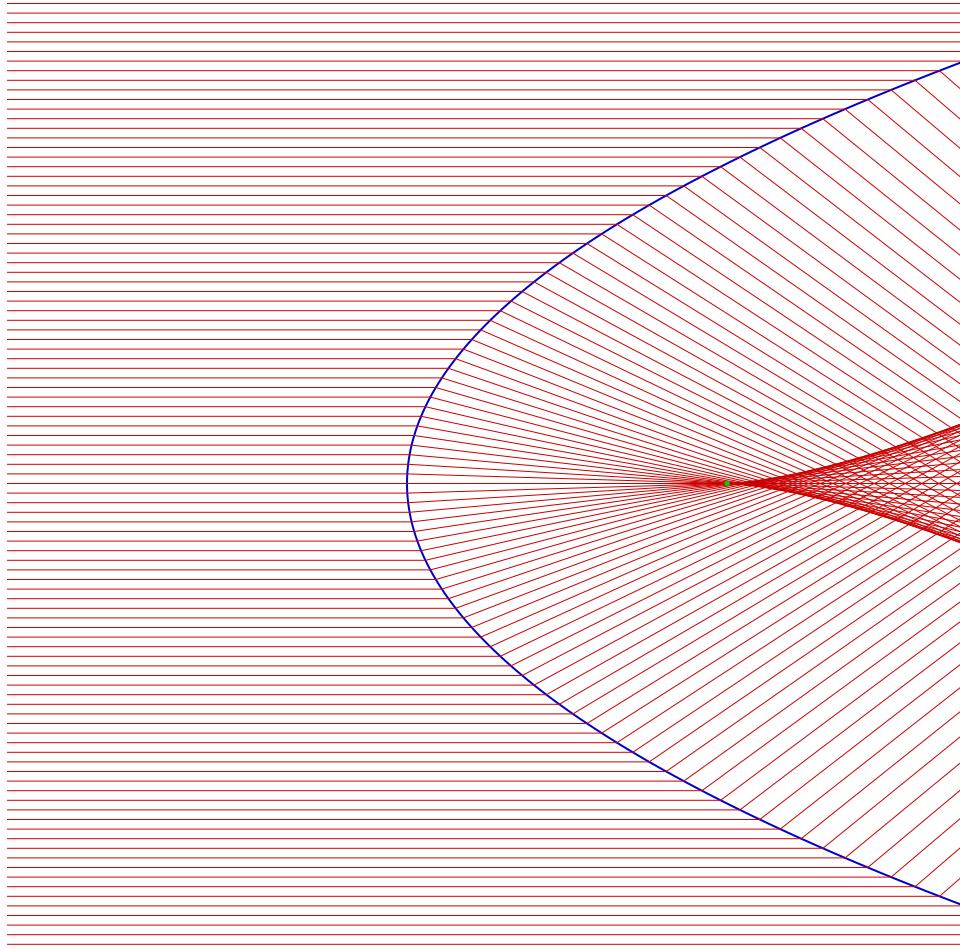
We can write this in our standard notation $f(\mathbf{r}) = 0$ for root-finding as

$$f(y, z) = z - z_c - ay^2 = 0. \quad (2.85)$$

Here z_c is an offset along the optical axis, and a controls the curvature of the ray. In the case where $a > 0$ (such that the parabola opens to the right), and the refractive index to the right is greater than to the left of the interface, the interface will act to focus incoming parallel rays.

2.8.3.1 Refraction

The diagram below shows the refracted rays for the case $a = 1/2$, $n_1 = 1$ to the left of the interface, and $n_2 = 2$ to the right.



The focusing action of the interface is clear, as well as the severe aberrations for rays far from the optical axis. Note that we have also marked the location of the effective focal point with a green dot. This is the common point on the optical axis through which horizontal rays very close to the optical axis will pass. The effective focal length is given by (Problem 2.25)

$$f_{\text{eff}} = \frac{n_2}{2a(n_2 - n_1)} \quad (2.86)$$

with respect to the location of the parabola along the optical axis; thus the absolute z location of the focal point is $z_c + f_{\text{eff}}$.

A program in the Octave¹ mathematical scripting language that traces rays refracting through this surface to generate this diagram is given below. Note that we are using the recipe and the coordinate-independent form of the refraction formula. We only worry about one refraction, and we don't bother with rays that reverse direction, because this doesn't happen here. For finding the incident point of rays on the parabola, we define the `surfacefunc` function, which implements Eq. (2.81). We then use the `fzero` built-in function to find the root of `surfacefunc`, and thus the intersection point. Since `fzero` expects a one-dimensional function, we must pass other parameters to the function via global variables. We also compute the normal vector in `gradfunc`, by explicitly programming the gradient $(\partial f / \partial y)\hat{y} + (\partial f / \partial z)\hat{z}$. In a more general-purpose program, we could just estimate the gradient by computing $f(\mathbf{r})$ at several different points, but the explicit form is pedagogically clearer here. This script is a good basis for ray tracing in other simple systems (e.g., Problem 2.27). The script is available at

http://atomoptics.uoregon.edu/~dsteck/teaching/octave/parabolic_surface_refract.m

and listed below.

```
% parabolic_surface_refract.m
%
% Trace incoming parallel rays for parabolic-surface "lens"
% defined by f(y,z) = z-zc - a*y^2 = 0.
%
% Represent rays by a position y at z, and also a direction unit vector
% rhat = [ry; rz]; z is the direction along the optical axis.
%
% Use coordinate-free form of Snell's Law in terms of surface normal vector.

global gY gRhat gZ0 gA gZc

a = 0.5;          % parabolic parameter
zc = -0.5;        % shift of parabola center
dy = 0.06;         % increment of initial y position of rays
ymax = 6*a;       % max y ray to trace
zmin = -6*a;      % launch z
zmax = 6*a;       % max z
n1 = 1;           % index to left of parabola
n2 = 2;           % index to right of parabola

% In the following function we need to pass extra parameters,
% hence the global statement.
%     gY -> y at z0
%     gRhat -> ray direction (unit) vector, in form [ry; rz]
%     gZ0 -> reference coordinate z0
%     gA -> copy of A
```

¹Octave is free, open-source, mostly compatible with MATLAB, and runs on most platforms: <http://octave.org>.

```

%      gZc -> copy of zc
% Note: global variables are labeled with a 'g' to avoid confusion
gA = a;
gZc = zc;

% function for finding intersections with the parabolic surface
function fout = surfacefunc(z)
    % z is the position along the optical axis
    global gY gRhat gZ0 gA gZc

    % extrapolate ray straight to z
    %   y = y0 + (z-z0)*ry/rz;
    ray_y = gY + (z-gZ0)*gRhat(1)/gRhat(2);

    % function to define surface via f(y,z) = (z-zc) - a*y^2 = 0
    fout = (z-gZc) - gA*ray_y^2;
end %function

% function for finding the normal vector to the surface
%   returns gradient of f = z-zc - a*y^2
function fout_vec = gradfunc(y,z)
    global gA
    fout_vec = [ ...
        -2*gA*y; ... % partial f/partial y
        1           ... % partial f/partial z
    ];
    fout_vec = fout_vec/sqrt(dot(fout_vec,fout_vec)); % make unit vector
end %function

% plot parabola, using parametric form
dt = 0.01;
t=(zmin:dt:zmax)';
plot(a*t.^2+zc, t,'b-'); % plot as blue line

% set axes using z dimensions, and square aspect ratio
axis([zmin, zmax, zmin, zmax], "square");
xlabel('z (mm)');
ylabel('y (mm)');
title(sprintf('ray trace of parabolic interface, n = %.2f',n2));
hold on; % subsequent plots will be overlayed

%%%% loop over rays
nrays = 2*floor(ymax/dy);
for ray = 0:nrays,
    y0 = (ray-nrays/2) *dy;
    yarr = [y0]; % array of y positions to plot
    zarr = [zmin]; % array of z positions

    % find first intersection with the parabolic surface
    %   look for intersection (root) between zmin and 10*zmax
    rhat = [0; 1]; % direction of initial ray, horizontal = zhat
    gY = y0; gRhat = rhat; gZ0 = zmin;

```

```

z = fzero('surfacefunc', [zmin; 10*zmax]);
y = gY;

% add point to plot
yarr = [yarr; y]; zarr = [zarr; z];

% compute refracted ray direction, update direction vector rhat
nhat = gradfunc(y,z); % surface normal vector
nhat3 = [0; nhat]; rhat3 = [0; rhat]; % add x dimension for cross product
crossthing = (n1/n2)*cross(nhat3,rhat3); % (n1/n2)*(n x r)
rhat = (n1/n2)*rhat ...
+ ( sign(dot(nhat,rhat))*sqrt(1-dot(crossthing,crossthing)) ...
- (n1/n2)*dot(nhat,rhat) )*nhat;

% extrapolate to the end of the optical axis
z0 = z; z = zmax;
yarr = [yarr; y + (z-z0)*rhat(1)/rhat(2)]; zarr = [zarr; z];

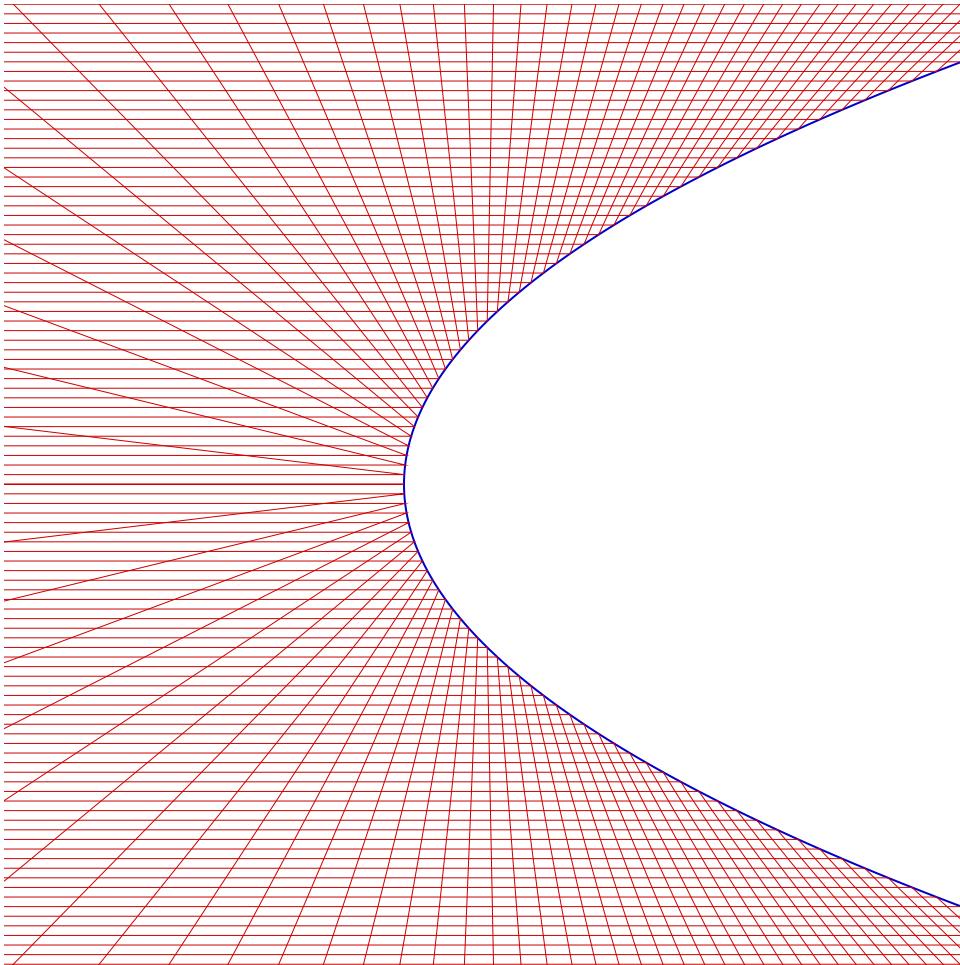
% plot ray
plot(zarr, yarr, 'r-'); % plot as red line
end %for

feff = n2/(2*a*(n2-n1)); % effective, paraxial focal length
% mark effective focal length, account for offset of surface
plot([feff+zc], [0], 'g.');
hold off;

```

2.8.3.2 Reflection

With relatively minor modifications to the script, we can use it to trace the reflected rays from the parabola.



We list only the changed section below, which is the last part of the loop over rays where we find the direction of the outgoing ray. We simply change from the refraction to the reflection formula. Also, we must now handle the possibility of a backwards-propagating ray, so we insert a condition based on the sign of the z -component of \hat{r}_r , and propagate the final ray to either z_{\min} or z_{\max} for backward- or forward-propagating rays, respectively.

```
% compute reflected ray direction, update direction vector rhat
nhat = gradfunc(y,z); % surface normal vector
rhat = rhat - 2*dot(nhat,rhat)*nhat;

% extrapolate to the end of the optical axis
% handle backwards rays based on sign of z component of rhat
if ( rhat(2) > 0 ), % forward case
    z0 = z; z = zmax;
    yarr = [yarr; y + (z-z0)*rhat(1)/rhat(2)]; zarr = [zarr; z];
else, % backward case
    z0 = z; z = zmin;
    yarr = [yarr; y + (z-z0)*rhat(1)/rhat(2)]; zarr = [zarr; z];
end %if
```

The full script to produce the above plot is available at

http://atomoptics.uoregon.edu/~dsteck/teaching/octave/parabolic_surface_reflect.m

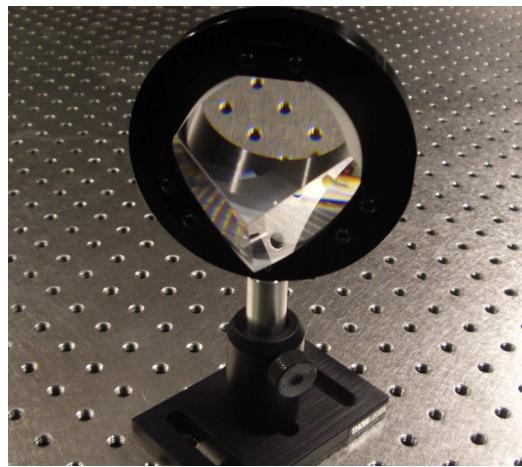
2.9 Exercises

Problem 2.1

A fish is 1 m beneath the surface of a pool of water. How deep does it *appear* to be, from the point of view of an observer above the pool? The refractive index of water is $n = 1.33$.

Problem 2.2

A corner-cube reflector is an arrangement of 3 planar mirrors to form, appropriately enough, the corner of a cube. The reflecting surfaces face the interior of the cube. Commercial corner cubes often use internal reflections to form the mirror surfaces, as in this photograph of the back side of a mounted corner-cube prism.



Show that any ray entering the corner cube is reflected such that the exiting ray is parallel (but opposite) to the incident ray. You may assume that the incident ray's direction is such that it reflects from all 3 surfaces.

Problem 2.3

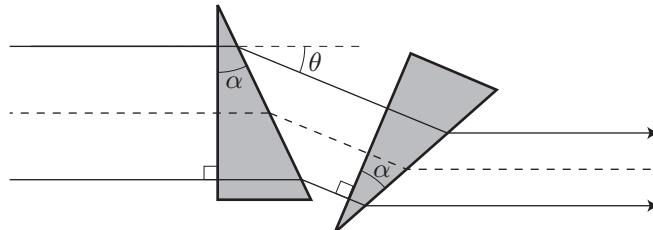
Suppose that a ray in air ($n = 1$) is incident on a planar window of uniform thickness T and refractive index n . If the angle of incidence is θ , show that the transmitted beam is parallel to the incident beam but has a horizontal displacement given by

$$\delta = T \sin \theta \left(1 - \frac{\cos \theta}{n \cos \theta'} \right), \quad (2.87)$$

where $n \sin \theta' = \sin \theta$.

Problem 2.4

An **anamorphic prism pair** is used to expand or shrink a beam in one dimension without deflecting its angle, as shown here.



The pair consists of two identical prisms with wedge angle α and refractive index n . The beam enters each prism at normal incidence to the front surface. As a model of the beam, it is useful to consider parallel rays as shown.

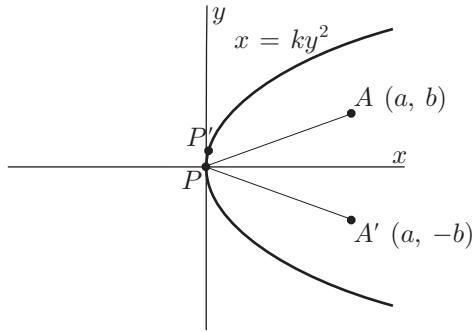
- Write down an expression that relates the deflection angle θ after the first prism to α and n .
- By how much is the beam reduced after the first prism? Write your answer in terms of n and α only.
- By how much is the beam reduced after the second prism?

Problem 2.5

A thin lens is submerged in water ($n = 1.33$). If the focal length is $f = 1$ m in air, what is the focal length in water? Assume the lens is made of fused silica ($n = 1.46$).

Problem 2.6

Consider² a reflection off the center P of a parabolic mirror ($x = ky^2$) as shown.



By symmetry of the reflection at the center, we can fix the endpoints of the reflected ray to have coordinates (a, b) and $(a, -b)$, with $a > 0$.

- Consider a point P' a small distance (with vertical component ε) from P . Show that the optical path length of $AP'A'$ is an extremum when P' coincides with P .
- Show that APA' is a *minimum* when $k < k_c$ and a *maximum* when $k > k_c$, where

$$k_c := \frac{a}{2(a^2 + b^2)}. \quad (2.88)$$

- The locus of all points B such that $ABA' = APA'$ is clearly an ellipse with foci A and A' . Show that the equation describing this ellipse is given by

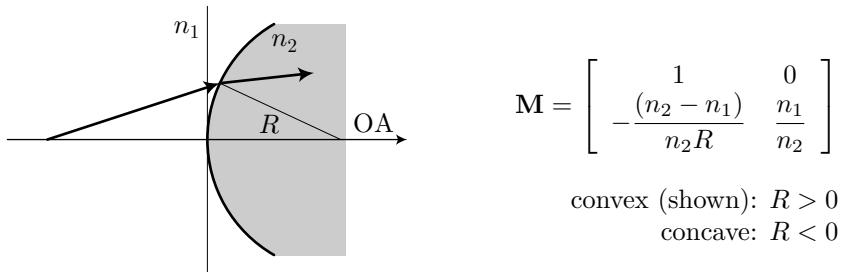
$$\frac{(x-a)^2}{\gamma^2 - b^2} + \frac{y^2}{\gamma^2} = 1, \quad (2.89)$$

where $2\gamma := ABA'$. Argue that this result is consistent with the result of part (b).

Problem 2.7

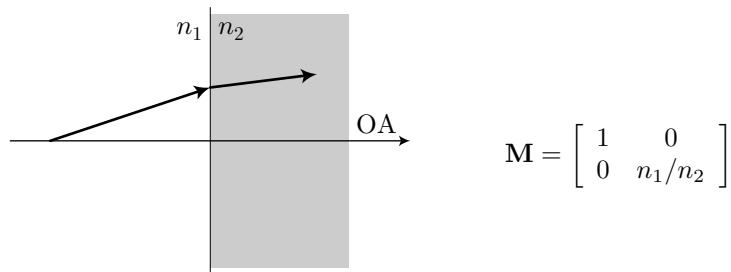
- Derive the $ABCD$ matrix for a refractive spherical boundary:

²Adapted from H. A. Buchdahl, *Introduction to Hamiltonian Optics* (Dover, 1993).



The convention is that $R > 0$ for a convex surface (as shown here) and $R < 0$ for a concave surface. Note that in the paraxial approximation, the height of the ray does not change across the boundary.

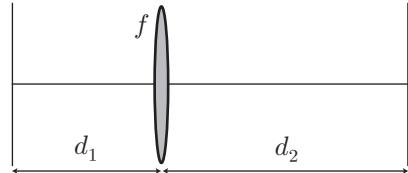
(b) Derive the $ABCD$ matrix for a refractive planar boundary:



(c) Find the determinants of the $ABCD$ matrices in parts (a) and (b).

Problem 2.8

(a) Derive the ray-transfer matrix for free-space propagation, followed by a thin lens, followed by more free-space propagation, as shown in the Figure.



(b) Show that applying the thin-lens law,

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f}, \quad (2.90)$$

all rays originating from a single point y_1 in the input plane reach the output plane at the single point y_2 , independent of the input angle y'_1 . Compute the **magnification** y_2/y_1 .

(c) Show that if $d_2 = f$, all parallel incident rays are focused to a single point in the output plane.

Problem 2.9

Let \mathbf{M} be the ray matrix for an arbitrary optical system where the input and output refractive indices are n_1 and n_2 , respectively. Prove that

$$\det(\mathbf{M}) = \frac{n_1}{n_2} \quad (2.91)$$

(as you saw in Problem 6(c)), using the following outline, which exploits the formal equivalence of ray optics to classical Hamiltonian mechanics.

Recall the action principle (Fermat's principle) for ray optics:

$$\delta \int n(x, y, z) ds = 0, \quad (2.92)$$

where $ds^2 = dx^2 + dy^2 + dz^2$ and $n(x, y, z)$ is a refractive-index profile that models the optical system. Compare to the action principle for classical mechanics. Take the coordinate z to be the “time” variable and the coordinate y to be the position coordinate. Let’s consider the two-dimensional case, so x is an ignorable coordinate, and note that z is also ignorable in the sense of being completely determined by x , y , and s . Then for the optical case, write down the Lagrangian. Show that the conjugate momentum p for the position y is $n dy/ds$, and then write down the Hamiltonian.

Now consider the following transformation relating the **canonical** coordinates before and after the optical system,

$$\begin{bmatrix} y_2 \\ p_2 \end{bmatrix} = \mathbf{M}' \begin{bmatrix} y_1 \\ p_1 \end{bmatrix}, \quad (2.93)$$

which of course is valid in the paraxial approximation (where it is also true that $s \approx z$). Because y and p are canonical variables and \mathbf{M}' represents “time evolution” of a Hamiltonian system, \mathbf{M}' represents a canonical transformation and in particular \mathbf{M}' is a **symplectic matrix**, which implies that $\det(\mathbf{M}') = 1$. This is essentially the content of Liouville’s theorem.

Using this result, transform to the standard (noncanonical) variables y and y' , and compute the determinant of \mathbf{M} .

A very brief review of variational principles in classical mechanics may help. Recall that the **action functional** is given by the integral

$$S[L] := \int_{t_1}^{t_2} L(q, \dot{q}; t) dt, \quad (2.94)$$

where the Lagrangian L is typically of the form $L = T(\dot{q}) - V(q)$ in particle mechanics. The variational principle (**Hamilton’s principle**) is $\delta S[L] = 0$, which for our purposes implies the Euler-Lagrange equation

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}^\alpha} - \frac{\partial L}{\partial q^\alpha} = 0 \quad (2.95)$$

under the condition that the endpoints of the variation are fixed ($\delta q(t_1) = \delta q(t_2) = 0$). The Hamiltonian is given by a Legendre transformation of the Lagrangian via

$$H(q, p; t) := \dot{q}^\alpha p_\alpha - L(q, \dot{q}; t), \quad (2.96)$$

where the conjugate momenta are defined by $p_\alpha := \partial L / \partial \dot{q}^\alpha$.

Problem 2.10

- (a) Suppose that two thin lenses of focal length f_1 and f_2 are placed in contact. Show that the combination acts as a thin lens with a focal length given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}. \quad (2.97)$$

- (b) The **optical power** of a lens is defined as $1/f$, where f is the focal length of the lens. Typically the lens power is measured in **diopters**, defined as $1/f$ where f is measured in meters (i.e., a lens with a 100 mm focal length has a power of 10 diopters). Based on your answer for part (a), why is the optical power a natural way to characterize a thin lens?

Problem 2.11

- Show that the effective focal length f_{eff} of two lenses having focal lengths f_1 and f_2 , separated by a distance d , is given by

$$\frac{1}{f_{\text{eff}}} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}. \quad (2.98)$$

Note that this system is no longer a thin lens, so for this to work out you should show that the effect of the two-lens optical system is equivalent to that of a single lens of focal length f_{eff} , with free-space propagation of distances d_1 and d_2 before and after the single lens, respectively, where

$$\frac{1}{d_1} = \frac{1}{d} - \frac{1}{f_1} + \frac{f_2}{f_1 d}, \quad \frac{1}{d_2} = \frac{1}{d} - \frac{1}{f_2} + \frac{f_1}{f_2 d}. \quad (2.99)$$

This problem is a simple model for one realization of a “zoom” or **variable focus** lens, e.g., for still photography. (A “true zoom” lens would also shift the location of the pair to maintain focus as the focal length changes, whereas a simpler variable focus lens merely changes the separation to achieve different focal lengths, possibly requiring a refocusing adjustment.)

Problem 2.12

Because of **dispersion**, the index of refraction varies slightly with the wavelength of light, and thus the focal length of a thin lens varies slightly with optical wavelength. This effect is called **chromatic aberration**. A common technique to correct for this aberration is to cement two lenses together of different materials to form an **achromatic doublet** or **achromat**.

For this problem, assume a simple linear model of the refractive-index variation:

$$n(\lambda) \approx n(\lambda_0) + \left(\frac{dn}{d\lambda} \right)_{\lambda=\lambda_0} (\lambda - \lambda_0), \quad (2.100)$$

where λ_0 is some wavelength in the center of the region of interest. The dispersion of an optical glass is often characterized by its refractive indices at three special wavelengths, the **Fraunhofer C, D, and F lines**, given by $\lambda_C = 656.3$ nm, $\lambda_D = 587.6$ nm, and $\lambda_F = 486.1$ nm, named after Fraunhofer’s catalog of the dark features in the solar spectrum. In terms of the three indices, we can define the **Abbé v-constant** by

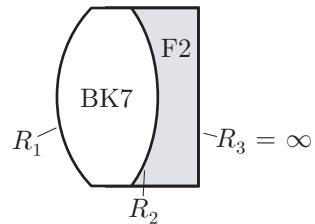
$$v_D := \frac{n_D - 1}{n_F - n_C}, \quad (2.101)$$

in terms of which we can write the refractive index as

$$n(\lambda) \approx 1 + (n_D - 1) \left[1 - \left(\frac{\delta\lambda}{v_D(\lambda_C - \lambda_F)} \right) \right], \quad (2.102)$$

where $\delta\lambda := \lambda - \lambda_D$.

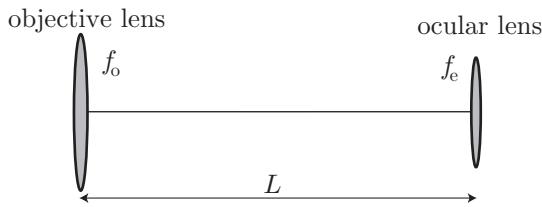
Design a thin, achromatic doublet with the following materials: BK7 borosilicate crown glass ($n_F = 1.52238$, $n_D = 1.51673$, $n_C = 1.51432$) and F2 flint glass ($n_F = 1.63208$, $n_D = 1.61989$, $n_C = 1.61503$). Use BK7 for the first section in the shape of a biconvex lens, and F2 for the second section in the shape of a plano-concave lens, as shown.



Obviously, the radii of curvature at the cemented interface should match. Make the thin-lens approximation and choose the two radii of curvature to achieve a lens with $f = 100$ mm over the visible spectrum.

Problem 2.13

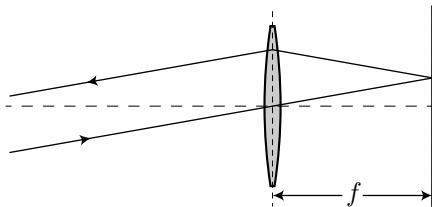
A **refracting telescope** is an optical system consisting of two thin lenses with a fixed space in between. Light first enters the **objective lens** of focal length f_o , propagates over a distance L , and goes through the **ocular lens** (eyepiece lens) of focal length f_e . The length of the telescope satisfies $L = f_o + f_e$.



- (a) Construct the $ABCD$ matrix for propagation through a telescope (from left to right in this diagram).
- (b) Show that a telescope produces **angular magnification**. That is, incoming rays with angle θ_1 from the optical axis exit the system at angle $-(f_o/f_e)\theta_1$, independent of the initial ray position y_1 .
- (c) Show that a telescope can also act as a **beam reducer** (or expander): i.e., show that a bundle of rays of diameter d traveling parallel to the optical axis has diameter $(f_e/f_o)d$ when exiting the eyepiece.
- (d) Sketch a **Keplerian telescope**, where both focal lengths are positive. Draw in the parallel rays corresponding to part (c). Also sketch a ray that is initially not parallel to the optical axis; use the thin lens law to justify the minus sign in the angular magnification of part (b).
- (e) Clearly, a telescope with positive angular magnification (i.e., a telescope that produces an upright image) has exactly one lens with negative focal length so that $(-f_o/f_e) > 0$. Such a refracting telescope is known as a **Galilean telescope**. Which of the two lenses can have negative focal length if the telescope produces a magnified (not reduced) image?

Problem 2.14

A retroreflector is any optic that reflects an incident ray, such that the exiting ray is parallel (but opposite) to the incoming ray. One version of a “cat’s eye” retroreflector uses a thin lens of focal length f and a mirror as shown.



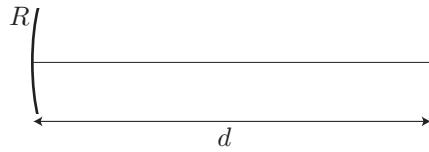
Set up the ray matrix for this optical system to prove that it is, indeed, a retroreflector.

Problem 2.15

- (a) A **ball lens** is a sphere of optical material (e.g., glass or sapphire) that is used as a lens, especially for coupling light into or out of an optical fiber. Within the paraxial approximation, derive an expression for the location where a beam of collimated incoming light (aimed at the sphere’s center) will come to a focus, assuming a sphere of diameter D and refractive index n , surrounded by vacuum. (Your answer should be in the form of a distance from the center of the ball lens, thus giving its **effective focal length**.)
- (b) What is the condition for incoming rays parallel to the optical axis to be imaged onto the back surface of the sphere? Argue that under these conditions, the sphere acts as a retroreflector (incoming rays are reflected back along the same direction of incidence). This is the traditional realization of a “cat’s eye” retroreflector, and is the reason, for example, that some painted lines on roads and road signs look so bright at night under headlights (small silica spheres embedded in the paint reflect headlight back to your eyes).

Problem 2.16

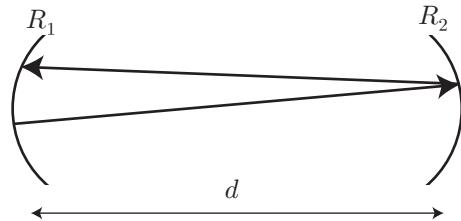
Consider a two-mirror resonator, as shown here. One mirror is concave ($R < 0$) and the other is flat.



- (a) What is the round-trip ray matrix for this resonator? Derive the matrix for a ray starting just to the right of the curved mirror, traveling to the right.
- (b) For what range of d is the cavity stable?
- (c) Derive the ray matrix for **two** round trips for the special case $R = -2d$. Sketch an example ray that illustrates your answer.

Problem 2.17

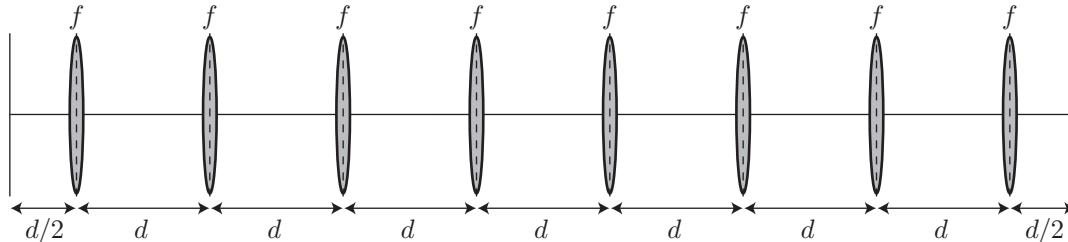
Suppose you have two convex, spherical mirrors in a gas laser resonator, with unknown and possibly different radii of curvature R_1 and R_2 .



Suppose also that you can vary the length d of the cavity, and that after much labor you find that the laser operates in the ranges $d < 50$ cm and 100 cm $< d < 150$ cm. What are the numerical values of R_1 and R_2 ? Keep in mind that gas lasers have relatively low gain per pass, and thus proper laser operation requires that the light makes many round trips inside the resonator before leaking out.

Problem 2.18

- (a) Consider a cavity consisting of two planar mirrors and identical thin lenses of focal length f , regularly spaced as shown.



For a given set of lenses, what is the range of d for which the cavity is stable?

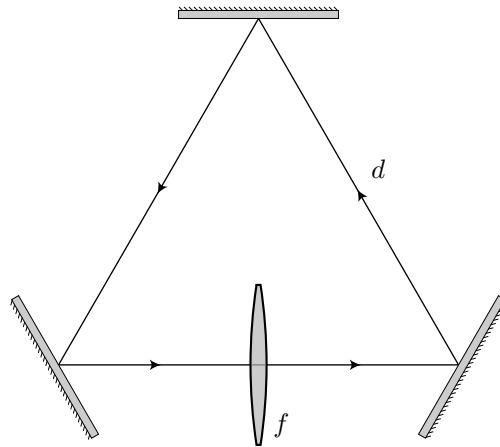
- (b) Write down the eigenvalues of the round-trip ray matrix for this cavity.

Problem 2.19

Consider a symmetric cavity consisting of two planar mirrors separated by 1 m. Suppose that a thin lens of focal length f is placed inside the cavity against one of the mirrors. For what range of f is the cavity stable?

Problem 2.20

Consider the ring cavity shown below, where the optical axis forms an equilateral triangle, with sides of length d , and a single lens of focal length f centered in the bottom leg of the cavity.



- (a) For what range of f is the cavity stable?
 (b) Discuss how moving the lens to the left or right influences the stability of the cavity.

Problem 2.21

Suppose an optical resonator is represented by the round-trip matrix \mathbf{M} , which satisfies

$$\mathbf{M} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{3}{2} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (2.103)$$

- (a) What are the two eigenvalues of \mathbf{M} ?
 (b) Is the resonator stable? *Explain.*
 (c) If your answer to (b) was “yes,” are there any rays that diverge after many round trips? If your answer to (b) was “no,” are there any rays that remain stable? In either case, give an example and give a physical interpretation of the rays.

Problem 2.22

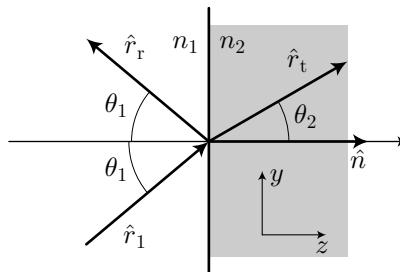
In the matrix formalism of optics, we have stuck to modeling only one dimension (y), transverse to the optical axis (z). Suppose we want to model *both* transverse directions x and y , by an analogous matrix formalism, with a four-vector of the form

$$\begin{bmatrix} x \\ x' \\ y \\ y' \end{bmatrix} \quad (2.104)$$

for the state of the vector at some point along the optical axis. Write down the ray-transfer matrix for this four-vector for free-space propagation over a distance d .

Problem 2.23

Using the setup in this ray-tracing diagram,



show that the unit vectors for the reflected and refracted rays may be written in the form of Eqs. (2.78),

$$\begin{aligned}\hat{r}_r &= \hat{r}_1 - 2 \operatorname{sgn}(\hat{n} \cdot \hat{r}_1) \cos \theta_1 \hat{n} \\ \hat{r}_t &= \frac{n_1}{n_2} \hat{r}_1 + \operatorname{sgn}(\hat{n} \cdot \hat{r}_1) \left(\cos \theta_2 - \frac{n_1}{n_2} \cos \theta_1 \right) \hat{n}.\end{aligned}\quad (2.105)$$

Problem 2.24

Starting with the results and setup of Problem 2.23, derive the coordinate-independent forms for the normal vectors of the refracted and reflected rays from Eqs. (2.79),

$$\begin{aligned}\hat{r}_r &= \hat{r}_i - 2(\hat{n} \cdot \hat{r}_i)\hat{n} \\ \hat{r}_t &= \left(\frac{n_1}{n_2} \right) \hat{r}_i + \left[\operatorname{sgn}(\hat{n} \cdot \hat{r}_i) \sqrt{1 - \left(\frac{n_1}{n_2} \right)^2 |\hat{n} \times \hat{r}_i|^2} - \frac{n_1}{n_2} (\hat{n} \cdot \hat{r}_i) \right] \hat{n}.\end{aligned}\quad (2.106)$$

Problem 2.25

For a parabolic refracting surface of the form (2.84)

$$z - z_c = ay^2, \quad (2.107)$$

with index n_1 to the left of the surface and n_2 to the right show that the effective focal length can be written

$$f_{\text{eff}} = \frac{n_2}{2a(n_2 - n_1)}. \quad (2.108)$$

Do this by considering incident, paraxial rays from the left, parallel to the optical axis, and where they focus by crossing the optical axis. In doing this, you will establish Eq. (2.86).

Problem 2.26

Suppose you have an optical system, and you want to find its effective focal length, which we define as the point where rays parallel and close to the optical axis cross the axis.

One way to do this with ray matrices is to propagate a parallel ray through the optical system (represented by \mathbf{M}), and then over a free-space distance d :

$$\begin{bmatrix} y_{\text{out}} \\ y'_{\text{out}} \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \mathbf{M} \begin{bmatrix} y_{\text{in}} \\ y'_{\text{in}} \end{bmatrix} \quad (2.109)$$

Then we simply find the d such that $y_{\text{out}} = 0$. We can regard this as a root-finding problem, $y_{\text{out}}(d) = 0$, and we can use a root-finder on the computer to do solve this. Note that in the paraxial approximation, this works for *any* input ray that is parallel to the optical axis, so we can choose y_{in} to be any (nonzero) number.

Note also that this is a rather *trivial* root-finding problem: given the output ray from the optical system \mathbf{M} , it is not a difficult trigonometry exercise to figure out the crossing point. However, the point of this exercise is to get used to using the root finder and multiplying together ray matrices in Octave, as a prelude for the following problem on computational ray-tracing.

The goal of this exercise is for you to write an Octave script to carry out the above calculation for a ball lens of diameter $D = 2$ mm and refractive index $n = 1.77$ (appropriate as an average index of sapphire around 500 nm). Reference your calculated focal length to the center of the ball, and compare your numerical answer to your analytic expression in Problem 2.15.

To get you started, study the example script below, which uses the same method to find the focal length of a thin lens of focal length f (which, as you might imagine, should turn out to be f). Then modify it to solve the ball-lens case. (To run it: start in the same directory as the script file, start Octave, and type the name of the script into the Octave shell.) You can download this script at

http://atomoptics.uoregon.edu/~dstreck/teaching/octave/thin_lens_matrix_root.m

and the code listing follows.

```
% thin_lens_matrix_root
%
% Trace rays through thin lens using ray matrices, and then find the
% focus via root-finder.

global gF

% parameters
f = 50; % focal length of thin lens in whatever units you like (let's say mm)

% we need to pass a parameters to our function via global variables;
% prepend global names with a 'g' and keep them separate for clarity
gF = f;

% function to return ray height at the end of the imaging system
% inputs:
%   d -> distance of propagation after the lens
% outputs:
%   yout -> height of ray after propagation through lens and distance d
%
function yout = rayfunc(d)

% get and rename global variables
global gF
f = gF;

thin_lens_M = [1, 0; -1/f, 1];
exterior_free_space_M = [1, d; 0, 1];

% note order of multiplication!
% use '...' to continue an expression on the next line
composite_M = ...
    exterior_free_space_M * ...
    thin_lens_M;

% now trace an initially parallel ray, height = 1
yin_vec = [1; 0];
yout_vec = composite_M * yin_vec;

% return the height of the final ray
yout = yout_vec(1);

end %function rayfunc

% find distance where ray crosses the axis, using 'fzero' root finder
% i.e., we want to find the distance d such that rayfunc(d) = 0
% Note that we need to supply an initial guess, say 10f here, just
```

```
% to make it not so trivial.
d_focus = fzero('rayfunc', 10*f);

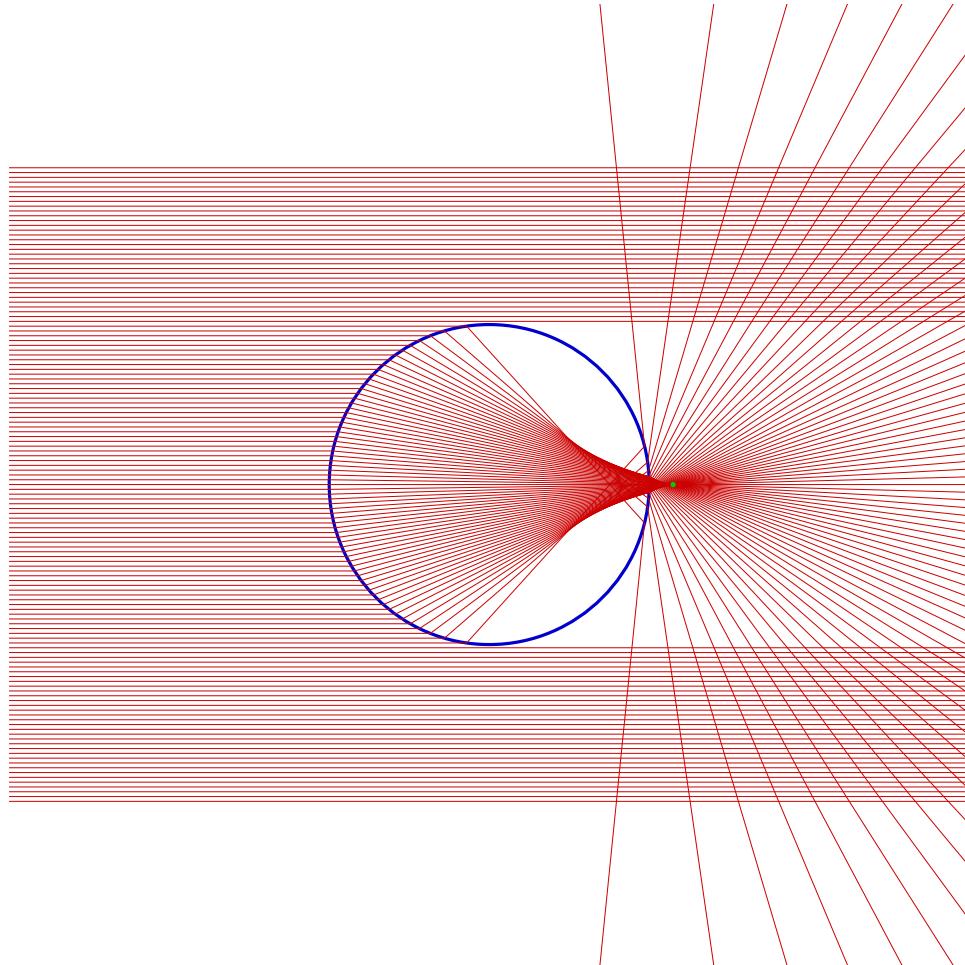
% print result
printf('numerically computed focal length = %f mm\n', d_focus);

% compare to what you know it should be
printf('analytically computed focal length = %f mm\n', f);
```

Problem 2.27

Study the code for refractive ray-tracing for a parabolic interface in the text (Section 2.8.3). Modify this code to trace rays through a ball lens (spherical dielectric in air) of radius 2 mm, and make plots for refractive indices $n = 1.33$ (water), $n = 1.77$ (an average index of sapphire around 500 nm), and $n = 2.42$ (diamond, around 589 nm) of the lens.

As an example of what you should get, the $n = 1.77$ case is shown here.



Some notes:

- You will need to handle refraction at *two* interfaces.
- As it turns out, you will not need to worry about total internal reflection.

- However, as you can see from the example, you will need to handle rays that are refracted into the backwards direction, so study the modifications for reflective ray tracing in the parabola example. Make sure to choose your incident rays to show off some of these backward rays.
- Be sure to mark the paraxial focal point of the lens, using the results of Problem 2.15.

Chapter 3

Fourier Analysis

Fourier analysis, the study of representing functions as sums of **harmonic functions** (sines and cosines), is one of the most important concepts in essentially every area of physics.

3.1 Periodic Functions: Fourier Series

Before getting into defining the Fourier transform, it is helpful to motivate it by first considering the simpler Fourier series for a periodic function (or equivalently, a function defined on a bounded domain). Suppose $f(t)$ is a periodic function with **period** T , so that it satisfies

$$f(t) = f(t + T), \quad (3.1)$$

for all t . The frequency corresponding to the period is given by

$$\omega = 2\pi\nu = \frac{2\pi}{T}, \quad (3.2)$$

where ν is the **frequency** and ω is the **angular frequency**. The distinction between these different frequency conventions can cause confusion when specific numbers are involved. For clarity, ν can be reported in Hz, and ω can be reported in rad/s or as $\omega/2\pi$ (in Hz).

The basic point is that we can think of the harmonic functions (sines and cosines) as being fundamental building blocks for functions. So let's try to build up $f(t)$ out of harmonic functions. Since we know that $f(t)$ is periodic, we will only use those harmonic functions that are also periodic with period T :

$$f(t) = a_0 + 2 \sum_{n=1}^{\infty} a_n \cos(n\omega t) + 2 \sum_{n=1}^{\infty} b_n \sin(n\omega t). \quad (3.3)$$

(Fourier series)

This expansion uses only expansion coefficients for positive n , but we can make things a bit easier to generalize if we define negative ones too:

$$a_{-n} := a_n, \quad b_{-n} := -b_n \quad (3.4)$$

for all nonnegative n (in particular, $b_0 = 0$ by our definition). Then for *any* integer n (positive and negative), we can define the complex Fourier coefficient by

$$c_n := a_n + ib_n. \quad (3.5)$$

This definition allows us to write things a bit more compactly. Let's rewrite the Fourier series:

$$f(t) = a_0 + b_0 + \sum_{n=1}^{\infty} a_n (e^{in\omega t} + e^{-in\omega t}) - i \sum_{n=1}^{\infty} b_n (e^{in\omega t} - e^{-in\omega t}). \quad (3.6)$$

In each term of the form $\exp(-in\omega t)$ we can make the transformation $n \rightarrow -n$ to simplify the sums:

$$f(t) = \sum_{n=-\infty}^{\infty} a_n e^{-in\omega t} + i \sum_{n=-\infty}^{\infty} b_n e^{-in\omega t}. \quad (3.7)$$

Using the complex coefficients, this simplifies greatly:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{-in\omega t}. \quad (3.8)$$

(complex Fourier series)

Now we can see that this Fourier series is a sum over complex harmonic functions with frequency $\omega_n = n\omega$. The convention is that the harmonic function of the form $\exp(-i\omega t)$ corresponds to a “positive” frequency ω , whereas the conjugate function $\exp(i\omega t) = \exp[-i(-\omega)t]$ corresponds to a “negative” frequency $-\omega$. It might sound strange to talk about positive and negative frequencies, but for a real function $f(t)$ the positive frequency “contributes as much as” its negative counterpart in the sense that the coefficients must obey the constraint $c_n = c_{-n}^*$. However, the series (3.8) is more general than the original series (3.3) in that the second series can also represent complex-valued functions.

Of course, given the c_n coefficients, we can obtain the coefficients in the original Fourier series (3.3):

$$\begin{aligned} a_n &= \frac{1}{2}(c_n + c_n^*) = \frac{1}{2}(c_n + c_{-n}) \\ b_n &= \frac{1}{2i}(c_n - c_n^*) = \frac{1}{2i}(c_n - c_{-n}). \end{aligned}$$

(real and complex coefficients related) (3.9)

But how do we obtain the c_n coefficients in the first place? First, observe that for some integer n' , we can evaluate the integral

$$\begin{aligned} \int_0^T e^{-in\omega t} e^{in'\omega t} dt &= \int_0^T e^{-i(n-n')\omega t} dt \\ &= \frac{1}{\omega} \int_0^{2\pi} e^{-i(n-n')x} dx \\ &= \frac{1}{-i\omega} \left[\frac{e^{-i(n-n')x}}{n-n'} \right]_0^{2\pi} \\ &= 0 \quad \text{if } n \neq n', \end{aligned} \quad (3.10)$$

where we defined $x := \omega t$. We need to be more careful with the $n = n'$ case, though, since there is a removable singularity here. If we define $s = n - n'$, then

$$\lim_{s \rightarrow 0} \left(\frac{e^{-i2\pi s} - 1}{s} \right) = \lim_{s \rightarrow 0} \left(\frac{[1 - i2\pi s + O(s^2)] - 1}{s} \right) = -2\pi i, \quad (3.11)$$

so that the integral (3.10) takes the value $2\pi/\omega = T$ if $n = n'$. We can rewrite this relation in the more meaningful form

$$\frac{1}{T} \int_0^T \left(e^{-in'\omega t} \right)^* e^{-in\omega t} dt = \delta_{nn'},$$

(orthonormality of harmonic functions) (3.12)

where $\delta_{nn'}$ is the Kronecker delta ($\delta_{nn'} = 1$ if $n = n'$ and 0 otherwise). This is the **orthogonality relation** for the harmonic functions. The harmonic functions are basis vectors in a vector space of functions, and the orthogonality relation is a special case of the inner product defined by the same integral:

$$\langle f_1, f_2 \rangle := \frac{1}{T} \int_0^T f_1^*(t) f_2(t) dt. \quad (3.13)$$

(inner product of periodic functions)

So we are still doing linear algebra, but the infinite-dimensional, continuous version instead of the finite, discrete version with matrices that we reviewed in the beginning.

Now consider the inner product of a basis vector with $f(t)$:

$$\langle e^{-in'\omega t}, f \rangle = \frac{1}{T} \int_0^T (e^{-in'\omega t})^* f(t) dt = \sum_{n=-\infty}^{\infty} \frac{c_n}{T} \int_0^T (e^{-in'\omega t})^* e^{-in\omega t} dt = \sum_{n=-\infty}^{\infty} c_n \delta_{nn'} = c_{n'}. \quad (3.14)$$

Thus, we can use the orthonormal properties of basis functions to project out the coefficients of $\exp(-in\omega t)$:

$$c_n = \frac{1}{T} \int_0^T e^{in\omega t} f(t) dt. \quad (3.15)$$

(complex Fourier coefficient)

Of course, as we noted above, we can also now calculate the a_n and b_n coefficients in terms of the c_n .

It can be shown for “reasonable” functions that

$$\sum_{n=-N}^N c_n e^{-in\omega t} \longrightarrow f(t) \text{ as } N \longrightarrow \infty \quad (3.16)$$

at each point t except possibly on a set of zero measure, and that the set of coefficients c_n that accomplish this is uniquely defined.

3.1.1 Example: Rectified Sine Wave

As a simple example, let’s compute the Fourier series for the function $|\sin \omega t|$. Note that because of the “rectification,” this function has a period of π/ω , so the effective frequency of this function is 2ω . Thus, letting $T = \pi/\omega$,

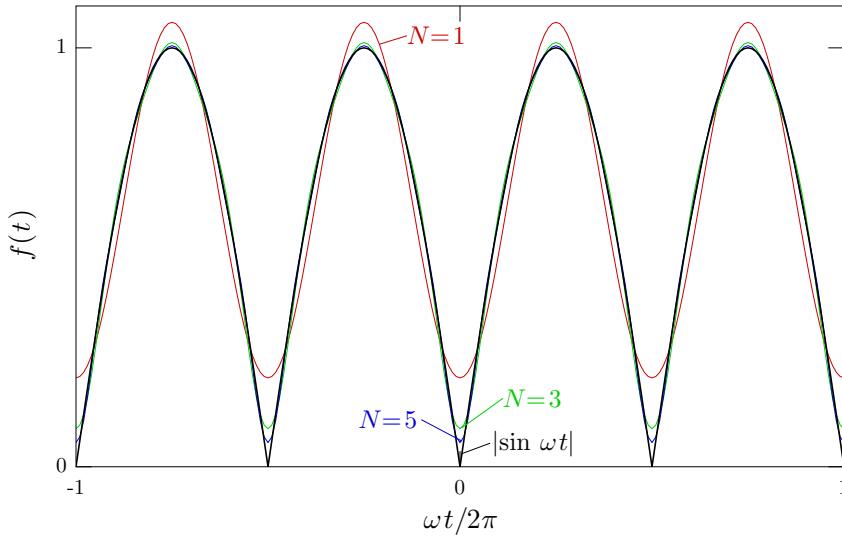
$$\begin{aligned} c_n &= \frac{1}{T} \int_0^T |\sin \omega t| e^{in(2\omega)t} dt \\ &= \frac{1}{T} \int_0^T \frac{1}{2i} (e^{i\omega t} - e^{-i\omega t}) e^{in(2\omega)t} dt \\ &= \frac{1}{2iT} \int_0^T (e^{i(2n+1)\omega t} - e^{i(2n-1)\omega t}) dt \\ &= \frac{1}{2\pi i} \int_0^\pi (e^{i(2n+1)x} - e^{i(2n-1)x}) dx \\ &= \frac{1}{2\pi i} \left[\frac{e^{i(2n+1)\pi} - 1}{i(2n+1)} - \frac{e^{i(2n-1)\pi} - 1}{i(2n-1)} \right] \\ &= \frac{1}{\pi(2n+1)} - \frac{1}{\pi(2n-1)} = \frac{2}{\pi(1-4n^2)}. \end{aligned} \quad (3.17)$$

Thus, we can write the series as

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{2}{\pi(1-4n^2)} e^{-i2n\omega t}. \quad (3.18)$$

As reflected in the initial setup of this problem, the *rectified* sine wave contains only the *even* harmonics of the original pure harmonic wave, and none of the initial frequency. This is a useful feature to keep in mind, for example, when designing a device to double the frequency of an input signal. Note that $c_n = c_{-n} = c_n^*$ in this example because $f(t)$ is real and even, whereas in the more general case of a real function we would only expect the less restrictive case $c_n = c_{-n}^*$.

As in the plot here, the Fourier series converges fairly rapidly as a function of the number N of frequencies included, even already for $N = 5$ terms (beyond the dc component).



For more terms, such as $N = 20$, the sum is visually indistinguishable from the original function. Other functions, particularly those with discontinuities, have series that do not converge so well as this example.

3.2 Aperiodic Functions: Fourier Transform

We can also use the same harmonic functions $e^{-i\omega t}$ to build up aperiodic functions. Recall that if T is the period, then we were using a discrete (countable) set of harmonic functions with a frequency spacing given by $\Delta\omega = 2\pi/T$. An aperiodic function corresponds to $T \rightarrow \infty$, so we must use a representation where $\Delta\omega \rightarrow 0$. Thus, we need a *continuous* spectrum to represent an aperiodic function, since there is much more “information” in the function than in the periodic case.

So let’s define the Fourier transform as a generalization of the Fourier series, but with a slightly different normalization:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{-in\omega t} \longrightarrow f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(\omega) e^{-i\omega t} d\omega$$

(generalization to inverse Fourier transform) (3.19)

Thus, $\tilde{f}(\omega)/2\pi$ is the amplitude of the frequency component $e^{-i\omega t}$ (the normalization coefficient depends on how we define the *density* of the continuum of basis functions, as we will see shortly). The other usual nomenclature is that $\tilde{f}(\omega)$ is the **Fourier transform** of $f(t)$. The above mathematical operation is the **inverse Fourier transform**, since we are finding $f(t)$ from its Fourier transform.

In the same way as before, we can do a projection to find the amplitudes (Fourier transform). The analogue of the Fourier-series projection is

$$c_n = \frac{1}{T} \int_0^T e^{in\omega t} f(t) dt \longrightarrow \tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt.$$

(generalization to Fourier transform) (3.20)

The pair of equations (3.19) and (3.20) is one of the most important tools in physics, so they deserve to be written again:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(\omega) e^{-i\omega t} d\omega, \quad \tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt.$$

(Fourier and inverse Fourier transforms) (3.21)

The functions $f(t)$ and $\tilde{f}(\omega)$ are said to be a **Fourier transform pair**, and their relationship is of fundamental importance in understanding linear systems in physics.

Again, if $f(t)$ is a real function, then the Fourier transform satisfies $\tilde{f}(\omega) = \tilde{f}^*(-\omega)$, in which case

$$f(t) = 2\operatorname{Re} \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(\omega) e^{-i\omega t} d\omega \right\}, \quad (3.22)$$

so that the positive and negative frequencies contribute equally in amplitude to form a real-valued function.

Note that there is an important alternate convention, where instead of ω we can use the standard frequency $\nu = \omega/2\pi$. This changes the density of the basis functions in a way that makes the normalizations more symmetric:

$$f(t) = \int_{-\infty}^{\infty} \bar{f}(\nu) e^{-i2\pi\nu t} d\nu, \quad \bar{f}(\nu) = \int_{-\infty}^{\infty} f(t) e^{i2\pi\nu t} dt.$$

(Fourier and inverse Fourier transforms, ν - t convention) (3.23)

Here, we have defined an alternate Fourier transform by $\bar{f}(\nu) := \tilde{f}(\omega/2\pi)$. It is a bit easier to remember this form to work out where the $(1/2\pi)$ goes in the ω form of the transform equations.

We haven't been too concerned with rigor here, but it is interesting to ask, when is it possible to have a Fourier transform? As with the Fourier series, the $f(t)$ must be a "reasonable" function for $\tilde{f}(\omega)$ to exist. If $f(t)$ is defined on the real line, then one possible set of sufficient conditions is as follows:

1. $\int_{-\infty}^{\infty} f(t) dt$ exists.
2. f has only a finite number of discontinuities and a finite number of maxima and minima in any finite interval.
3. f has no infinite discontinuities.

Naturally, functions useful in physics, e.g., to model wave phenomena, tend to be reasonable in precisely this sense.

Finally, let's reemphasize the connection to linear algebra: the Fourier transform is a linear transformation between vector spaces of *functions*. If we denote the Fourier transform by the symbol " \mathcal{F} ," we can write the above definitions in the compact form

$$\tilde{f}(\omega) = \mathcal{F}[f(t)], \quad f(t) = \mathcal{F}^{-1}[\tilde{f}(\omega)]. \quad (3.24)$$

Then again, linearity of the Fourier transform means that

$$\mathcal{F}[\alpha f(t) + \beta g(t)] = \alpha \tilde{f}(\omega) + \beta \tilde{g}(\omega) \text{ iff } \mathcal{F}[f(t)] = \tilde{f}(\omega), \mathcal{F}[g(t)] = \tilde{g}(\omega) \quad (3.25)$$

for all $\alpha, \beta \in \mathbb{R}$. Again, this is like a matrix transformation, but in the infinite, continuous limit.

3.2.1 Example: Fourier Transform of a Gaussian Pulse

One of the most useful Fourier transforms that can be easily calculated is of the Gaussian pulse,

$$f(t) = Ae^{-\alpha t^2}. \quad (3.26)$$

Using the Fourier transform definition (3.20),

$$\tilde{f}(\omega) = \int_{-\infty}^{\infty} Ae^{-\alpha t^2 + i\omega t} dt. \quad (3.27)$$

To evaluate the integral, we need to perform a mathematical trick, *completing the square* in the exponent. That is, let's rewrite the argument in the exponent in the form $a(t-b)^2 + c = at^2 - 2abt + c - ab^2$. Equating powers of t , we find the new coefficients:

$$\begin{aligned} t^2: \quad & a = \alpha \\ t^1: \quad & -2ab = i\omega \implies b = \frac{-i\omega}{2\alpha} \\ t^0: \quad & c - ab^2 = 0 \implies c = -\frac{\omega^2}{4\alpha}. \end{aligned} \quad (3.28)$$

Substituting the new form of the exponent and letting $t \rightarrow t + b$,

$$\tilde{f}(\omega) = \int_{-\infty}^{\infty} Ae^{-at^2+c} dt = Ae^c \sqrt{\frac{\pi}{a}} = A\sqrt{\frac{\pi}{\alpha}} \exp\left(-\frac{\omega^2}{4\alpha}\right), \quad (3.29)$$

where we evaluated the integral by comparison to the standard normalized form of the Gaussian (you should memorize this, by the way),

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = 1, \quad (3.30) \quad (\text{Gaussian normalization})$$

by taking the **standard deviation** σ to be $1/\sqrt{2\alpha}$. Hence we see that **the Fourier transform of a Gaussian is a Gaussian**.

Notice that while the standard deviation of the original Gaussian is

$$\sigma_t = \frac{1}{\sqrt{2\alpha}}, \quad (3.31)$$

the standard deviation of the Fourier transform is

$$\sigma_{\omega} = \sqrt{2\alpha}. \quad (3.32)$$

Since the standard deviation is a measure of the width of a function (more precisely the square root of the variance of the function), we can see that as a increases, the width of the original pulse decreases, but the width of the transform increases. This reflects a general property of Fourier transforms, that

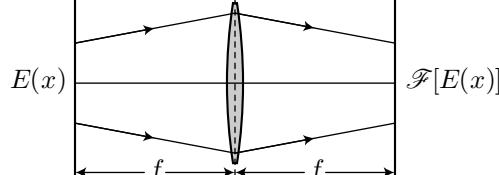
$$\delta t \sim \frac{1}{\delta \omega}, \quad (3.33) \quad (\text{uncertainty relation})$$

where δt is the “width” of $f(t)$ and $\delta \omega$ is the width of $\tilde{f}(\omega)$ (for root-mean-square widths and Gaussian functions, there is a strict equality here). This is precisely the same **uncertainty principle** that is a fundamental principle in quantum mechanics.

3.3 The Fourier Transform in Optics

We will spend essentially the rest of the course using the Fourier transform to understand wave optics. But first let's review a few of the physically important Fourier transforms that you should know, and briefly discuss examples of situations where they come up.

The major use of the Fourier transform, as we will see, is in **Fourier optics**. The central effect behind Fourier optics is that within the paraxial approximation, a thin lens acts as a Fourier-transform computer.



That is, given a scalar electric field $E(x)$ one focal length before the lens, the field after the lens is related to $\mathcal{F}[E(x)]$.

But Fourier transforms show up everywhere in optics as well as in the rest of physics. Some of the most important and useful ones are summarized here. **You should memorize these.**

1. The Fourier transform of a Gaussian is a Gaussian.

$$\mathcal{F}\left[e^{-t^2/2}\right] = e^{-\omega^2/2} \quad (3.34) \quad (\text{Fourier transform of Gaussian})$$

Recall that the spherical mirrors of resonators act as lenses. We will see that Gaussian beams are electromagnetic field modes of a spherical-mirror resonator. A simple way to see this is that the modes must be Fourier transform of itself in order to resonate (repeat itself) in the cavity. The Gaussian function is the most localized function that has this property.

2. The Fourier transform of an exponential is a Lorentzian.

$$\mathcal{F}[e^{-|t|}] = \frac{2}{1+\omega^2} \quad (3.35) \quad (\text{Fourier transform of exponential})$$

Atoms decay exponentially due to spontaneous emission. If N_e is the number of atoms in the excited state, the decay law is $N_e(t) = N_e(0) \exp(-\Gamma t)$. It turns out that the Fourier transform of the time dependence of the emission gives the radiation spectrum for spontaneous emission. For atoms obeying the exponential decay law, the emission spectrum is Lorentzian.

3. The Fourier transform of a square pulse is a sinc function.

$$\mathcal{F}[\text{rect}(t)] = \text{sinc}(\omega/2) := \sin(\omega/2)/(\omega/2) \quad (\text{Fourier transform of square pulse}) \quad (3.36)$$

The rectangular function is defined by

$$\text{rect}(t) := \begin{cases} 1 & \text{if } |t| < 1/2 \\ 1/2 & \text{if } |t| = 1/2 \\ 0 & \text{if } |t| > 1/2. \end{cases} \quad (3.37)$$

The far-field diffraction pattern of a uniformly illuminated slit is a sinc function for this reason, a fact that is related to the Fourier-transform property of a lens as mentioned above.

4. The Fourier transform of a constant is a delta function.

$$\mathcal{F}\left[\frac{1}{2\pi}\right] = \delta(\omega) \quad (3.38) \quad (\text{Fourier transform of constant})$$

We haven't yet defined the delta function but we will do so shortly. For now it suffices to think of the delta function $\delta(t)$ as a unit-area pulse in the limit as δt tends to zero. In other words, this relation is the extreme limit of the uncertainty principle.

5. The Fourier transform of a delta function is a constant.

$$\mathcal{F}[\delta(t)] = 1 \quad (3.39) \quad (\text{Fourier transform of delta function})$$

This is the opposite extreme of the uncertainty principle: an arbitrarily short pulse has an arbitrarily broad spectrum.

3.4 Delta Function

As we just mentioned, a delta function is an idealized limit of a very short pulse. But this is a sloppy definition. To do this in a mathematically well-defined way, we need to consider a sequence of functions $h_n(t)$ that have the following properties:

1. The h_n are "reasonable" (e.g., simply peaked around $t = 0$).
2. The width of h_n converges to zero as $n \rightarrow \infty$.
3. The h_n are normalized: $\int_{-\infty}^{\infty} h_n(t) dt = 1$ for all n .

For example, we can use Gaussian functions. In normalized form, we can write

$$h_n(t) = \frac{1}{\sqrt{2\pi\sigma_n}} \exp\left(-\frac{t^2}{2\sigma_n^2}\right), \quad (3.40)$$

and if we choose $\sigma_n = 1/\sqrt{2\pi n}$, the functions become

$$h_n(t) = n \exp(-\pi n^2 t^2). \quad (3.41)$$

Then consider the integral

$$\int_{-\infty}^{\infty} h_n(t) f(t) dt, \quad (3.42)$$

for an arbitrary test function $f(t)$, which we again assume to be “reasonable.” We will define the **delta function** $\delta(t)$ such that

$$\int_{-\infty}^{\infty} \delta(t) f(t) dt := \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} h_n(t) f(t) dt, \quad (\text{definition of delta function}) \quad (3.43)$$

or in inner-product notation,

$$\langle \delta, f \rangle := \lim_{n \rightarrow \infty} \langle \delta, h_n \rangle \quad (\text{definition of delta function}) \quad (3.44)$$

Notice that the order of the integral sign and the limit are important; i.e., the integral and the limit *don't commute*. This is because $\lim_{n \rightarrow \infty} h_n(t)$ does not exist. Thus, **the delta function only makes sense as part of the argument of an integral**. But as physicists, we often write

$$\delta(t) = \lim_{n \rightarrow \infty} h_n(t) \quad (\text{sloppy!}) \quad (\text{“definition” of delta function}) \quad (3.45)$$

as a shorthand for the above (more careful) definition.

Similarly, note that $\lim_{n \rightarrow \infty} h_n(t) = 0$ for all $t \neq 0$, so it is useful to think of $\delta(t)$ as a “unit-area function that is zero everywhere but ∞ at $t = 0$,” as long as you are careful about it.

Now we will review and derive a few properties of the delta function. The list that follows is by no means exhaustive.

1. $\delta(t)$ is **normalized**:

$$\int_{-\infty}^{\infty} \delta(t) dt = 1 \quad (\text{normalization of delta function}) \quad (3.46)$$

Proof. Since the functions $h_n(t)$ are normalized,

$$\int_{-\infty}^{\infty} \delta(t) dt = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} h_n(t) dt = 1. \quad (3.47)$$

2. **Projection property** of $\delta(t)$:

$$\int_{-\infty}^{\infty} \delta(t) f(t) dt = f(0). \quad (\text{projection property of delta function}) \quad (3.48)$$

Proof. Assuming f has a Taylor expansion about $t = 0$,

$$f(t) = f(0) + f'(0)t + \frac{1}{2}f''(0)t^2 + \dots \quad (3.49)$$

Noting that

$$\int_{-\infty}^{\infty} h_n(t) t^m dt = n \int_{-\infty}^{\infty} t^m \exp(-\pi n^2 t^2) dt = \frac{1 + (-1)^m}{2} \Gamma\left(\frac{1+m}{2}\right) \pi^{-(m+1)/2} n^{-m} \quad (3.50)$$

for the Gaussian $h_n(t)$'s defined above, we can write

$$\int_{-\infty}^{\infty} h_n(t) f(t) dt = f(0) + \frac{f''(0)}{2\pi n^2} + O\left(\frac{1}{n^4}\right). \quad (3.51)$$

As $n \rightarrow \infty$, all the higher order terms vanish, leaving just $f(0)$.

3. **Shifted projection property** of $\delta(t)$:

$$\int_{-\infty}^{\infty} \delta(t-a) f(t) dt = f(a).$$

(shifted projection property of delta function) (3.52)

This can be proved by letting $t \rightarrow t+a$ in the integral, and then using Eq. (3.48) to evaluate the result.

4. **Fourier transform** of $\delta(t)$:

$$\mathcal{F}[\delta(t-a)] = e^{i\omega a}. \quad (3.53)$$

(Fourier transform of delta function)

The proof of this follows trivially from the previous property and the definition (3.20) of the Fourier transform.

5. **Integral representation** of $\delta(t)$:

$$\delta(t-a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega(t-a)} d\omega.$$

(integral representation of delta function) (3.54)

This important representation of the delta function follows from the inversion of the previous property,

$$\mathcal{F}^{-1}[e^{i\omega a}] = \delta(t-a), \quad (3.55)$$

and the application of the inverse transform integral (3.19).

This final property above, the integral representation (3.54) of $\delta(t)$ is useful, for example, for showing the orthogonality of the harmonic basis functions. As in the periodic case, we can define two suitable functions $f(t)$ and $g(t)$ to be **orthogonal** if

$$\int_{-\infty}^{\infty} f^*(t) g(t) dt = 0, \quad (3.56)$$

(orthogonality of functions)

which is just $\langle f, g \rangle = 0$ in inner-product notation. For harmonic functions, $\exp(-i\omega t)$ and $\exp(-i\omega' t)$ are orthogonal because

$$\int_{-\infty}^{\infty} \exp(i\omega t) \exp(-i\omega' t) dt = 2\pi\delta(\omega - \omega'),$$

(orthogonality of harmonic functions) (3.57)

which is zero if $\omega \neq \omega'$. Also, since $\delta(t)$ is normalized, we see that in the form $\exp(-i\omega t)/2\pi$ [or alternately, $\exp(i2\pi\nu t)$], the harmonic functions can form an orthonormal set.

Subject to the conditions on $h_n(t)$ above, the delta function (and the properties we have derived) are *independent* of the particular choice of $h_n(t)$. However, sometimes, some ambiguous integrals of delta functions sometimes crop up, such as

$$\int_0^{\infty} dt \delta(t) = ?, \quad (3.58)$$

which has no well-defined value. Depending on the functions $h_n(t)$ used to define $\delta(t)$, this function can take on essentially any value (though more reasonably, one could expect it to be between zero and unity). This is one case where the more rigorous definition helps in physics; the physical model leading to the delta function defines the choice of $h_n(t)$, and thus the value of this integral.

3.5 Exercises

Problem 3.1

Compute the Fourier transform $\tilde{f}(\omega)$ of the rectangular-pulse function

$$f(t) = \begin{cases} 1, & |t| < a/2 \\ 0 & \text{elsewhere.} \end{cases} \quad (3.59)$$

Show from your answer that as the pulse width a decreases, the width of $\tilde{f}(\omega)$ increases.

Problem 3.2

Suppose that $f(t)$ is a real, odd [$f(-t) = -f(t)$], periodic function; what constraints are satisfied by the Fourier coefficients c_n for this function?

Problem 3.3

If c_n is the n th Fourier coefficient for the periodic function $f(t)$, write down an expression for the n th Fourier coefficient for $f(t + \alpha)$ in terms of c_n .

Problem 3.4

(a) Compute the Fourier series for a square wave of unit period, given by

$$f(t) = \operatorname{sgn}[\sin(2\pi t)] \quad (3.60)$$

where $\operatorname{sgn}(x)$ is the sign of x , i.e.,

$$\operatorname{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases} \quad (3.61)$$

(b) Make plots of the truncated Fourier series,

$$f_N(t) = \sum_{n=-N}^N c_n e^{-in\omega t}, \quad (3.62)$$

along with the exact function for $N = 1, 3, 5, 21$, and 101 to check how good the finite approximations are. You should notice a ringing behavior near the discontinuities, where the approximant $f_N(t)$ overshoots the actual function $f(t)$. This is called the *Gibbs phenomenon*, and persists at an essentially constant (nonzero) value even for arbitrarily large N . From your plots, estimate the magnitude of the Gibbs overshoot.

(c) We said that the functions $f_N(t)$ converge to $f(t)$. But the Gibbs phenomenon says that for any N , there is always some t such that $f_N(t)$ is at least some minimum distance away from $f(t)$. Explain how these two statements are consistent.

Problem 3.5

One application of Fourier series is in synthesizing waveforms or images with relatively little information. For example, the JPEG image compression algorithm achieves much of its compression by removing harmonic components, and (analog) electronic organs can use relatively few harmonics (on the order of 10) to make a passable imitation of a musical instrument.¹

A common feature of truncated Fourier series is that relatively few terms are needed to make a fairly good approximant, but for a high accuracy approximant, many terms may be necessary. Mathematically, the initial convergence of the Fourier series is rapid, but the convergence rate becomes very slow

¹A nice discussion of this example appears in T. W. Körner, *Fourier Analysis* (Cambridge, 1988).

by the time many terms have accumulated. (This is even the case for continuous functions, where the Gibbs phenomenon is absent.)

As a simple model, suppose you want to synthesize a triangle wave,

$$f(t) := \begin{cases} 1 - 4|t|/T, & |t| < T/2 \\ f(t+T) = f(t-T) & \text{elsewhere,} \end{cases} \quad (3.63)$$

using as few harmonics as possible.

(a) Compute the Fourier series for the triangle wave as defined above.

(b) Consider the partial sum,

$$f_N(t) = \sum_{n=-N}^N c_n e^{-in\omega t}, \quad (3.64)$$

and let's define the *fractional* error of the partial sum to be normalized root-mean-square error:

$$\varepsilon_N := \frac{\sqrt{\int_0^T |f_N(t) - f(t)|^2 dt}}{\sqrt{\int_0^T |f(t)|^2 dt}}. \quad (3.65)$$

How large should N be in the partial sum to obtain an accuracy of 10%? 1%? 0.1%? 0.01%? (A computer may be of great help in solving this problem.)

Problem 3.6

Let $f(t)$ be a function with Fourier transform $\tilde{f}(\omega)$. Show that the Fourier transform of $f(t) \cos(\alpha t)$ is $[\tilde{f}(\omega + \alpha) + \tilde{f}(\omega - \alpha)]/2$.

Problem 3.7

Prove the following form of the Poisson sum rule:

$$\sum_{n=-\infty}^{\infty} \delta(t - n) = \sum_{n=-\infty}^{\infty} \cos(2\pi nt). \quad (3.66)$$

Problem 3.8

Let $\tilde{f}(\omega)$ be the Fourier transform of $f(t)$. Show that the Fourier transform of $g(t) := f(t) \cos^2(\alpha t)$ is

$$\tilde{g}(\omega) = \frac{1}{2}\tilde{f}(\omega) + \frac{1}{4} [\tilde{f}(\omega + 2\alpha) + \tilde{f}(\omega - 2\alpha)]. \quad (3.67)$$

Problem 3.9

Show that if $f(t)$ and $\tilde{f}(\omega)$ form a Fourier transform pair, then so do $f(\alpha t)$ and $\tilde{f}(\omega/\alpha)/|\alpha|$ (for $\alpha \neq 0$).

Problem 3.10

(a) Compute the Fourier transform of $f(t) = U_H(t) e^{-at}$, where $U_H(t)$ is the Heaviside step function [i.e., $U_H(t > 0) = 1$ and $U_H(t < 0) = 0$], and $a > 0$.

(b) Use your result from (a) to compute the Fourier transform of $g(t) = U_H(t) te^{-at}$ for $a > 0$. (Try to do this without any integration other than what you have already done.)

Chapter 4

Review of Electromagnetic Theory

4.1 Maxwell Equations in Vacuum

Electromagnetism is the fundamental theory that underlies all of wave optics, so before proceeding with this more general theory of optics, we will stop and review the basics. **Maxwell's equations** form the basis of electromagnetism. We will write Maxwell's equations in free space as

$$\begin{aligned}\nabla \cdot \mathbf{E} &= 0 && \text{(Gauss' law in free space)} \\ \nabla \cdot \mathbf{H} &= 0 && \text{(no magnetic monopoles)} \\ \nabla \times \mathbf{E} &= -\mu_0 \frac{\partial \mathbf{H}}{\partial t} && \text{(Faraday's law)} \\ \nabla \times \mathbf{H} &= \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} && \text{(Ampère's law with Maxwell's correction in free space)}\end{aligned}\quad (\text{Maxwell's equations}) \quad (4.1)$$

where $\mathbf{E}(\mathbf{r}, t)$ is the **electric field**, $\mathbf{H}(\mathbf{r}, t)$ is the **magnetic field**,

$$\epsilon_0 \simeq 8.85 \times 10^{-12} \frac{\text{F}}{\text{m}} \quad (4.2)$$

is the **electric permittivity of the vacuum**, and

$$\mu_0 := 4\pi \times 10^{-7} \frac{\text{N}}{\text{A}^2} \quad (\text{magnetic permeability of the vacuum}) \quad (4.3)$$

is the **magnetic permeability of the vacuum**. Also recall the notation

$$\nabla := \hat{x} \frac{\partial}{\partial x} + \hat{y} \frac{\partial}{\partial y} + \hat{z} \frac{\partial}{\partial z} \quad (4.4)$$

for the vector derivative operator that we used in writing down the Maxwell equations.

Actually, we won't use Maxwell's equations very often in the form we wrote. Rather, we will use them to derive the **wave equation**, the fundamental equation of optics. To do this, we first apply $(\nabla \times)$ to Faraday's law:

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu_0 \frac{\partial}{\partial t} (\nabla \times \mathbf{H}). \quad (4.5)$$

Now we will use the vector identity

$$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}, \quad (4.6)$$

which holds for any vector \mathbf{A} to evaluate the left-hand side of this equation, and we will use Ampère's law to evaluate the right-hand side:

$$\nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\mu_0 \frac{\partial}{\partial t} \left(\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right). \quad (4.7)$$

The first term vanishes, as we can see from Gauss' law. Then cleaning things up a bit, we arrive at the standard form of the wave equation:

$$\nabla^2 \mathbf{E} - \frac{1}{c_0^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \quad (4.8)$$

(wave equation)

Here,

$$c_0 = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \quad (4.9)$$

(speed of light in vacuum)

is the **speed of light in vacuum**. This is actually now a defined quantity, given by

$$c_0 := 2.997\,924\,58 \times 10^8 \text{ m/s.} \quad (4.10)$$

(definition of speed of light in vacuum)

Since μ_0 and c_0 are both exactly defined quantities, the exact value of ϵ_0 is defined by

$$\epsilon_0 := \frac{1}{\mu_0 c_0^2}. \quad (4.11)$$

(electric permittivity of the vacuum)

Finally, we note that the magnetic field \mathbf{H} satisfies the same wave equation as the electric field.

4.2 Intensity

One of the most important features of electromagnetic waves (time-dependent solutions of the Maxwell or wave equations) is that they transport energy. The flow of energy is described by the **Poynting vector**, defined by

$$\mathbf{S} := \mathbf{E} \times \mathbf{H}. \quad (4.12)$$

(Poynting vector)

The Poynting vector points in the direction of energy flow; note that it is orthogonal to both the electric and magnetic field vectors. Energy transport and the Poynting vector are explored further in Problem 4.4.

More commonly, energy transport is quantified by the **intensity** of the electromagnetic wave. The intensity is defined as simply the magnitude of the time-averaged Poynting vector,

$$I := |\langle \mathbf{S} \rangle|, \quad (4.13)$$

(intensity)

where the average is taken over time scales long compared to an optical cycle (a few fs) but short compared to other time scales of interest (the angle brackets here denote the time average). Physically, the intensity is a measure of what a realistic optical detector would register in response to the field. Detectors can register time variation of the field energy in up into the MHz or GHz range, but not optical frequencies (which are in the range of hundreds of THz).

4.3 Maxwell Equations in Media

Maxwell's equations take on a slightly different form for electromagnetic fields in a medium:

$$\begin{aligned}\nabla \cdot \mathbf{D} &= 0 \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t}.\end{aligned}$$

(Maxwell's equations in magnetodielectric media) (4.14)

Here, \mathbf{D} is the **electric flux density** or **electric displacement**, and \mathbf{B} is the **magnetic flux density**. These equations are still valid only in a source-free region (i.e., free of charges and currents).

The relation between the \mathbf{E} and \mathbf{D} fields is determined by the electric properties of the medium:

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}. \quad (4.15) \quad (\text{electric flux density in terms of polarization})$$

Here, \mathbf{P} is the **polarization density**, defined to be the electric dipole moment per unit volume. Usually the polarization is induced by the electric field, though some materials can be fabricated with permanent polarization (**electrets**).

Similarly the relation between the \mathbf{H} and \mathbf{B} fields is related to the magnetic properties of the medium:

$$\mathbf{B} = \mu_0 \mathbf{H} + \mathbf{M}.$$

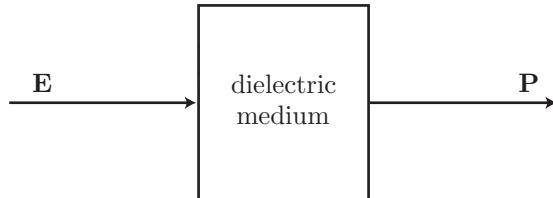
(magnetic flux density in terms of magnetization) (4.16)

Here, \mathbf{M} is the **magnetization density**, defined as the magnetic dipole moment per unit volume. Sometimes the magnetization is induced by an external magnetic field, but permanent magnets are relatively commonplace.

In free space, $\mathbf{P} = \mathbf{M} = 0$, so $\mathbf{D} = \epsilon_0 \mathbf{E}$ and $\mathbf{B} = \mu_0 \mathbf{H}$, in which case we recover the free-space Maxwell equations (4.1) from Eqs. (4.14). In optics, we usually deal with **dielectric** materials. Thus, from now on we will assume $\mathbf{M} = 0$, so that $\mathbf{B} = \mu_0 \mathbf{H}$.

4.4 Simple Dielectric Media

We will now discuss dielectric media in more detail, but here we will only treat the simplest possible such media. For our purposes here it will be useful to think of the electric field \mathbf{E} as something *externally applied* to the medium, while the polarization \mathbf{P} is the *response* of the medium to the field.



Thus, we can think of the polarization as a function of the electric field, as well as space and time: $\mathbf{P}(\mathbf{r}, t) = \mathbf{P}(\mathbf{E}(\mathbf{r}, t); \mathbf{r}, t)$.

The simple dielectric medium has a number of constraints on its properties, which greatly simplify things.

1. **Linearity:** $|\mathbf{P}| \propto |\mathbf{E}|$, so that the superposition principle still holds in the medium.

2. **Nondispersivity:** the speed of light does not depend on the frequency of the electromagnetic waves. In the time domain, this means that there is no hysteresis, since it amounts to saying that $\mathbf{P}(t)$ is a function of only $\mathbf{E}(t)$, never $\mathbf{E}(t')$ for $t' < t$. This is always an idealization, but sufficient for many purposes.
3. **Homogeneity:** $\mathbf{P}(\mathbf{E})$ is independent the position \mathbf{r} inside the medium.
4. **Isotropy:** $\mathbf{P}(\mathbf{E})$ is independent of the orientation of the medium with respect to the field, so there is no “preferred direction” in the medium. The only way for this to work is for $\mathbf{P} \parallel \mathbf{E}$.
5. **Spatial nondispersivity (locality):** $\mathbf{P}(\mathbf{r})$ is a function of only $\mathbf{E}(\mathbf{r})$, never $\mathbf{E}(\mathbf{r}')$ for $\mathbf{r}' \neq \mathbf{r}$.

To satisfy all these conditions, the simple dielectric has a polarization that can be written as a simple proportionality:

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E}. \quad (4.17) \quad (\text{polarization in terms of susceptibility})$$

The constant χ is the **electric susceptibility**, which completely characterizes the simple medium. Thus, we can rewrite Eq. (4.15) as

$$\mathbf{D} = \epsilon \mathbf{E}, \quad (4.18) \quad (\text{electric flux density in simple dielectric})$$

where

$$\epsilon := \epsilon_0(1 + \chi) \quad (4.19) \quad (\text{electric permittivity of dielectric})$$

is the **electric permittivity** of the medium, and the ratio

$$\frac{\epsilon}{\epsilon_0} = 1 + \chi \quad (4.20) \quad (\text{dielectric constant})$$

is the **dielectric constant**.

If we rederive the wave equation using the Maxwell equations for a dielectric, we find that the equation has the same form as before,

$$\nabla^2 \mathbf{E} - \frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0, \quad (4.21) \quad (\text{wave equation for dielectric medium})$$

(\mathbf{H} also satisfies this equation) except that c_0 is replaced by

$$c := \frac{c_0}{n}, \quad (4.22) \quad (\text{speed of light in dielectric medium})$$

which gives the speed of light in the dielectric medium in terms of the **index of refraction**, defined as the square root of the dielectric constant:

$$n := \sqrt{\frac{\epsilon}{\epsilon_0}}. \quad (4.23) \quad (\text{index of refraction})$$

Thus we see how the dielectric properties of the medium relate to the refractive index that we used in ray optics.

Finally, note that in a simple medium (as well as in vacuum), each component of \mathbf{E} and \mathbf{H} *separately* satisfies the wave equation. Thus, we can write a **scalar wave equation**,

$$\nabla^2 E - \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2} = 0, \quad (4.24) \quad (\text{scalar wave equation})$$

where E now stands for any component of \mathbf{E} or \mathbf{H} . Note that this equation only holds in cases like simple media where the different vector components don't get mixed. For example, you have to be more careful when analyzing a refractive interface, which in general mixes different field components except in a special basis. However, even when the scalar wave equation is incorrect, it can be a useful approximation, and is certainly easier to handle than the full vector case.

4.5 Monochromatic Waves and Complex Notation

In general, any function of the form $E(x, t) = E(x \pm ct)$ is a solution to the scalar wave equation in a homogeneous medium. These solutions represent functions that do not change shape but propagate rightward ($x - ct$) or leftward ($x + ct$) with speed c . There are many possible functions of this form, and to handle more general media, we can study **monochromatic waves**, which have a harmonic time dependence with a single angular frequency ω :

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}) \cos(\omega t + \phi), \quad \mathbf{H}(\mathbf{r}, t) = \mathbf{H}(\mathbf{r}) \cos(\omega t + \phi'). \quad (\text{monochromatic fields}) \quad (4.25)$$

Here, ϕ and ϕ' are constant phase offsets. As in the Fourier analysis case, we can instead use complex harmonic functions and define the **complex monochromatic fields** $\mathbf{E}^{(\pm)}$ by breaking up the real field into its positive- and negative-frequency parts:

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}) \frac{e^{-i\phi}}{2} e^{-i\omega t} + \mathbf{E}(\mathbf{r}) \frac{e^{i\phi}}{2} e^{i\omega t} =: \mathbf{E}^{(+)}(\mathbf{r}) e^{-i\omega t} + \mathbf{E}^{(-)}(\mathbf{r}) e^{i\omega t}. \quad (\text{complex monochromatic fields}) \quad (4.26)$$

Recall that we are defining $\mathbf{E}^{(\pm)}$ to go with $e^{\mp i\omega t}$ since by convention $e^{-i\omega t}$ corresponds to the *positive* frequency ω and $e^{i\omega t} = e^{-i(-\omega)t}$ corresponds to the *negative* frequency $(-\omega)$. Of course, we can do the same for all of the other fields \mathbf{H} , \mathbf{D} , \mathbf{B} , \mathbf{P} . The physical field is just the sum of the positive- and negative-frequency parts. But notice that these parts are complex conjugates, as is required to get a real (physical) field. Thus, we can always write the physical field as $E^{(+)}$ with its conjugate:

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}^{(+)}(\mathbf{r}) e^{-i\omega t} + \text{c.c.} = 2\text{Re} \left\{ \mathbf{E}^{(+)}(\mathbf{r}) e^{-i\omega t} \right\}. \quad (4.27)$$

Since all of the time dependence in $\mathbf{E}^{(+)}$ is of the form $e^{-i\omega t}$, keeping track of only $\mathbf{E}^{(+)}$ will greatly simplify electromagnetism calculations.

In particular, the Maxwell equations take on a simpler form for complex monochromatic fields. Because the time dependence is of the simple form $e^{-i\omega t}$, and because

$$\frac{\partial}{\partial t} e^{-i\omega t} = -i\omega e^{-i\omega t}, \quad (4.28)$$

we can always make the formal identification

$$\frac{\partial}{\partial t} \equiv -i\omega \quad (4.29)$$

for a monochromatic field. Then the Maxwell equations become

$$\begin{aligned} \nabla \cdot \mathbf{D}^{(+)} &= 0 \\ \nabla \cdot \mathbf{B}^{(+)} &= 0 \\ \nabla \times \mathbf{E}^{(+)} &= i\omega \mathbf{B}^{(+)} \\ \nabla \times \mathbf{H}^{(+)} &= -i\omega \mathbf{D}^{(+)}, \end{aligned} \quad (\text{Maxwell's equations, monochromatic field}) \quad (4.30)$$

with $\mathbf{D}^{(+)} = \epsilon_0 \mathbf{E}^{(+)} + \mathbf{P}^{(+)}$ and $\mathbf{B}^{(+)} = \mu_0 \mathbf{H}^{(+)}$.

Note that all of this complex notation stuff works out because the wave and Maxwell equations are all *linear*. That is, the real and imaginary parts (equivalently, positive- and negative-frequency parts) satisfy these equations separately, so they are equivalent to the regular Maxwell equations. The imaginary part is just an extra piece that helps to simplify calculations, which doesn't do any harm so long as you get rid of it at the end.

In fact, in writing down the Maxwell equations for monochromatic fields, we are implicitly taking a Fourier transform of the general field. Again, this follows from the linearity of the field equations along with the harmonic time dependence of the harmonic fields. We can write the solutions to the monochromatic

Maxwell equations as $\mathbf{E}^{(+)}(\mathbf{r}, \omega)$ (and so on for $\mathbf{H}^{(+)}$, $\mathbf{D}^{(+)}$, $\mathbf{B}^{(+)}$, etc.), and thereby obtain the general solution as an inverse Fourier transform,

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{E}^{(+)}(\mathbf{r}, \omega) e^{-i\omega t} d\omega,$$

(arbitrary field in terms of complex monochromatic components) (4.31)

where we can identify $\mathbf{E}^{(-)}(\mathbf{r}, \omega) = \mathbf{E}^{(+)}(\mathbf{r}, -\omega)$. In other words, the monochromatic Maxwell equation establishes the forms of the *harmonic basis functions*. We can then build up arbitrary solutions out of these basis functions.

4.5.1 Alternate Complex Notation

Another common convention for complex notation involves working with a *single* complex field (that we will write as \mathcal{E}) with time dependence $e^{-i\omega t}$, such that the real part is the physical field:

$$\mathbf{E}(\mathbf{r}, t) = \operatorname{Re} [\mathcal{E}(\mathbf{r}) e^{-i\omega t}] . \quad (4.32)$$

This is the same idea as before, but rather than thinking of a complex field as always accompanied by its conjugate, the idea here is to carry around an *extra* imaginary field $\mathcal{E}(\mathbf{r}) \sin(\omega t + \phi)$, where ϕ is the phase of \mathcal{E} . This extra, imaginary field is discarded at the end of the calculation by taking the real part of the answer.

This alternate notation is common in textbooks on optics and electrodynamics, and the $\mathbf{E}^{(\pm)}$ notation is more common in quantum optics. The $\mathbf{E}^{(\pm)}$ notation has some advantages, including a better emphasis on the physical field \mathbf{E} (since $\mathbf{E}^{(\pm)}$ is *part* of \mathbf{E}), and it is a bit less prone to causing confusion when computing *nonlinear* quantities of the field, such as the intensity. We will stick to the $\mathbf{E}^{(\pm)}$ notation here.

4.6 Intensity in Complex Notation

Quantities that are not linear in the fields are a bit more subtle. Still, for monochromatic waves, the Poynting vector and intensity are relatively simple:

$$\begin{aligned} \mathbf{S} &= \mathbf{E} \times \mathbf{H} \\ &= (\mathbf{E}^{(+)} e^{-i\omega t} + \mathbf{E}^{(-)} e^{i\omega t}) \times (\mathbf{H}^{(+)} e^{-i\omega t} + \mathbf{H}^{(-)} e^{i\omega t}) \\ &= \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} + \mathbf{E}^{(-)} \times \mathbf{H}^{(+)} + \mathbf{E}^{(+)} \times \mathbf{H}^{(+)} e^{-i2\omega t} + \mathbf{E}^{(-)} \times \mathbf{H}^{(-)} e^{i2\omega t}. \end{aligned} \quad (4.33)$$

For intensity calculations, we time-average the 2ω terms, so

$$\langle \mathbf{S} \rangle = \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} + \mathbf{E}^{(-)} \times \mathbf{H}^{(+)} = \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} + \text{c.c.} = 2\operatorname{Re} \{ \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} \} . \quad (4.34)$$

The optical intensity then is simply the magnitude of the Poynting vector,

$$I = |\mathbf{E}^{(+)} \times \mathbf{H}^{(-)} + \text{c.c.}| = |2\operatorname{Re} \{ \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} \}| , \quad (\text{intensity of monochromatic waves}) \quad (4.35)$$

but now in terms of the complex field components.

4.6.1 Complex Notation for Simple Dielectric Media

Since $\mathbf{D} = \epsilon \mathbf{E}$ and $\mathbf{B} = \mu_0 \mathbf{H}$ for a simple dielectric medium, we can rewrite the Maxwell equations for a monochromatic field in a simple dielectric as

$$\begin{aligned} \nabla \cdot \mathbf{E}^{(+)} &= 0 \\ \nabla \cdot \mathbf{H}^{(+)} &= 0 \\ \nabla \times \mathbf{E}^{(+)} &= i\omega \mu_0 \mathbf{H}^{(+)} \\ \nabla \times \mathbf{H}^{(+)} &= -i\omega \epsilon \mathbf{E}^{(+)}. \end{aligned}$$

(Maxwell's equations, simple dielectric) (4.36)

This is essentially the same set as the original Maxwell equations, but with $\partial/\partial t \rightarrow -i\omega$. Thus, the same replacement applies in the vector wave equation (4.21), and we obtain the **vector Helmholtz equation**

$$(\nabla^2 + k^2) \mathbf{E} = 0, \quad (4.37)$$

(vector Helmholtz equation)

where $k = \omega/c = \omega\sqrt{\epsilon\mu_0} = nk_0$ is the **wave number** in the medium, and $k_0 = \omega/c_0$ is the wave number in vacuum. In a simple dielectric, each component of each field again satisfies this equation separately, and so each satisfies the **scalar Helmholtz equation**

$$(\nabla^2 + k^2) E = 0, \quad (scalar \text{ Helmholtz equation}) \quad (4.38)$$

(scalar Helmholtz equation)

This is the time-independent version of the scalar wave equation (4.24), where we must keep in mind that a solution to the Helmholtz equation has an implied time dependence of $e^{-i\omega t}$.

4.7 Plane Waves

The **plane wave** is one of the simplest and most important solutions of the scalar wave equation. We can define the plane wave as the monochromatic field

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} e^{i\mathbf{k}\cdot\mathbf{r}}. \quad (4.39)$$

Recall that this expression has an implied harmonic time dependence:

$$E^{(+)}(\mathbf{r}, t) = E_0^{(+)} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}. \quad (4.40)$$

(scalar plane wave)

Here, $E_0^{(+)}$ is a complex constant scalar and \mathbf{k} is the **wave vector**. In terms of the components (k_x, k_y, k_z) , the inner product $\mathbf{k} \cdot \mathbf{r}$ is simply $k_x x + k_y y + k_z z$.

The plane wave satisfies the scalar Helmholtz equation. We can see this by observing that

$$\hat{x} \frac{\partial}{\partial x} E^{(+)} = \hat{x} i k_x E^{(+)}, \quad (4.41)$$

with similar relations holding for y and z . As for the time dependence of the monochromatic field, we can make the formal identification $\partial/\partial x \equiv ik_x$ for the plane wave. Combining all three components,

$$\nabla E^{(+)} = i\mathbf{k} E^{(+)}. \quad (4.42)$$

Iterating the differentiation once,

$$\nabla^2 E^{(+)} = -\mathbf{k} \cdot \mathbf{k} E^{(+)} = -k^2 E^{(+)}. \quad (4.43)$$

Thus, we recover the Helmholtz equation, where we can identify the **wave number** we defined before with the magnitude of the wave vector \mathbf{k} .

Now let's examine the phase of the plane wave. If we define ϕ to be the phase of the complex amplitude $E_0^{(+)}$,

$$E_0^{(+)} = |E_0^{(+)}| e^{i\phi}, \quad (4.44)$$

then we can write

$$E^{(+)}(\mathbf{r}) = |E_0^{(+)}| e^{i(\mathbf{k}\cdot\mathbf{r} + \phi)}. \quad (4.45)$$

Since $e^{i\phi} = e^{i(\phi+2\pi)}$, the wave repeats itself every 2π in phase. Thus, the **wave fronts** (surfaces of constant phase) are given by

$$\mathbf{k} \cdot \mathbf{r} = k_x x + k_y y + k_z z = 2\pi q + \phi \quad (4.46)$$

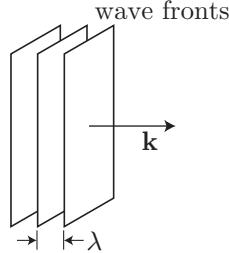
(plane-wave fronts)

for integer q . These surfaces of constant $\mathbf{k} \cdot \mathbf{r}$ are planes perpendicular to \mathbf{k} , separated by the **wavelength**

$$\lambda := \frac{2\pi}{k} \quad (4.47)$$

(wavelength)

in space (i.e., λ is the distance along \mathbf{k} corresponding to a phase shift of 2π).



Let's simplify a bit by orienting the z -axis along \mathbf{k} , so that $\mathbf{k} = k\hat{z}$. Then the plane wave has the form

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} e^{ikz}. \quad (4.48)$$

The corresponding real wave is

$$E(\mathbf{r}, t) = E^{(+)}(\mathbf{r}) e^{-i\omega t} + \text{c.c.} = |E_0^{(+)}| e^{i(kz - \omega t + \phi)} + \text{c.c.} = 2|E_0^{(+)}| \cos(kz - \omega t + \phi) \quad (4.49)$$

In standard form, the real plane wave is

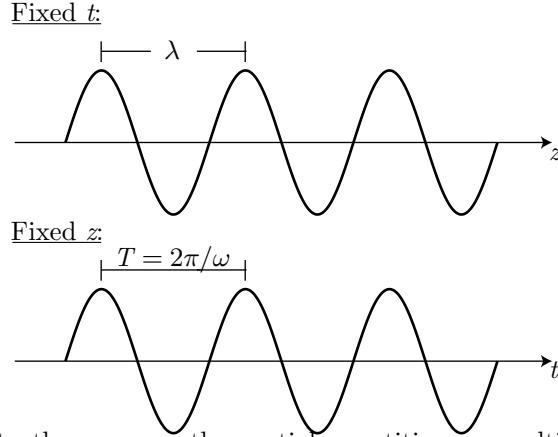
$$E(\mathbf{r}, t) = E_0 \cos(kz - \omega t + \phi), \quad (4.50)$$

(real plane wave)

where $E_0 := 2|E_0^{(+)}|$. Note that we can also write the plane wave in the form

$$E(\mathbf{r}, t) = E_0 \cos[k(z - ct) + \phi] \quad (4.51)$$

where again $c = \omega/k$ is the speed of light. As time evolves, the points of stationary phase—constant $(z - ct)$ —move with **phase velocity** c .



In a medium, as compared to the vacuum, the spatial quantities are multiplied by the refractive index, whereas the time quantities are unchanged:

$$\begin{aligned} c &= c_0/n \\ \lambda &= \lambda_0/n \\ k &= nk_0 \\ \omega &= \omega_0. \end{aligned} \quad (4.52)$$

Here, as before, the subscripted quantities refer to the vacuum.

4.8 Vector Plane Waves

In the general case, the electric field is a vector, and sometimes it is insufficient to treat the electromagnetic field as a scalar. Thus we should treat the **polarization** or *orientation* of the field explicitly—along with the direction of *motion* that we have already treated—and write

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}^{(+)}(\mathbf{r}) e^{-i\omega t} + \text{c.c.} = \mathbf{E}_0^{(+)} e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} + \text{c.c.}, \quad (4.53)$$

and since each component of this vector field separately satisfies the (homogeneous-medium) scalar wave equation (4.24), this field satisfies the vector wave equation (4.8). We can also guess that this form holds for the magnetic field, since each of its components satisfies the scalar wave equation:

$$\mathbf{H}(\mathbf{r}, t) = \mathbf{H}^{(+)}(\mathbf{r}) e^{-i\omega t} + \text{c.c.} = \mathbf{H}_0^{(+)} e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} + \text{c.c.}, \quad (4.54)$$

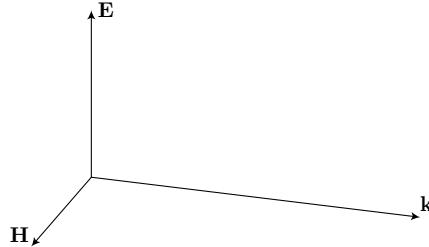
Using the Maxwell equations from Eqs. (4.36),

$$\begin{aligned} \nabla \times \mathbf{H}^{(+)} &= -i\omega\epsilon\mathbf{E}^{(+)} \\ \nabla \times \mathbf{E}^{(+)} &= i\omega\mu_0\mathbf{H}^{(+)}, \end{aligned} \quad (4.55)$$

we can use the identification $\nabla \equiv i\mathbf{k}$ for the plane wave to write

$$\begin{aligned} \mathbf{k} \times \mathbf{H}^{(+)} &= -\omega\epsilon\mathbf{E}^{(+)} \\ \mathbf{k} \times \mathbf{E}^{(+)} &= \omega\mu_0\mathbf{H}^{(+)}. \end{aligned} \quad (4.56)$$

By inspecting this pair of equations, we see that \mathbf{k} , \mathbf{H}_0 , and \mathbf{E}_0 are all *mutually orthogonal* (i.e., they determine a basis in three dimensions), and they must be arranged as shown below.



Since \mathbf{E}_0 and \mathbf{H}_0 are orthogonal to \mathbf{k} , plane waves are often called **Transverse Electromagnetic (TEM) waves**.

Taking the magnitudes of the above two Maxwell equations, we find

$$\begin{aligned} H_0 &= \left(\frac{\omega\epsilon}{k} \right) E_0 \\ H_0 &= \left(\frac{k}{\omega\mu_0} \right) E_0, \end{aligned} \quad (4.57)$$

where again $E_0 := 2|E_0^{(+)}|$ and similarly $H_0 := 2|H_0^{(+)}|$. For these two equations to be consistent, we require

$$\frac{\omega\epsilon}{k} = \frac{k}{\omega\mu_0}, \quad (4.58)$$

which leads to

$$k = \omega\sqrt{\mu_0\epsilon} = \frac{\omega}{c} = \frac{n\omega}{c_0} = nk_0, \quad (4.59)$$

which is the same as the condition for the wave to satisfy the Helmholtz equation.

4.8.1 Wave Impedance

Now if we consider the ratio of the electric and magnetic fields,

$$\frac{E_0}{H_0} = \frac{\omega\mu_0}{k} = \frac{\mu_0 c_0}{n} = \frac{1}{n} \sqrt{\frac{\mu_0}{\epsilon_0}}, \quad (4.60)$$

we can define this constant ratio of E_0 to H_0

$$\eta := \frac{1}{n} \sqrt{\frac{\mu_0}{\epsilon_0}}, \quad (4.61) \quad (\text{impedance of dielectric medium})$$

which has the dimensions of impedance. Thus η is the **wave impedance of the medium**. We can also write

$$\eta = \frac{\eta_0}{n}, \quad (4.62) \quad (\text{wave impedance modified by refractive index})$$

where

$$\eta_0 := \sqrt{\frac{\mu_0}{\epsilon_0}} = \mu_0 c_0 = \frac{1}{\epsilon_0 c_0} \approx 377 \Omega \quad (4.63) \quad (\text{impedance of vacuum})$$

is the **wave impedance of the vacuum**.

Returning to the Poynting vector,

$$\langle \mathbf{S} \rangle = \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} + \text{c.c.} = \left(E_0^{(+)} H_0^{(-)} + \text{c.c.} \right) \hat{k} = \left(\frac{|E_0^{(+)}|^2}{\eta} + \text{c.c.} \right) \hat{k} = \left(\frac{E_0^2}{4\eta} + \text{c.c.} \right) \hat{k} = \frac{E_0^2}{2\eta} \hat{k}, \quad (4.64)$$

where $\hat{k} := \mathbf{k}/k$. Thus, for a plane wave, we can write the intensity in terms of the real amplitude E_0 as

$$I = \frac{E_0^2}{2\eta}. \quad (4.65) \quad (\text{intensity in terms of real field amplitude})$$

Alternately, we can also write the intensity

$$I = \frac{2|E^{(+)}|^2}{\eta} \quad (4.66) \quad (\text{intensity in terms of complex amplitude})$$

in terms of the complex field $|E^{(+)}|$.

As an example, a typical light bulb uses 100 W of electric power. Assuming a light conversion efficiency of $\sim 2.5\%$ and a diameter of ~ 5 cm, the intensity at the surface is the optical power per unit area:

$$I = \frac{P}{A} = \frac{2.5 \text{ W}}{4\pi(2.5 \text{ cm})^2} = 0.03 \frac{\text{W}}{\text{cm}^2}. \quad (4.67)$$

The electric field is thus

$$E_0 = \sqrt{2\eta_0 I} = 5 \frac{\text{V}}{\text{cm}} \quad (4.68)$$

at the surface.

A laser, as we will see later, has much higher intensity and electric field even for a much more modest power. A low-power He-Ne laser might typically have $P \sim 2$ mW of output power and a beam diameter of $d \sim 0.8$ mm. Making the simplistic estimate $I = P/d^2$, we find an intensity of 0.3 W/cm^2 and an electric field of 15 V/cm, values larger than the light bulb despite a much smaller power rating.

4.9 Exercises

Problem 4.1

- (a) Prove the vector identity

$$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} \quad (4.69)$$

for an arbitrary vector \mathbf{A} . (Hint: try doing this first for only one Cartesian component of this equation and then argue that the other components follow in the same way.)

- (b) Use this identity and Maxwell's equations to show that the magnetic field $\mathbf{H}(\mathbf{r}, t)$ in free space satisfies the wave equation

$$\nabla^2 \mathbf{H} - \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} \mathbf{H} = 0, \quad (4.70)$$

which is the same equation satisfied by the electric field $\mathbf{E}(\mathbf{r}, t)$. Here, c_0 is the speed of light in vacuum.

Problem 4.2

For a plane-wave solution of the wave equation,

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \cos(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi_0), \quad (4.71)$$

we can write the constant \mathbf{E}_0 as $\hat{e}E_0$, where the unit vector \hat{e} is the *polarization* and E_0 is a scalar constant. Show that the wave cannot be polarized along the \mathbf{k} direction. Prove this directly; do not make any assumption about the form of the magnetic field.

Problem 4.3

Recall the wave equation for the electric field in a simple, linear dielectric,

$$\nabla^2 \mathbf{E} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \mathbf{E} = 0. \quad (4.72)$$

Consider the particular solution,

$$\mathbf{E}(\mathbf{r}, t) = \hat{z}E_0 \cos(kx - \omega t + \phi_0), \quad (4.73)$$

where \mathbf{E}_0 and ϕ_0 are constants and \hat{z} is the unit vector along the z -direction.

- (a) Show explicitly that this solution satisfies the wave equation provided $\omega/k = c$.
- (b) Show explicitly that the solution is invariant under transformations of the form $t \rightarrow t + T$, where $T = 2\pi/\omega$, and thus show that the solution is *periodic in time* with period T (i.e., the time required for the wave to repeat itself at fixed position is T).
- (c) Show that the wave is spatially periodic with period λ , where $\lambda = 2\pi/k$.
- (d) Using a similar argument, show that ω/k is the *phase velocity* (the velocity at which the wave peaks travel) of the wave.
- (e) Show that the corresponding solution for the magnetic field is

$$\mathbf{H}(\mathbf{r}, t) = -\hat{y} \frac{E_0}{\mu_0 c} \cos(kx - \omega t + \phi_0). \quad (4.74)$$

- (f) Write down the complex form $\mathbf{E}^{(+)}$ of the wave in terms of the above real wave. Be explicit about the dependence on all the parameters in the real wave.

Problem 4.4

The energy density (energy stored per unit volume) of an electromagnetic field in a simple, linear dielectric is

$$w = \frac{1}{2} (\epsilon \mathbf{E} \cdot \mathbf{E} + \mu_0 \mathbf{H} \cdot \mathbf{H}) \quad (4.75)$$

for real field \mathbf{E} and \mathbf{H} .

Consider a plane wave in complex notation of the form

$$\mathbf{E}^{(+)}(\mathbf{r}, t) = \hat{z}E_0^{(+)}e^{i(kx-\omega t)}, \quad (4.76)$$

with $E_0^{(+)}$ a complex constant.

(a) Show that the time-averaged energy density is given by

$$\langle w \rangle = \frac{1}{4} (\epsilon E_0^2 + \mu_0 H_0^2), \quad (4.77)$$

where $E_0 = 2|E_0^{(+)}|$. What is the significance of the extra factor of $1/2$ in this expression? (Hint: it's not the definition of E_0 .)

(b) What is the real wave corresponding to the complex wave given above? Be explicit about the parameter dependence.

(c) Calculate w , the Poynting vector \mathbf{S} , and the intensity I for the real plane wave of part (b).

(d) Show, for the real plane wave of part (b), that $\mathbf{S} = cw\hat{x}$. With the interpretation that the Poynting vector represents an *energy flux density*, argue that this means that the electromagnetic energy of a plane wave propagates with velocity c .

Problem 4.5

A laser with $\lambda = 200\pi$ nm points along the $-x$ -direction, and its electric field is polarized in the z -direction. Assume the laser output is a plane wave in free space with intensity 1 W/cm^2 . Write down expressions for the electric **and** magnetic fields for the laser beam, giving numerical values for any parameters you use. Don't worry too much about arithmetic; just reduce your numerical answers to reasonably simple fractions, powers, etc.

Problem 4.6

Write down an expression for **both** the electric and magnetic field for a monochromatic plane wave in vacuum with the following properties: the wave vector is $(-k, k, 0)/\sqrt{2}$, the electric field amplitude is E_0 , and the polarization is along the positive z -direction. Your expressions should only involve these parameters and fundamental constants, and you should *explain* each aspect of what you write down. (Write down the physical fields, not the complex versions.)

Problem 4.7

Two possible definitions of the intensity in terms of the Poynting vector are the *magnitude* of the *time-averaged* Poynting vector,

$$I := |\langle \mathbf{S} \rangle|, \quad (4.78)$$

and the *time-average* of the *magnitude* of the Poynting vector,

$$I' := \langle |\mathbf{S}| \rangle = \langle S \rangle. \quad (4.79)$$

The first definition is the one we use here, while the second definition is common in textbooks on optics and electromagnetism. The point of this problem is to examine whether the order of the two operations matters, and if so which ordering is more sensible.

Consider the composite field

$$\mathbf{E}^{(+)}(\mathbf{r}, t) = \mathbf{E}_1^{(+)}(\mathbf{r}, t) + \mathbf{E}_2^{(+)}(\mathbf{r}, t), \quad (4.80)$$

where the first field is an optical plane wave propagating in the z -direction and polarized in the x -direction,

$$\mathbf{E}_1^{(+)}(\mathbf{r}, t) = \hat{x}E_{10}^{(+)}e^{i(kz-\omega t)}, \quad (4.81)$$

while the second is a dc field, polarized in the z -direction:

$$\mathbf{E}_2^{(+)}(\mathbf{r}, t) = \hat{z} E_{20}^{(+)}. \quad (4.82)$$

To make algebra easier, take the special case $z = 0$, with $E_{10}^{(+)}$ and $E_{20}^{(+)}$ both being real numbers.

- (a) Write down the magnetic field $\mathbf{H}_1^{(+)}(\mathbf{r}, t)$ that is consistent with $\mathbf{E}_1^{(+)}(\mathbf{r}, t)$.
- (b) We will take the magnetic field corresponding to the second electric field to be zero, $\mathbf{H}_2^{(+)}(\mathbf{r}, t) = 0$. Explain why this is consistent with Maxwell's equations.
- (c) Compute the Poynting vector for the combined field.
- (d) Compute the time average of the Poynting vector, and then the intensity according to the first definition.
- (e) Compute the magnitude of the Poynting vector. The time average will then be difficult to compute; to make it easier, expand your result to lowest nontrivial order in

$$\alpha := \frac{E_{20}^{(+)}}{E_{10}^{(+)}}. \quad (4.83)$$

Then compute the time average, and hence the intensity according to the second definition.

- (f) Give a *physical* argument for why one definition of the intensity is better than the other. (Consider what intensity is intended to describe, and remember that an alternative is not preferable merely because it is easier to calculate!)

Problem 4.8

In the Coulomb-gauge formulation of electromagnetism, there is a field \mathbf{A} (the vector potential) that satisfies $\nabla \cdot \mathbf{A} = 0$, and defines the electromagnetic fields via $\mathbf{E} = -\partial \mathbf{A} / \partial t$ and $\mathbf{B} = \mu_0 \mathbf{H} = \nabla \times \mathbf{A}$. Use these relations and Maxwell's equations to show that \mathbf{A} satisfies the same wave equation as \mathbf{E} (and \mathbf{B} and \mathbf{H}). For simplicity, stick to vacuum.

Problem 4.9

- (a) The electromagnetic fields can be generally defined in terms of the scalar potential ϕ and vector potential \mathbf{A} as

$$\mathbf{E} = -\nabla\phi - \partial_t \mathbf{A}, \quad \mathbf{B} = \nabla \times \mathbf{A}. \quad (4.84)$$

Use Maxwell's equations to derive a set of two coupled wave equations for the potentials **in a dielectric medium** described by permittivity $\epsilon(\mathbf{r})$.

- (b) Then show that in the generalized Lorenz gauge, specified by the gauge condition

$$\nabla \cdot \epsilon \mathbf{A} + \epsilon_0^2 \mu_0 \partial_t \phi = 0, \quad (4.85)$$

the two wave equations decouple.

Chapter 5

Interference

5.1 Superposition of Two Plane Waves

Interference occurs when multiple waves are added together, and it is one of the most obvious effects not accounted for by geometrical optics. First, we will consider the simplest case of adding two scalar, monochromatic waves $E_1^{(+)}(\mathbf{r})$ and $E_2^{(+)}(\mathbf{r})$. Both have the same frequency, and thus an implicit time dependence of $e^{-i\omega t}$. The superposition of the two waves is just the sum

$$E^{(+)}(\mathbf{r}) = E_1^{(+)}(\mathbf{r}) + E_2^{(+)}(\mathbf{r}). \quad (5.1)$$

Clearly the combined field $E^{(+)}(\mathbf{r})$ is also monochromatic.

Recall from Eq. (4.66) that the intensity of each component wave is given by

$$I_{1,2} = \frac{2|E_{1,2}^{(+)}|^2}{\eta}. \quad (5.2)$$

To simplify our notation a bit, we can introduce the fields $U_{1,2}(\mathbf{r})$, defined by

$$U_{1,2} := \sqrt{\frac{2}{\eta}} E_{1,2}^{(+)}(\mathbf{r}), \quad (5.3) \quad (\text{scaled electric fields})$$

so that the intensity is simply the square of the U field without any extra coefficients:

$$I_{1,2} = |U_{1,2}|^2. \quad (5.4) \quad (\text{intensity in terms of scaled electric fields})$$

In this notation, the intensity of the superposition is simply

$$I = |U|^2 = |U_1 + U_2|^2 = |U_1|^2 + |U_2|^2 + U_1^* U_2 + U_1 U_2^*. \quad (5.5)$$

Writing out explicitly the phases of $U_{1,2}$,

$$U_{1,2} = \sqrt{I_{1,2}} e^{i\phi_{1,2}}, \quad (5.6)$$

we find that

$$\begin{aligned} I &= I_1 + I_2 + \left(\sqrt{I_1 I_2} e^{i(\phi_2 - \phi_1)} + \text{c.c.} \right) \\ &= I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(\phi_2 - \phi_1). \end{aligned} \quad (\text{combined intensity of two waves}) \quad (5.7)$$

Notice that all the phase dependence is in the factor $\cos(\phi_2 - \phi_1)$, which ranges from -1 to 1 . Putting these extreme values into this expression, we see that the intensity ranges from $(\sqrt{I_1} - \sqrt{I_2})^2$ to $(\sqrt{I_1} + \sqrt{I_2})^2$.

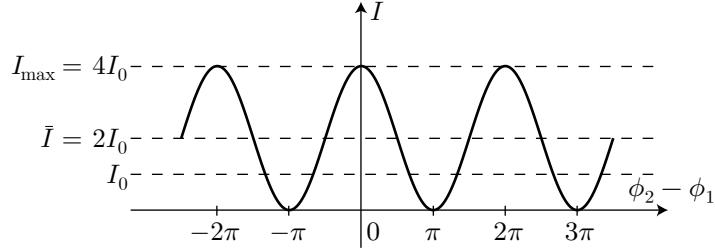
A useful and simple special case is where the intensities are equal:

$$I_1 = I_2 =: I_0. \quad (5.8)$$

In this case the intensity becomes

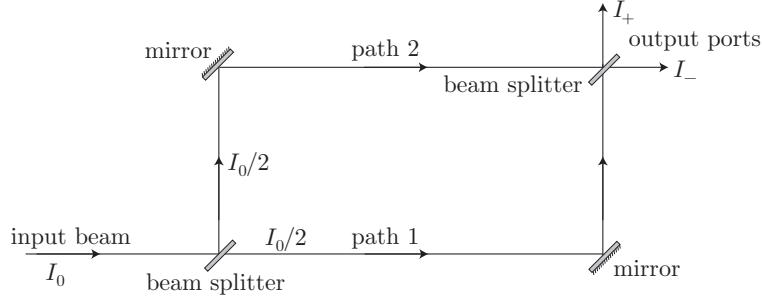
$$I = 2I_0 [1 + \cos(\phi_2 - \phi_1)]. \quad (5.9)$$

If $\phi_2 = \phi_1$, the intensity attains its maximum value $I = 4I_0$, and this situation is **fully constructive interference**. If $\phi_2 - \phi_1 = \pi$, the intensity attains its minimum value $I = 0$, and this situation is **fully destructive interference**. Comparing to the intensity range above, we see that the case of equal intensities is the case of *maximum intensity variation*. The intensity \bar{I} averaged over all phases is $2I_0$, as we might expect on average for adding two beams of intensity I_0 from energy conservation. Of course, adding two waves should always conserve energy *regardless* of the phase, and we will see how this works out when we consider interferometers below, in particular in *how* waves are combined using beam splitters.



5.2 Mach-Zehnder Interferometer

An **interferometer** is a device that uses the interference effect to do something useful. One of the simplest interferometers is the **Mach-Zehnder interferometer**, shown here.



A beam of intensity I_0 is split into two components by a **beam splitter** (a beam splitter is simply a device that deflects part of the intensity of an optical beam, while allowing the rest to pass). For simplicity we will assume a 50% lossless beam splitter, so that half the intensity goes into either arm of the interferometer. The two components are then recombined on the second beam splitter and the intensities I_+ and I_- can be monitored at their respective output ports.

The wave in path 1, just before the second beam splitter, is

$$U_1 = \sqrt{\frac{I_0}{2}} e^{i\phi_1}, \text{ where } \phi_1 = k_1 d_1 = \frac{2\pi n_1 d_1}{\lambda_0}, \quad (5.10)$$

where d_1 is the length of path 1, n_1 is the refractive index of path 1, and λ_0 is the optical wavelength (in vacuum). Similarly, for the wave in path 2 just before the second beam splitter

$$U_2 = \sqrt{\frac{I_0}{2}} e^{i\phi_2}, \text{ where } \phi_2 = k_2 d_2 = \frac{2\pi n_2 d_2}{\lambda_0}. \quad (5.11)$$

Then the intensities of the two output ports are given by squaring the sums of the intensities. At the first port,

$$I_+ = \left| \frac{U_1}{\sqrt{2}} + \frac{U_2}{\sqrt{2}} \right|^2 = \frac{I_0}{2} [1 + \cos(\phi_2 - \phi_1)]. \quad (5.12)$$

At the second port, we also add the fields, but *with an extra minus sign*:

$$I_- = \left| \frac{U_1}{\sqrt{2}} - \frac{U_2}{\sqrt{2}} \right|^2 = \frac{I_0}{2} [1 - \cos(\phi_2 - \phi_1)]. \quad (5.13)$$

Why the extra minus sign in I_- ? Well, first of all, we can see that energy conservation works in this case, since the total output intensity is $I_+ + I_- = I_0$, which is the same as the input intensity. This wouldn't have worked out if we omitted the extra minus sign. But we can do a better job of justifying it by considering an important time-reversal argument.

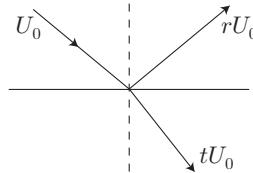
5.3 Stokes Relations

The Stokes relations follow elegantly from the time-reversal invariance of electromagnetism. We can see this invariance directly from the electromagnetic wave equation:

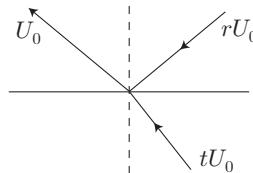
$$\nabla^2 \mathbf{E} - \frac{1}{c_0^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \quad (5.14)$$

Since time appears only as a second derivative, the equation is invariant under the replacement $t \rightarrow -t$.

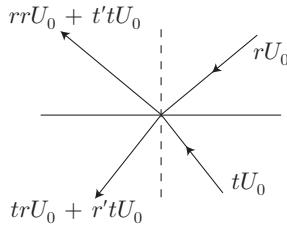
Consider a beam of light at an interface between two media (e.g., a dielectric interface). From geometrical optics, we know that part of the beam will transmit (refract) and part will reflect, although we don't know how much of the light will go in each direction. So we'll just define coefficients r and t , such that if U_0 is the amplitude of the incident field, then the amplitude of the reflected field is rU_0 , and the amplitude of the transmitted field is tU_0 .



In standard terminology, r is the **field reflection coefficient** and t is the **field transmission coefficient**. Time-reversal invariance states that this should also work in *reverse*: if two waves of amplitude rU_0 and tU_0 are incident from opposite sides of the same interface, they should combine to form a *single* beam that travels away from the interface. Of course, this happens by constructive interference. In the other opposite location where you might expect to see a beam going away from the interface, there really is none due to complete destructive interference. Note that *in principle* this is possible and it works for the purposes of this derivation. *In practice*, it is often very difficult to implement ideal time-reversed situations such as this.



Now let's consider all the reflections and transmissions from the two incident beams that form the two outgoing beams in the time-reversed case. On the top side, the outgoing beam comprises a reflection of the rU_0 and a transmission from the tU_0 beam. So we can write this field as $rrU_0 + t'tU_0$. The primed coefficients r' and t' denote the reflection and transmission coefficients for a ray *incident from below*, which are possibly different from r and t . Similarly, the outgoing beam on the bottom comprises a transmission of the rU_0 beam and a reflection of the tU_0 beam, so we can write this field as $trU_0 + r'tU_0$.



But the last two diagrams must be consistent with each other, so for the top beam we require that

$$r^2 U_0 + t't U_0 = U_0, \quad (5.15)$$

and the bottom beam must vanish, so

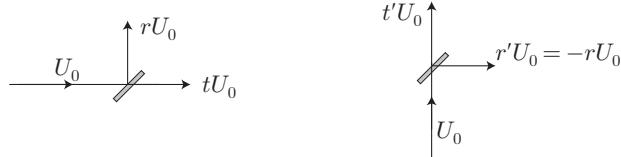
$$trU_0 + r'tU_0 = 0. \quad (5.16)$$

Simplifying these two equations a bit, we arrive at the **Stokes relations**

$$\begin{aligned} r' &= -r \\ t't &= 1 - r^2 \end{aligned} \quad (5.17) \quad (\text{Stokes relations})$$

in standard form.

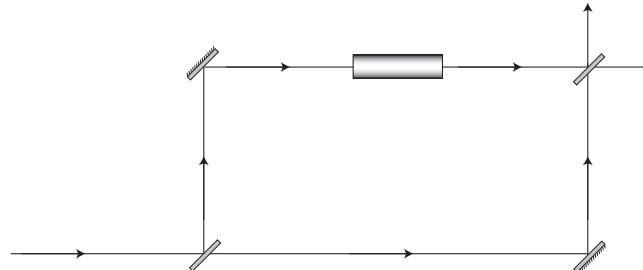
Now we can use the Stokes relations to analyze what happens at a physical beam splitter. A physical beam splitter is typically something like a slab of glass, where the splitting occurs at one interface and the no splitting occurs at the other (often this is accomplished with special dielectric coatings).



If the reflection coefficient is r for the beam that impinges on the beam splitter from the outside, the beam in the other arm of the interferometer impinges on the same interface from within the beam splitter. Thus, for the other arm, the reflection coefficient must be $-r$. We can't work out exactly what happens to the transmission coefficients from the Stokes relations, but it works out that they are equal. Since $r^2 = 0.5$ for the 50% splitters (the split ratio refers to *intensity*), we have that $t't = 1 - r^2 = 0.5$, so there is no sign change between t and t' .

5.4 Mach-Zehnder Interferometer: Applications

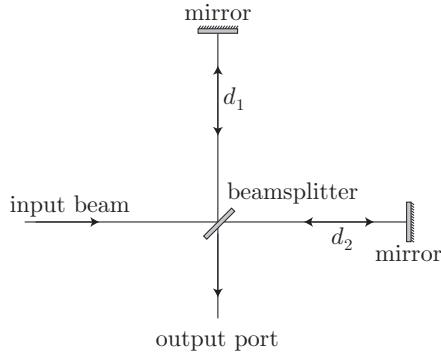
Since the Mach-Zehnder interferometer has two spatially separated but otherwise equivalent paths, it is typically used to measure an optical phase shift (path-length change) due to an object in one arm.



The other arm acts as a “phase reference” for the phase-shifted beam. For example, you could use this technique to measure the index of refraction of gas in a cell by measuring the phase shift as a function of the gas pressure. Or you can visualize a phase object such as a flame or a gas flow, where the shape of the optical fringes reflects the spatial index profile of the object.

5.5 Michelson Interferometer

The Mach-Zehnder interferometer can be “folded” so that the two beam splitters coincide and can be replaced by a single splitter. This new configuration is the **Michelson interferometer**.



Note that in this diagram, the input port is also one of the output ports, but often in this configuration you would only monitor the other output port.

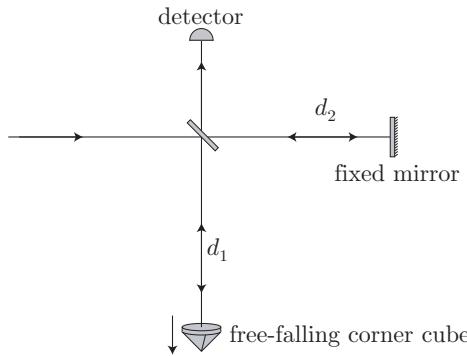
Just as in the Mach-Zehnder case, the Michelson interferometer measures the path-length difference between the two arms. Assuming the same refractive index for both arms and a 50/50 beam splitter, the output intensity is

$$I_{\text{out}} = \frac{I_0}{2} \left[1 + \cos \left(\frac{2\pi n(d_2 - d_1)}{\lambda} \right) \right].$$

(output intensity of Michelson interferometer) (5.18)

This interferometer is very useful for measuring lengths. For example, one arm can act as a reference distance, while the output changes depending on how far the other “probe” mirror moves. Also, if *two* different laser beams are coupled into the same interferometer, measuring the fringes of both lasers as a mirror is moved gives a measure of the relative wavelength—useful as an optical “wave meter” or “λ-meter” for measuring the wavelength of an unknown laser in terms of a known reference laser.

One example is the **interferometric gravimeter**, which is a Michelson interferometer where one of the mirrors is actually a corner-cube reflector (a prism that retroreflects the incident beam in the same direction independent of small changes in its orientation). The corner cube acts as a freely falling mass, and by monitoring the output it is possible to get a real-time position measurement of the cube’s position.



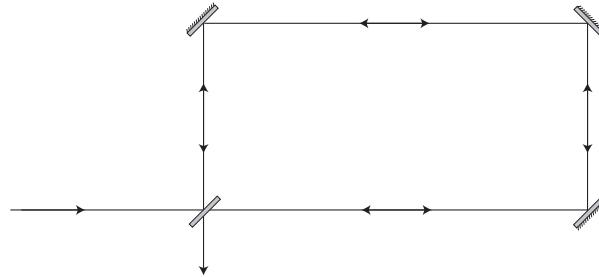
This method is still the state-of-the-art method for sensitive measurements of g .

Another important application of the Michelson interferometer was its original version and namesake in the Michelson-Morley experiment. The idea here was to search for an “aether,” the fundamental medium through which light propagates, which would define a “rest frame” for light. The interferometer would show a fringe shift if it moved through the aether (because it was stationary on Earth) depending on its orientation with respect to its velocity. No effect was observed, the conclusion being that there is no preferred rest frame for light propagation.

Yet another example is the LIGO (Laser Interferometric Gravitational Wave Observatory) experiment, which uses some of the largest and most impressive Michelson interferometers built thus far. The arms of the LIGO interferometers are 4 km long, with *many* technical refinements to detect displacements of mirrors due to gravitational waves. The strain sensitivities achieved so far are better than 10^{-21} , corresponding to detecting the displacement of a macroscopic mirror to a precision of much less than a nuclear diameter.

5.6 Sagnac Interferometer

The **Sagnac interferometer** is another variation on the Mach-Zehnder configuration, where the second beam splitter is replaced by a mirror.



The mirror causes the beams to continue around, so really there are not two distinct “arms” of this interferometer. Rather, two beams traverse the same ring but in opposite directions. Since there is no path-length difference, there is no sense of distance measurement in this system. However, if the interferometer *rotates*, the symmetry between the two beams is broken due to the Doppler effect, giving an effective phase shift between the two beams. Thus, the Sagnac interferometer is a very effective rotation sensor. One common instrument that uses this effect is the **ring-laser gyroscope**, which uses this interferometer as the cavity for a laser for sensitive detection of rotation.

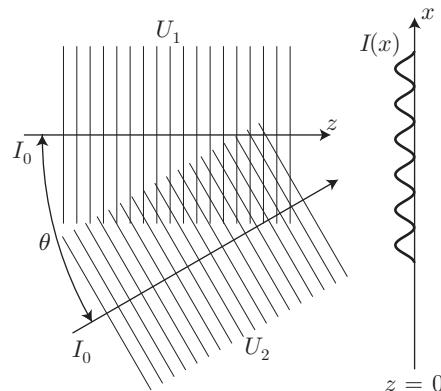
5.7 Interference of Two Tilted Plane Waves

A simple but useful effect to understand is the interference pattern between two plane waves with slightly different angles. Consider a plane wave propagating in the z -direction,

$$U_1 = \sqrt{I_0} e^{ikz}, \quad (5.19)$$

and a second wave of equal intensity propagating at an angle θ with respect to the first,

$$U_2 = \sqrt{I_0} e^{i[k(\cos \theta)z + k(\sin \theta)x]}. \quad (5.20)$$



For a screen at $z = 0$, the phase difference between the two waves is just $(k \sin \theta)x$, so from Eq. (5.9), the intensity on the screen is

$$I_{\text{screen}} = 2I_0 \{1 + \cos[(k \sin \theta)x]\}. \quad (\text{interference pattern of two tilted plane waves}) \quad (5.21)$$

Thus, the interference pattern varies sinusoidally with spatial period

$$\frac{2\pi}{k \sin \theta} = \frac{\lambda}{\sin \theta}. \quad (5.22)$$

This is useful as a simple model of misalignment in a two-beam interferometer. Obviously this result is a useful “interferometer” in its own right: this gives you an accurate method for measuring angles or printing accurate periodic patterns lithographically (e.g., for holographically fabricating diffraction gratings).

5.8 Multiple-Wave Interference

A useful generalization of the two-beam interference theme is the interference due to N waves:

$$U = U_1 + U_2 + \cdots + U_N. \quad (5.23)$$

In general, this is a very complicated problem, but we can easily treat the special case of equal amplitudes and constant phase differences:

$$U_n = \sqrt{I_0} e^{i(n-1)\phi}, \quad n = 1, 2, \dots, N. \quad (5.24)$$

The algebra simplifies if we define $h := e^{i\phi}$, so that

$$U_n = \sqrt{I_0} h^{n-1}. \quad (5.25)$$

Then

$$U = \sqrt{I_0} (1 + h + h^2 + \cdots + h^{N-1}) = \sqrt{I_0} \frac{1 - h^N}{1 - h} = \sqrt{I_0} \frac{1 - e^{iN\phi}}{1 - e^{i\phi}}. \quad (5.26)$$

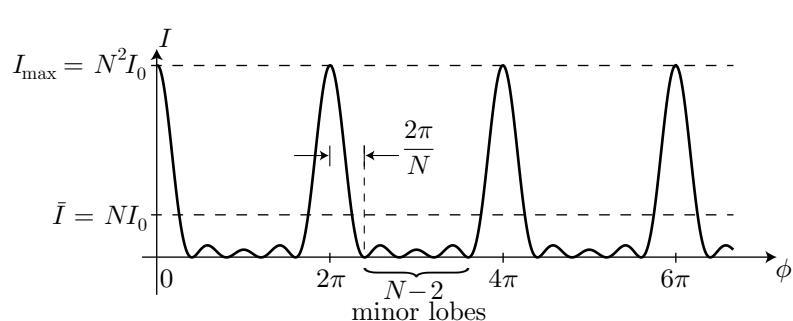
The corresponding intensity is the squared magnitude of this expression:

$$I = I_0 \left| \frac{e^{-iN\phi/2} - e^{iN\phi/2}}{e^{-i\phi/2} - e^{i\phi/2}} \right|^2. \quad (5.27)$$

Here, we multiplied through by a factor of $e^{-iN\phi/2}/e^{-i\phi/2}$ inside the magnitude, which doesn't change anything. Finally, we have

$$I = I_0 \frac{\sin^2(N\phi/2)}{\sin^2(\phi/2)}. \quad (\text{interference of } N \text{ waves, equal phase difference}) \quad (5.28)$$

This corresponds to an interference pattern with tall, narrow peaks of height $N^2 I_0$ with $N - 2$ sub-lobes in between.



The corresponding average intensity is

$$\bar{I} = \frac{1}{2\pi} \int_0^{2\pi} I d\phi = NI_0, \quad (5.29)$$

again as we expect from energy conservation. This problem is a useful model for diffraction through N thin, equally spaced slits, or a diffraction grating where N rulings are illuminated. However, these problems are better dealt with in the more general and powerful formalism of Fourier optics (Chapter 12).

5.9 Exercises

Problem 5.1

The *Young double slit* experiment is another example of a simple interferometer. Two slits in a planar wall are illuminated by the same light source. Interference fringes appear in the light on a distant, planar screen (parallel to the first wall), because the path lengths of the two beams to the screen vary slightly with the propagation direction of the waves.

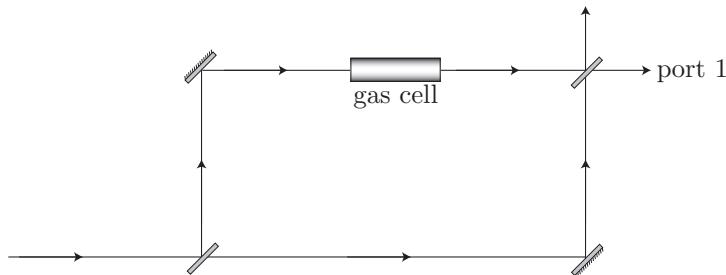
- Sketch the setup and show that the angular distance between successive dark bands (areas of destructive interference) is $\Delta\theta = \lambda/a$, where λ is the wavelength of the light and a is the separation between slits. Because the screen is distant, you can assume the two waves propagating to any particular point on the screen are approximately parallel.
- Suppose that the two slits are 0.5 mm apart and diffract monochromatic light onto a screen 10 m away. You measure the fringe spacing (distance between dark bands) on the screen to be 1.266 cm. What is the wavelength of the light?
- What type of laser could plausibly be the source of the light in part (b)?

Problem 5.2

Consider a Michelson interferometer illuminated by laser light. Assume that the incident light is well described by a plane wave. Suppose that one of the retroreflecting mirrors has a small misalignment of angle $\delta\theta$. Describe the interference pattern on a screen at the output of the interferometer, relating aspects of the interference pattern to $\delta\theta$ as appropriate. Also describe how the pattern changes as either mirror moves.

Problem 5.3

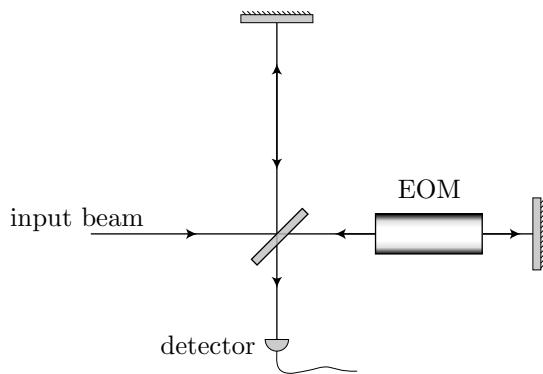
Consider a Mach-Zehnder interferometer with a gas cell of length 10 cm in one arm as shown. The input light is $\lambda = 200\pi$ nm.



The cell is initially evacuated, and all of the interferometer light exits output port 1. As the cell is pressurized, you observe that port 1 goes dark then becomes bright again a total of 10 complete cycles. What is the refractive index of the gas at the final pressure?

Problem 5.4

Consider a Michelson interferometer, operating at wavelength λ , containing an **electro-optic modulator** (EOM). For the purposes of this problem, think of the EOM as a slab of glass of thickness d , whose refractive index n can be changed slightly via an electrical bias signal.

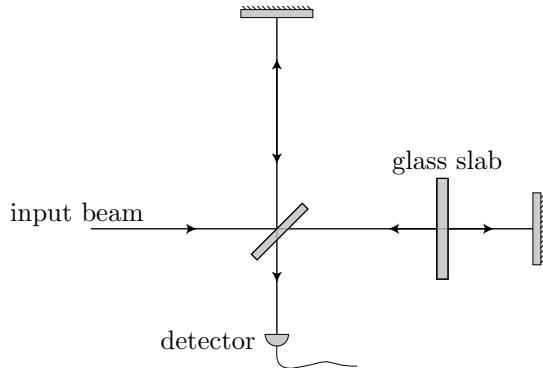


Suppose the output of the interferometer on the detector is at the maximum of a bright fringe (at refractive index n_0 of the EOM). By how much does the refractive index change so that the output goes to minimum intensity? (Compute the smallest change that does this.) Compute a numerical value for the refractive-index change for $d = 10 \text{ cm}$ and $\lambda = 1 \mu\text{m}$.

Incidentally, this is one method for making fast optical switches (as used in telecommunications, but these are usually Mach-Zehnder interferometers implemented with optical waveguides).

Problem 5.5

Consider a Michelson interferometer, operating at wavelength λ . A slab of glass of thickness d and refractive index n is inserted in one arm, parallel to the closest mirror as shown.



Derive an expression for the number of fringes the output goes through as the slab is rotated (clockwise in the diagram) through an angle θ .

Problem 5.6

Consider a Michelson interferometer, with mirrors at distances d_1 and d_2 from the 50/50 beam splitter. Suppose a Gaussian beam enters the interferometer (assume the beam and mirrors are all ideally aligned). Consider the intensity pattern on a screen at the output, where s is the distance from the beam splitter to the screen.

Show that there will be a “bull’s-eye” intensity pattern of concentric interference fringes on the screen, provided $d_1 \neq d_2$. Compare the fringe spacing near the center vs. near the edge of the pattern.

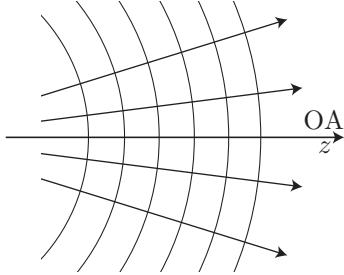
Hint: derive the expression for the intensity pattern due to two interfering waves of arbitrary intensity profiles $I_{1,2}$ and phase $\phi_{1,2}$, and then replace with Gaussian-beam expressions as appropriate. Try to focus only on the important stuff.

Chapter 6

Gaussian Beams

6.1 Paraxial Wave Equation

Just as paraxial rays greatly simplified ray optics by allowing us to use linear propagation equations, **paraxial waves** are much simpler to treat than waves in the general case. A paraxial wave is one where the curves normal to the wave fronts are paraxial rays.



Paraxial waves propagate (more or less) along the optical axis, so they should be fairly close to a plane wave. Thus, we can factor out the plane-wave part of the paraxial wave. For a scalar electric field propagating along the z -direction,

$$E^{(+)}(\mathbf{r}) = \psi(\mathbf{r})e^{ikz}. \quad (6.1)$$

Here, ψ is the **envelope** and e^{ikz} is the **carrier wave**. For a paraxial wave, the envelope ψ should be smooth on the scale of λ .

Plugging this into the wave (Helmholtz) equation, $(\nabla^2 + k^2)E^{(+)} = 0$, we find

$$\left(\nabla_{\tau}^2 + i2k \frac{\partial}{\partial z} + \frac{\partial^2}{\partial z^2} \right) \psi = 0, \quad (6.2)$$

where

$$\nabla_{\tau}^2 := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad (6.3)$$

is the **transverse Laplacian**. If ψ is smooth (i.e., varies slowly) on the scale of λ , then

$$\frac{\partial \psi}{\partial z} \ll \frac{\psi}{\lambda}, \quad (6.4)$$

which we can rewrite as

$$\frac{\partial \psi}{\partial z} \ll k\psi. \quad (6.5)$$

We can similarly write

$$\frac{\partial^2 \psi}{\partial z^2} \ll k \frac{\partial \psi}{\partial z}, \quad (6.6)$$

and thus we can neglect the $\partial^2/\partial z^2$ term in Eq. (6.2). This is called the **slowly varying envelope approximation**, and the resulting wave equation is the **paraxial wave equation**:

$$\left(\nabla_{\tau}^2 + i2k \frac{\partial}{\partial z} \right) \psi = 0. \quad (6.7)$$

(paraxial wave equation)

Remember that in treating the field in terms of ψ , we have already factored out the plane-wave part e^{ikz} of the phase.

It is interesting to note that this is exactly the same as the two-dimensional Schrödinger equation in quantum mechanics for the wave function $\psi(x, y, t)$:

$$-\frac{\hbar^2}{2m} \nabla_{\tau}^2 \psi + V(x, y) \psi = i\hbar \frac{\partial \psi}{\partial t}. \quad (\text{Schrödinger wave equation}) \quad (6.8)$$

We can see this equivalence by rewriting this equation as

$$\left(\nabla_{\tau}^2 - \frac{2m}{\hbar^2} V(x, y) + \frac{i2m}{\hbar} \frac{\partial}{\partial t} \right) \psi = 0, \quad (6.9)$$

and then making the formal identifications $z \rightarrow t$ and $k \rightarrow m/\hbar$. The potential $V(x, y)$ turns out to be related to the refractive index profile, though in a somewhat subtle way—the obvious correspondence is to the Schrödinger equation in free space.

6.2 Gaussian Beams

The simplest waves are the plane wave, which is of the form $\cos(kx - \omega t)$, and the spherical wave, which is of the form $\cos(kR - \omega t)/R$. These are useful models, so why would we want to study something more complicated like a Gaussian beam? The Gaussian beam is the simplest model of a *directed beam* that satisfies Maxwell's equations, at least in the paraxial approximation. It also turns out that the outputs of spherical-mirror resonators and lasers are often Gaussian beams. The Gaussian beam is just like the Gaussian wave packet in quantum mechanics, being just complex enough to be interesting but simple enough to be tractable.

Let's start by just writing it down (we will take it for granted that this satisfies the *paraxial* wave equation, at least until we study Fourier optics):

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} \frac{w_0}{w(z)} \exp \left[-\frac{r^2}{w^2(z)} \right] \exp \left[ikz - i \tan^{-1} \left(\frac{z}{z_0} \right) \right] \exp \left[ik \frac{r^2}{2R(z)} \right]. \quad (\text{Gaussian beam}) \quad (6.10)$$

This is the **Gaussian** or **TEM_{0,0}** (for “transverse electromagnetic”, as for the plane wave; the subscript comes from the hierarchy of Hermite–Gaussian beams, which we will come to later) beam. In this expression, we have used polar coordinates (r, z) with $r = \sqrt{x^2 + y^2}$; $E_0^{(+)}$ is as usual an overall field-amplitude constant; z_0 is a constant called the **Rayleigh length** (or “Rayleigh range”);

$$w_0 := \sqrt{\frac{\lambda z_0}{\pi}} \quad (6.11)$$

(beam waist parameter)

is the **beam waist parameter** or **beam radius**, (that's a “double-u,” not an “omega,” by the way), where the inverted relation

$$z_0 = \frac{\pi w_0^2}{\lambda} \quad (6.12)$$

(Rayleigh length in terms of w_0)

for the Rayleigh length in terms of the beam waist parameter is also useful;

$$w(z) := w_0 \sqrt{1 + \left(\frac{z}{z_0} \right)^2} \quad (6.13)$$

(beam waist along optical axis)

is the beam waist as a function of z ; and

$$R(z) := z \left[1 + \left(\frac{z_0}{z} \right)^2 \right] \quad (6.14) \quad (\text{wave-front radius of curvature})$$

is the radius of curvature of the wave fronts, as we will see.

This pretty much looks like a big mess, right? It's not nearly as bad as it looks, we just have to break it down and analyze each piece to see what is going on. First, let's try rewriting the Gaussian beam solution in a way that's partitioned into three smaller factors:

$$\begin{aligned} E^{(+)}(\mathbf{r}) &= E_0^{(+)} \frac{w_0}{w(z)} \exp \left[-\frac{r^2}{w^2(z)} \right] && (\text{amplitude factor}) \\ &\times \exp \left[ikz - i \tan^{-1} \left(\frac{z}{z_0} \right) \right] && (\text{longitudinal phase factor}) \\ &\times \exp \left[ik \frac{r^2}{2R(z)} \right]. && (\text{radial phase factor}) \end{aligned} \quad (6.15) \quad (\text{Gaussian beam, organized form})$$

We will now analyze each factor separately.

6.2.1 Amplitude Factor

The first factor in Eq. (6.15) is the amplitude factor, which completely describes the intensity profile of the Gaussian beam. In this sense, this is the most useful factor for understanding the behavior of the Gaussian beam. There are several important features to understand here:

Intensity Profile. The intensity is just the squared modulus of Eq. (6.15), but the other two factors make no contribution:

$$I(r, z) = I_0 \left[\frac{w_0}{w(z)} \right]^2 \exp \left[-\frac{2r^2}{w^2(z)} \right]. \quad (6.16) \quad (\text{intensity profile of Gaussian beam})$$

Here, we have defined $I_0 := 2|E_0^{(+)}|^2/\eta$ as the maximum intensity of the Gaussian beam (which occurs at the center $\mathbf{r} = 0$). Clearly, the intensity falls off in the radial direction like a Gaussian function, hence the name “Gaussian beam.” Also, note that $w(z)$ is one measure of the width of the Gaussian, and it reduces to the minimum value w_0 at $z = 0$. There are differing conventions for quoting the width of the beam (and even for what w_0 represents), so when using w_0 or $w(z)$ to represent the size of a Gaussian beam, for clarity it is good to call it the “ $1/e^2$ radius” of the beam.

Total Power. The total power of the Gaussian beam is given by integrating the intensity over a transverse plane:

$$P = \int_0^\infty I(r, z) 2\pi r dr = \frac{I_0}{2} (\pi w_0^2). \quad (6.17) \quad (\text{total power of Gaussian beam})$$

This result holds independent of z . Thus we can rewrite the intensity profile as

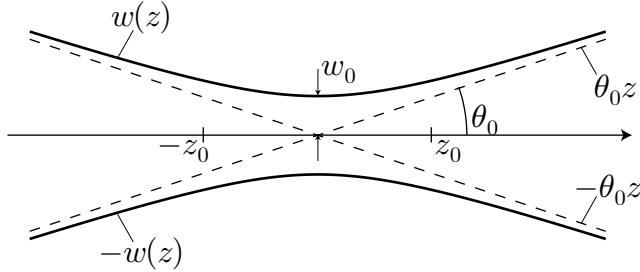
$$I(r, z) = \frac{2P}{\pi w^2(z)} \exp \left[-\frac{2r^2}{w^2(z)} \right]. \quad (6.18) \quad (\text{intensity profile in terms of total power})$$

It is useful to consider some similar integrals over a disc of finite radius, as a measure of how “large” is a Gaussian beam. For example, a circle of radius $R = w(z)$ contains 86% of the optical power of a Gaussian beam. A circle of radius $R = 1.5w(z)$ contains 99% of the power of a Gaussian beam.

Beam Divergence. Note that the beam waist $w(z)$ traces out a hyperbolic curve in z . Near the focus at $z = 0$, the waist achieves its minimum value w_0 . At large distances from the focus, the hyperbola approaches its asymptotes, given by $(w_0/z_0)z$. The Rayleigh length z_0 marks the crossover between these two regimes; thus,

$$\begin{aligned} w(z) &\sim w_0 && \text{for } |z| \ll z_0 \\ w(z) &\sim \left(\frac{w_0}{z_0}\right)z && \text{for } |z| \gg z_0 \end{aligned} \quad (6.19)$$

for the regions near to and far away from the focus, respectively.



From this we see that in the far field, the beam propagates in the form of a cone of half angle θ_0 , where

$$\tan \theta_0 := \frac{w_0}{z_0} = \frac{\lambda}{\pi w_0}. \quad (6.20)$$

In the paraxial regime, we have

$$\theta_0 \approx \frac{w_0}{z_0} = \frac{\lambda}{\pi w_0}. \quad (6.21) \quad (\text{far-field divergence half-angle})$$

Thus, a *tightly focused* beam (small w_0) diverges *more quickly*, whereas a larger beam (large w_0) diverges less. This is a manifestation of the uncertainty principle in optics. Also, a larger wavelength causes more far-field divergence: generally speaking, diffraction effects are more important for longer wavelengths.

As for the intensity, in the far field we have

$$I(r = 0, z) \approx I_0 \left(\frac{z_0}{z}\right)^2, \quad (6.22)$$

so the Gaussian beam is consistent with the inverse-square intensity law in the far field (as is also the case for the spherical wave).

Depth of Focus. As we just mentioned, the best focus occurs at $z = 0$. At a distance of one Rayleigh length z_0 from the focus, the beam waist increases by 41% compared to the focal value:

$$w(\pm z_0) = \sqrt{2} w_0. \quad (6.23)$$

Again, z_0 marks the transition from focused behavior to far-field divergence. So we can say that the beam is roughly focused between $\pm z_0$, so the “depth of focus” for a Gaussian beam is $2z_0$, where again

$$2z_0 = \frac{2\pi w_0^2}{\lambda}. \quad (6.24) \quad (\text{depth of focus})$$

From this relation we can see explicitly that tightly focused beams have a smaller depth of focus. This is related to a general principle in imaging optics, which is that larger apertures produce smaller depths of focus (as occurs in photography).

6.2.2 Longitudinal Phase Factor

Recall that the longitudinal phase factor has the form

$$\exp \left[ikz - i \tan^{-1} \left(\frac{z}{z_0} \right) \right]. \quad (6.25)$$

(longitudinal phase factor)

The first term in the phase is simply the phase of a plane wave ikz propagating in the same direction and with the same optical frequency as the Gaussian beam. The second term is called the **Gouy phase shift**¹ and represents a small departure from planarity. The longitudinal phase is dominated by the ikz term, but don't go underestimating the importance of the Gouy term. It represents a phase *retardation* compared to the plane wave (in that the total phase here changes more *slowly* than in the plane wave). The \tan^{-1} form implies a monotonically decreasing phase shift (relative to the plane-wave part), amounting to a total of $-\pi$ change in phase over all z . Gouy effects are generic to focusing-beam-type solutions to the wave equation. As we will see, the Gouy phase is important in computing the resonant frequencies of optical resonators. Also, it implies that the phase velocity of a Gaussian beam is slightly larger than c , since the spacing between wave fronts is slightly larger than λ .

6.2.3 Radial Phase Factor

Recall that the radial phase factor is of the form

$$\exp \left[ik \frac{r^2}{2R(z)} \right], \quad (6.26)$$

(radial phase factor)

where

$$R(z) = z \left[1 + \left(\frac{z_0}{z} \right)^2 \right]. \quad (6.27)$$

(radius of curvature)

This factor gives the dependence of the phase on r , whereas the longitudinal phase factor simple gave the on-axis phase behavior. Thus the entire phase factor is

$$\exp \left[ikz - i \tan^{-1} \left(\frac{z}{z_0} \right) + ik \frac{r^2}{2R(z)} \right]. \quad (6.28)$$

(total Gaussian phase)

To understand this, we need to compare this phase to the phase of a spherical wave,

$$E^{(+)}(\mathbf{r}) = \frac{E_0^{(+)}}{R} e^{ikR} = \frac{E_0^{(+)}}{R} e^{ik\sqrt{r^2+z^2}}, \quad (6.29)$$

where $R = |\mathbf{r}|$. In the paraxial case, $r \ll z$, so

$$\begin{aligned} E^{(+)}(\mathbf{r}) &= \frac{E_0^{(+)}}{R} \exp \left[ik|z| \sqrt{1 + \frac{r^2}{z^2}} \right] \\ &\simeq \frac{E_0^{(+)}}{R} \exp \left[ik|z| \left(1 + \frac{r^2}{2z^2} \right) \right] \\ &= \frac{E_0^{(+)}}{R} \exp \left[ik|z| + ik \frac{r^2}{2|z|} \right] \\ &\simeq \frac{E_0^{(+)}}{R} \exp \left[ik|z| + ik \frac{r^2}{2R} \right]. \end{aligned} \quad (6.30)$$

¹C. R. Gouy, "Sur une propriété nouvelle des ondes lumineuses," *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris)* **110**, 1251 (1890); C. R. Gouy, "Sur la Propagation Anomale Des Ondes," *Annales de Chimie et de Physique, 6e série* **24**, 145 (1891).

Comparing this phase to Eq. (6.28), we see that in the paraxial approximation, the spherical wave and Gaussian beam have the same form except for the absolute value of z , which accounts for the fact that the spherical wave propagates outward from $R = 0$; the Gouy phase, which represents only a small correction (and in some sense “patches” together the wave fronts on either side of the focus, due to the reversal of half of the wave fronts as compared to the spherical wave); and the fact that the radius of curvature changes with z for the Gaussian beam. Thus, $R(z)$ in the Gaussian beam represents the **radius of curvature** of the wave fronts.

Examining the radius of curvature, we see that near and far away from the focus,

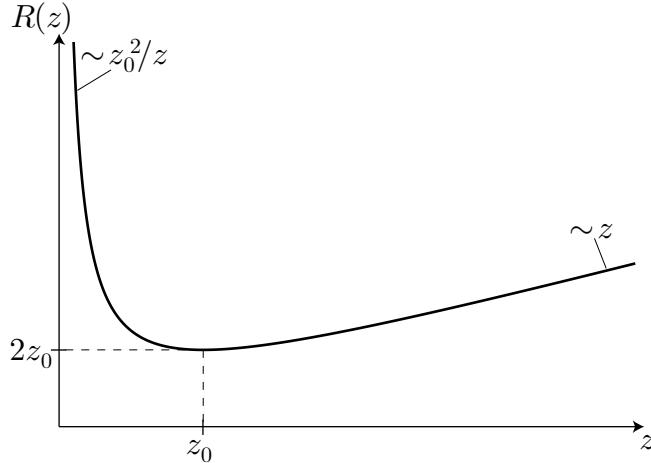
$$\begin{aligned} R(z) &\sim \frac{z_0^2}{z} \quad \text{for } |z| \ll z_0 \\ R(z) &\sim z \quad \text{for } |z| \gg z_0 \end{aligned} \tag{6.31}$$

Again, $z \approx R$ in the paraxial approximation, so the *Gaussian beam approaches a spherical wave* for large $|z|$.

We can also see where $R(z)$ is minimum (i.e., the curvature is maximum) by differentiating R :

$$\frac{dR}{dz} = \frac{d}{dz} \left(z + \frac{z_0^2}{z} \right) = 1 - \frac{z_0^2}{z^2}. \tag{6.32}$$

Thus, $dR/dz = 0$ when $z = \pm z_0$, so maximum wave-front curvature occurs at the Rayleigh length. Closer to the focus, the wave fronts become flat (as we could guess from symmetry considerations), and farther out, the wave-front curvature decreases as for a spherical wave. The maximum radius of curvature is $R(\pm z_0) = 2z_0$.



In a cavity, the boundary conditions imposed by the cavity mirrors require that the curvature of the spherical mirrors and the curvature of the wave fronts match. This allows the wave to map back on itself. This is one reason why Gaussian beams can exist in resonators. But in particular, recall that for a confocal cavity, the length d of the cavity is the radius of curvature of both mirrors $R_{1,2}$. A Gaussian beam with $2z_0 = |R_{1,2}|$ has wave fronts that exactly match the mirror curvatures at $z = \pm z_0$. Hence, z_0 is also referred to as the **confocal parameter**.

6.3 Specification of Gaussian Beams

Despite the complex form of the Gaussian beam, relatively little information is needed to completely specify it. For example, if we know where $z = 0$ is located, the value of w_0 , and the optical wavelength λ all the other parameters are uniquely fixed. Alternately, it is sufficient to know w_0 and $R(z)$ at some distance z , or it is sufficient to know $w(z)$ and $R(z)$ at some distance z . From a Fourier-transform point of view, this is because a Gaussian wave packet is uniquely determined by its center and width. By construction of the Gaussian beam, the center is zero. The divergence is set by the width of the Fourier transform of the wave packet, which we know is fixed via the uncertainty principle. Again, the tradeoff between tighter focus and larger divergence angle is precisely the manifestation of the uncertainty principle in the Gaussian beam.

6.4 Vector Gaussian Beams

We have thus far only treated Gaussian beams as solutions to the scalar wave equation. It turns out that for the full vector electromagnetic field case, the only difference is in the overall amplitude. That is, supposing we have a Gaussian beam propagating along the z -direction and polarized in the x -direction, we can obtain the vector field from the scalar field by the simple replacement

$$E_0^{(+)} \longrightarrow E_0^{(+)} \left(-\hat{x} + \frac{x}{z - iz_0} \hat{z} \right).$$

(modification to obtain vector Gaussian beam, linear polarization) (6.33)

Note that the polarization is mostly along the x -direction in the region where the paraxial approximation holds, since $|x| \ll |z + iz_0|$. Thus, the field is mostly transverse. Optical beams are *nearly* transverse (TEM), but the transverse property is not completely necessary (as it is for the plane wave), as long as $\nabla \cdot \mathbf{E} = 0$. Writing out this condition,

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0. \quad (6.34)$$

The first two terms represent a *transverse divergence*, so that

$$\nabla_T \cdot \mathbf{E} + \partial E_z / \partial z = 0. \quad (6.35)$$

Putting in approximate numbers,

$$\frac{E_T}{2w_0} + kE_z = 0, \quad (6.36)$$

where E_T is the transverse field $\sqrt{E_x^2 + E_y^2}$, and $2w_0$ is the beam diameter. Thus, for these terms to cancel, they must be on the same order, and thus

$$\left| \frac{E_z}{E_T} \right| \sim \frac{1}{2kw_0} = \frac{\lambda}{4\pi w_0}. \quad (6.37)$$

So as long as the beam intensity falls off on a length scale large compared to λ , which happens for a Gaussian beam if $w_0 \gg \lambda$, the field is mostly transverse (i.e., E_z is small). In the paraxial approximation, the field is transverse to good approximation, justifying the label of TEM_{0,0} for the Gaussian beam.

6.5 ABCD Law

The **ABCD Law** is a quick method for propagating Gaussian beams. We'll begin by just stating it, and then we'll justify it later.

It is convenient to define a complex q parameter by

$$q(z) := z - iz_0. \quad (6.38) \quad (\text{q parameter for Gaussian beams})$$

It is worth reiterating that the longitudinal coordinate z is measured with respect to the beam *focus*. Note that q determines the Rayleigh length and the location of the beam focus, and therefore determines the entire geometry of the Gaussian beam. Alternately, we can compute the inverse of q ,

$$\frac{1}{q} = \frac{1}{z - iz_0} = \frac{z + iz_0}{z^2 + z_0^2} = \frac{z}{z^2 + z_0^2} + i \frac{z_0}{z^2 + z_0^2}. \quad (6.39)$$

Recalling that $w_0 = \sqrt{\lambda z_0 / \pi}$, $w(z) = w_0 \sqrt{1 + (z/z_0)^2}$, and $R(z) = z[1 + (z_0/z)^2]$, we can write $1/q$ in the more useful form

$$\frac{1}{q(z)} = \frac{1}{R(z)} + \frac{i\lambda}{\pi w^2(z)}. \quad (6.40) \quad (\text{inverse of q parameter})$$

Again, q determines both the spot size and the radius of curvature at a given location, so q completely determines the Gaussian beam geometry.

We can see that this simplifies things greatly by rewriting the Gaussian beam in a compact form in terms of q :

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} \frac{q_0}{q(z)} \exp\left[\frac{ikr^2}{2q(z)}\right] \exp(ikz). \quad (\text{Gaussian beam in terms of } q \text{ parameter}) \quad (6.41)$$

Here, we have used the notation $q_0 := q(0) = -iz_0$. The Gouy phase ends up being just the phase angle of $1/q$, while q also represents the width and curvature of the beam.

Here is the other nice thing about q : it transforms through an optical system according to a simple law related to the *geometrical optics* $ABCD$ matrix for the system. In terms of the matrix elements, the $ABCD$ law reads

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D}, \quad (\text{ABCD Law}) \quad (6.42)$$

where q_1 and q_2 are the q parameters before and after the optical system, respectively. or in alternate form,

$$\frac{1}{q_2} = \frac{C + D/q_1}{A + B/q_1}. \quad (\text{ABCD Law}) \quad (6.43)$$

Let's check this for a couple of simple examples, and then use them to argue that the $ABCD$ law is generally valid.

6.5.1 Free-Space Propagation

Recall that the matrix for propagation over distance d in free space is

$$\mathbf{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}. \quad (6.44)$$

Thus, applying the $ABCD$ law gives

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D} = q_1 + d. \quad (6.45)$$

Writing out the real and imaginary parts explicitly,

$$z_2 - iz_{02} = (z_1 + d) - iz_{01}. \quad (6.46)$$

The imaginary part of this equation states that the Rayleigh length is unchanged, $z_{02} = z_{01}$, and thus the shape of the beam is unaffected. The real part states that $z_2 = z_1 + d$, that is, the position along the beam is shifted by d , as we expect for a propagation of the same distance. Thus, the $ABCD$ law gives a sensible result.

Note that the argument carries through for a dielectric medium of refractive index n , by repeating the argument with the replacement $d \rightarrow d/n$. However, this point is a bit subtle: propagation over a distance d through a medium of refractive index n is described by the *same* matrix as in free space, independent of n . However, propagation through a refractive *slab* of thickness d (surrounded by vacuum) is given by the composite matrix for two planar, refractive interfaces separated by propagation over distance d :

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & n \end{bmatrix} \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/n \end{bmatrix} = \begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix} \quad (6.47)$$

Again, *this* matrix works for the Gaussian beam, since the propagation distance is effectively reduced by n , due to the refraction at the boundaries modifying the divergence angle.

6.5.2 Thin Optic

Recall that for a thin optic, the position of the ray does not change ($y_2 = y_1$). Thus, the matrix for a thin optic has the general form

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ C & D \end{bmatrix}. \quad (6.48)$$

This condition is also equivalent to the condition that the Gaussian beam waist does not change across the optic ($w_{02} = w_{01}$), so that power density on each side of the optic is consistent.

In ray optics, the matrix affects the transmitted angle,

$$y'_2 = Cy_1 + Dy'_1. \quad (6.49)$$

In the paraxial approximation, the Gaussian beam is equivalent to a spherical wave of radius R_1 or R_2 before or after the optic, respectively. For the (paraxial) rays normal to the wave fronts, we can write

$$y'_{1,2} \approx \frac{y_{1,2}}{R_{1,2}}. \quad (6.50)$$

Putting this into Eq. (6.49), we find

$$\frac{1}{R_2} = C + \frac{D}{R_1}. \quad (6.51)$$

Using $1/R = 1/q - i\lambda/\pi w^2$ and $D = \det(\mathbf{M}) = n_1/n_2 = \lambda_2/\lambda_1$, we find

$$\frac{1}{q_2} = C + \frac{D}{q_1} \implies q_2 = \frac{q_1}{Cq_1 + D}, \quad (6.52)$$

so we recover the *ABCD* law in this restricted case.

6.5.3 Cascaded Optics

As in ray optics, the *ABCD* law works for cascaded (composite) systems. It is possible to verify directly by substitution that if

$$q_3 = \frac{A_2 q_2 + B_2}{C_2 q_2 + D_2} \quad \text{and} \quad q_2 = \frac{A_1 q_1 + B_1}{C_1 q_1 + D_1}, \quad (6.53)$$

then

$$q_3 = \frac{A q_1 + B}{C q_1 + D} \quad (6.54)$$

provided that the new matrix elements are obtained by matrix multiplication in the proper order:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix}. \quad (6.55)$$

Thus, we can handle Gaussian-beam propagation through any composite system via the composite ray matrix.

6.5.4 Factorization of a General Matrix

We showed that the *ABCD* law works in the case of propagation in free space (and thus also a homogenous medium), as well as for a thin optic. We now know that the *ABCD* law works for any cascade of propagations and thin optics. Even for continuous media, we can always in principle break any medium up into a cascade of infinitesimal propagations and weak, thin optics. In principle, the *ABCD* law should then work for *any* optical system.

We can argue this more precisely without so much hand-waving, however. Consider a general ray matrix,

$$\mathbf{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}. \quad (6.56)$$

We can write down an explicit factorization of this matrix in terms of propagation and thin-optic matrices. Consider a system comprising (1) a propagation in free space over distance d_1 ; (2) a thin lens of focal length f ; (3) a planar refractive interface to a homogenous medium of index n ; and finally (4) a propagation over distance d_2/n in the medium. The ray matrix is

$$\mathbf{M}' = \begin{bmatrix} 1 & d_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/n \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \begin{bmatrix} 1 & d_1 \\ 0 & 1 \end{bmatrix}. \quad (6.57)$$

Then these two systems are equivalent ($\mathbf{M} = \mathbf{M}'$) provided that

$$\begin{aligned} n &= \frac{1}{\det(\mathbf{M})} \\ f &= -\frac{\det(\mathbf{M})}{C} \\ d_1 &= \frac{D - \det(\mathbf{M})}{C} \\ d_2 &= \frac{A - 1}{C} \end{aligned} \quad (6.58)$$

We have already verified that the *ABCD* law works for any of the component matrices, thus by the cascading rule it works for the matrix \mathbf{M} .

6.5.5 “Deeper Meaning” of the *ABCD* Law

It seems amazing that a solution to the wave equation can be propagated using *classical* rules. But really what this is saying is that, in some sense, wave phenomena are absent in the paraxial approximation.

As we will see, this transformation rule applies in more general situations as long as the paraxial approximation holds. But let's focus on the Gaussian beam for concreteness. The Gaussian beam stays Gaussian through any optical system as long as the paraxial approximation is valid. As soon as the paraxial approximation breaks down (i.e., nonlinear terms become important), the beam will become non-Gaussian. For example, one would not expect a beam in an unstable resonator to remain Gaussian. However, within the paraxial approximation, a *bundle of geometrical rays* with a Gaussian density profile (in both y and y' , in a way that satisfies the uncertainty principle) follows the Gaussian beam exactly. Showing this mathematically really requires the Wigner transform, so we won't get into this here. But the important point is that Gaussian beam optics is really just ray optics; Gaussian beams do not show any manifestly wave-like behavior, at least within the paraxial approximation. All the diffraction-type effects can be mimicked by an appropriate ray ensemble.

Again, since paraxial wave propagation is equivalent to quantum mechanics, this is equivalent to the relation between quantum and classical mechanics for linear systems. For linear systems (i.e., free-space propagation and the harmonic oscillator), quantum and classical mechanics turn out to be equivalent. That is, you can attribute any quantum effects in a linear system to the *initial condition*, not to the time evolution. Nonlinear systems dynamically generate quantum effects, even for a classical initial state.

6.5.6 Example: Focusing of a Gaussian Beam by a Thin Lens

As a more useful example of the *ABCD* law in action, let's consider the focusing of a Gaussian beam by a lens. We will consider the special case where the lens is placed at the waist of the beam to be focused, so that the initial radius of curvature $R_1 = \infty$.

The ray matrix is

$$\begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix}, \quad (6.59)$$

and the initial beam parameter is, starting from Eq. (6.40),

$$\frac{1}{q_1} = \frac{1}{R_1} + \frac{i\lambda}{\pi w_1^2} = \frac{i\lambda}{\pi w_{01}^2} = \frac{i}{z_{01}}, \quad (6.60)$$

where we recall the initial Rayleigh length is $z_{01} = \pi w_{01}^2/\lambda$. Remember that we can *only* represent the q parameter simply in terms of the Rayleigh length at the focus of the beam. Applying the *ABCD* law,

$$\frac{1}{q_2} = \frac{C + D/q_1}{A + B/q_1} = -\frac{1}{f} + i \frac{\lambda}{\pi w_{01}^2}. \quad (6.61)$$

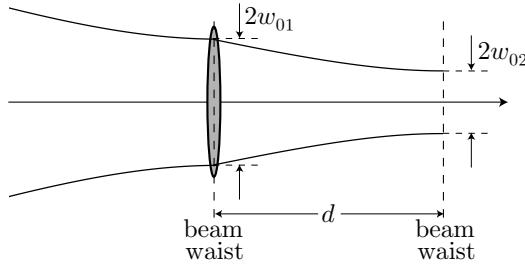
We should then compare this relation to the general form (6.40) for $1/q$:

$$\frac{1}{q_2} = \frac{1}{R_2} + \frac{i\lambda}{\pi w_2^2}. \quad (6.62)$$

Equating the real and imaginary parts, we find that the radius of curvature at the output is $R_2 = -f$, corresponding to a focusing beam if $f > 0$. Also, $w_2 = w_{01}$, which means that the beam size is unchanged—again, as we expect for any thin optic.

6.5.7 Example: Minimum Spot Size by Lens Focusing

We can do a similar analysis to compute the minimum spot size obtained by a focusing lens for a Gaussian beam. The setup is the same as for the last example, but we will also consider a free-space propagation over a distance d after the lens.



Thus, the system matrix is

$$\mathbf{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{f} & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} = \begin{bmatrix} 1 - \frac{d}{f} & d \\ -\frac{1}{f} & 1 \end{bmatrix}. \quad (6.63)$$

Applying the *ABCD* law,

$$\frac{1}{q_2} = \frac{C + D/q_1}{A + B/q_1} = \frac{-1 + if/z_{01}}{f - d + idf/z_{01}}. \quad (6.64)$$

Separating the real and imaginary parts,

$$\frac{1}{q_2} = \frac{-(f-d) + df^2/z_{01}^2}{(f-d)^2 + (fd/z_{01})^2} + i \frac{(f-d)(f/z_{01}) + df/z_{01}}{(f-d)^2 + (fd/z_{01})^2}. \quad (6.65)$$

From Eq. (6.40), we can equate the real and imaginary parts of this expression with $1/R_2$ and $1/z_{02}$, respectively (recall that we can only make the latter identification under the assumption that we are at the focus). Thus,

$$R_2 = \frac{(f-d)^2 + (fd/z_{01})^2}{-(f-d) + df^2/z_{01}^2}. \quad (6.66)$$

But plane 2 is at the beam waist, so $R_2 = \infty$ and thus the denominator of this expression must vanish:

$$-(f-d) + df^2/z_{01}^2 = 0. \quad (6.67)$$

Solving this for the distance, we find

$$d = \frac{f}{1 + f^2/z_{01}^2}. \quad (6.68)$$

(focus distance for Gaussian beam)

Note that the focus is *not* located a distance f away as in classical optics—the focus occurs at a slightly smaller distance. We only recover the geometrical result $d = f$ in the limit of a large input beam, $z_{01} \rightarrow \infty$ (i.e., the plane-wave limit). Roughly speaking, a beam with a finite z_{01} (or w_{01}) is still “like” a converging beam at the focus. This isn’t really a large effect. For some sample numbers, let’s consider $f = 0.5$ m, $\lambda = 2\pi \times 10^{-7}$ m (close to a He-Ne laser line), and $w_{01} = 1$ cm. In this case, $z_{01} = \pi w_{01}^2/\lambda = 500$ m. Since $z_{01} \gg f$, the effect is very small (only a 1 ppm reduction in the focal distance).

Now we can come to the main point, the minimum spot size. The imaginary part of Eq. (6.65) gives

$$z_{02} = \frac{\pi w_{02}^2}{\lambda} = \frac{(f-d)^2 + (fd/z_{01})^2}{(f-d)(f/z_{01}) + fd/z_{01}}. \quad (6.69)$$

We will consider the limit $z_{01} \gg f$, which simplifies the algebra here, and is the most important case in practical optical systems when trying to obtain a small spot size (i.e., using a lens to focus a large, collimated beam, where the large spot size implies a correspondingly large Rayleigh length). Then $d \approx f$ and

$$z_{02} \approx \frac{fd}{z_{01}} \approx \frac{f^2}{z_{01}}. \quad (6.70)$$

Thus,

$$w_{02} \approx \frac{\lambda f}{\pi w_{01}} \approx \frac{3}{\pi} \lambda (f/\#), \quad (6.71)$$

(ideal minimum spot size)

where $f/\#$ (the “ f -number”) is given by

$$(f/\#) := \frac{f}{D}, \quad (6.72)$$

(definition of $f/\#$)

with D denoting the diameter of the lens. In writing down this expression, we are assuming a diameter of $2 \cdot 1.5 \cdot w_{01}$, which passes 99% of the total incident beam power.

This calculation is for an ideal lens, where we can ignore aberrations. A simple approximate correction to account for aberrations² is to multiply this result by 4/3:

$$w_{02} \approx \frac{4}{\pi} \lambda (f/\#). \quad (6.73)$$

(more realistic minimum spot size)

However, this is a heuristic formula based on experience with typical lenses, and not justifiable on fundamental grounds. (In particular, spherical aberrations should lead to larger corrections for larger $(f/\#)$ and thus smaller spot sizes.)

6.6 Hermite–Gaussian Beams

If we don’t require axial symmetry, we can obtain more general solutions to the paraxial wave equation. The **Hermite–Gaussian modes** form one possible set of solutions

$$\begin{aligned} E_{l,m}^{(+)}(\mathbf{r}) &= E_0^{(+)} \frac{w_0}{w(z)} \sqrt{\frac{1}{2^{l+m} l! m!}} H_l \left[\frac{\sqrt{2}x}{w(z)} \right] H_m \left[\frac{\sqrt{2}y}{w(z)} \right] \\ &\times \exp \left[-\frac{r^2}{w^2(z)} \right] \exp \left[ikz - i(1+l+m) \tan^{-1} \left(\frac{z}{z_0} \right) \right] \exp \left[ik \frac{r^2}{2R(z)} \right]. \end{aligned} \quad (6.74)$$

(TEM _{l,m} mode)

²As recommended by the Melles-Griot catalog.

These are also referred to as **TEM_{*l,m*} modes**. Here, a **mode** refers to field profiles that, when confined in resonators (as we will see) have profiles that are time-invariant. (So far, we are only considering propagation in free space, in which case the analogous property is that the intensity profiles of these modes are invariant along the beam except for an overall scaling, which is described by $w(z)$ here.)

To facilitate comparison with the Gaussian beam, we can write the general Hermite–Gaussian mode in terms of the simple Gaussian beam as

$$E_{l,m}^{(+)}(\mathbf{r}) = E_{\text{Gauss}}^{(+)}(\mathbf{r}) \sqrt{\frac{1}{2^{l+m} l! m!}} H_l \left[\frac{\sqrt{2}x}{w(z)} \right] H_m \left[\frac{\sqrt{2}y}{w(z)} \right] \exp \left[-i(l+m) \tan^{-1} \left(\frac{z}{z_0} \right) \right], \quad (\text{TEM}_{l,m} \text{ mode in terms of Gaussian beam}) \quad (6.75)$$

where $E_{\text{Gauss}}^{(+)}(\mathbf{r})$ is the Gaussian beam from Eq. (6.10). The differences compared to the Gaussian beam are: (1) the presence of the **Hermite polynomials** $H_n(z)$; (2) the normalization factor just before the Hermite polynomials, chosen so that the total beam power for a given field $E_0^{(+)}$ is the same for all l and m ; and (3) the factor of $(1 + l + m)$ in the Gouy phase.

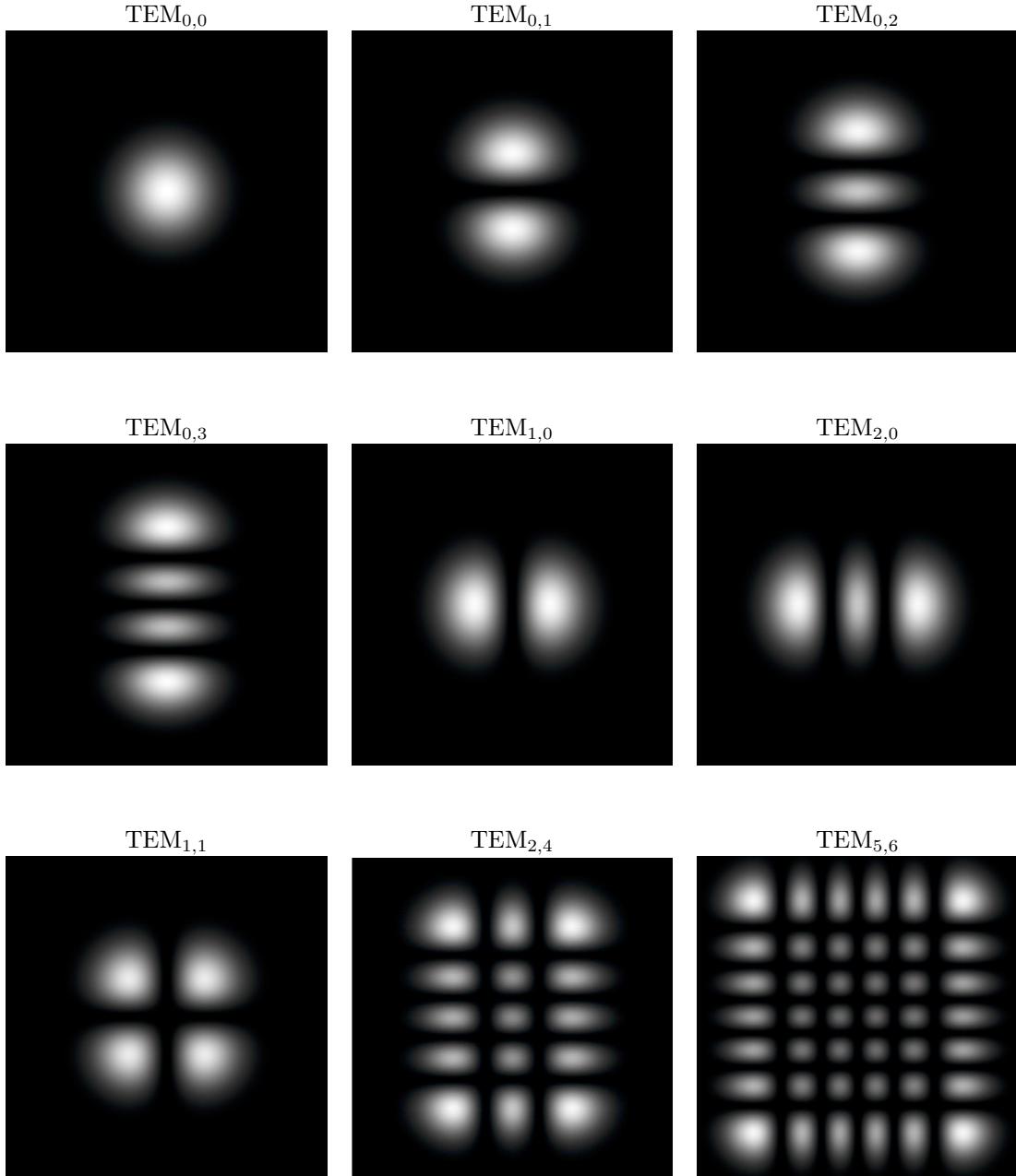
The Hermite polynomial $H_n(z)$ is a polynomial of degree n . We can write them out explicitly as

$$\begin{aligned} H_0(z) &= 1 \\ H_1(z) &= 2z \\ H_2(z) &= 2(2z^2 - 1) \\ H_n(z) &= (-1)^n e^{z^2} \frac{d^n}{dz^n} e^{-z^2} \quad (\text{explicit formula}) \\ H_{n+1}(z) &= 2zH_n(z) - 2nH_{n-1}(z). \quad (\text{recurrence relation}) \end{aligned} \quad (6.76)$$

For our purposes here, it is most important to know that $H_n(z)$ is an n th degree polynomial with n zeros on the real axis. Note that since $H_0(z) = 1$, we see that for $l = m = 0$ we recover the Gaussian beam, as is consistent with our earlier notation. Hence, the Gaussian beam is also called the TEM_{0,0} beam. In general, the intensity pattern of the TEM_{*l,m*} mode has l dark bands across the x -direction and m dark bands across the y -direction. Alternately, the intensity pattern has a grid of $l+1$ bright spots in the x -direction and $m+1$ in the y -direction. Note that the Hermite polynomials diverge for large $|x|$ or $|y|$. However, the Gaussian factor always cuts off the intensity far away from the optical axis, so the intensity is still concentrated (relatively) close to the optical axis.

Like the Gaussian beam, the Hermite–Gaussian beams have intensity patterns that stay the same as they propagate in space, except for the overall scale. They have the same far-field divergence characteristics as for the Gaussian beam (except for the different spatial profile, which again is constant up to stretching over the entire extent of the beam). The higher-order modes are wider for the same value of w_0 . However, all the same transformation rules apply to the Hermite–Gaussian beams, including the *ABCD* law.

Again, there is an important connection with quantum mechanics. The Hermite–Gaussian beams are equivalent to the eigenfunctions of the quantum harmonic oscillator. Thus, the Hermite–Gaussian modes form a complete set—that is, *any* beam can be represented as a linear combination of Hermite–Gaussian modes, and its spatial evolution is a manifestation of the interferences between the component modes.

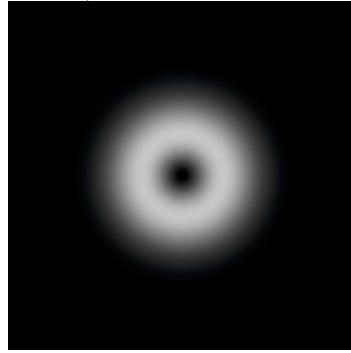


6.6.1 Doughnut and Laguerre–Gaussian Modes

Before finishing, it is worth mentioning one more beam, the equal superposition of $\text{TEM}_{0,1}$ and $\text{TEM}_{1,0}$ beams. This beam has an axially symmetric intensity pattern, where the intensity vanishes at the origin. This beam is called the **doughnut mode** or $\text{TEM}_{0,1}^*$ mode. This mode is important in describing the output of lasers as the dominant component for slightly distorted (but still axially symmetric) Gaussian

beams.

$$\text{TEM}_{0,1}^* = \text{TEM}_{0,1} + \text{TEM}_{1,0}$$



The doughnut mode is a particular example of another class of beams, the **Laguerre–Gaussian beams**, where the general form is given by ³

$$E_{l,m}^{(+)}(\mathbf{r}) = E_{\text{Gauss}}^{(+)}(\mathbf{r}) \sqrt{\frac{2l!}{\pi(l+|m|)!}} \left(\frac{\sqrt{2}r}{w(z)} \right)^{|m|} L_l^{|m|} \left[\frac{2r^2}{w^2(z)} \right] \exp \left[-i(2l+|m|) \tan^{-1} \left(\frac{z}{z_0} \right) \right] e^{-im\phi},$$

($\text{TEM}_{l,m}^*$ mode in terms of Gaussian beam) (6.77)

where $l \geq 0$ and $-l \leq m \leq l$. These beams are radially symmetric in intensity, so that the only azimuthal dependence (on the azimuthal angle ϕ) is in the phase. These are the analogous solutions to the paraxial wave equation to the Hermite–Gaussian beams, but when separated in cylindrical rather than Cartesian coordinates.

The polynomials $L_l^{|m|}(z)$ here are the **associated Laguerre polynomials**

$$\begin{aligned} L_0^m(z) &= 1 \\ L_1^m(z) &= -z + m + 1 \\ L_2^m(z) &= \frac{z^2}{2} - (m+2)z + \frac{(m+2)(m+1)}{2} \\ L_n^m(z) &= \frac{z^{-m}e^z}{n!} \frac{d^n}{dz^n} (e^{-z} z^{n+m}) \end{aligned} \quad (\text{explicit formula}), \quad (6.78)$$

here defined for $m \geq 0$. Again, $L_l^{|m|}(z)$ is a polynomial of degree l , and thus has l zeros. Here, these are manifested in increasing numbers of dark circles or dark spots at the center (for odd l) for higher-order beams.

³For an operator-based derivation of both the Laguerre–Gaussian and Hermite–Gaussian beams, see Francesco Pampaloni and Jörg Enderlein, “Gaussian, Hermite-Gaussian, and Laguerre-Gaussian beams: A primer,” arXiv.org preprint physics/0410021 (2004).

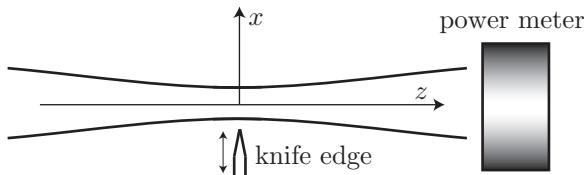
6.7 Exercises

Problem 6.1

A common technique in the laboratory for measuring the beam waist parameter of a Gaussian beam is illustrated in the diagram. A Gaussian beam is incident on an optical power meter, which registers the total power of the incident beam. A knife edge can be translated in the transverse direction to block part of the beam (i.e., if the position of the knife edge is x_{knife} , then the parts of the beam with $x < x_{\text{knife}}$ is blocked from reaching the power meter). The “10-90” rule is to measure the knife edge position $x_{10\%}$ where the power meter reads 10% of the total beam power, and then the position $x_{90\%}$ where the power meter reads 90% of the total beam power. Then the beam radius $w(z)$ at the knife-edge location z along the beam is given by

$$w(z) = \alpha |x_{10\%} - x_{90\%}|, \quad (6.79)$$

where α is some constant factor. Calculate the numerical value of α .



Problem 6.2

Suppose you have a 500 mW, $\text{TEM}_{0,0}$ (i.e., Gaussian) beam from a continuous-wave Ti:sapphire laser ($\lambda = 850 \text{ nm}$). The measured beam radius at the output coupler (mirror) is $w = 0.7 \text{ mm}$, and the beam radius in the laser rod is $w = 330 \mu\text{m}$.

- (a) What is the optical distance between the output coupler and the laser rod? Assume that the length of the rod is negligible (it is actually around 20 mm) and that the focus of the beam occurs in the center of the rod.
- (b) How far will this beam propagate (past the output coupler) before the spot size is 1 cm?
- (c) What is the radius of curvature of the phase front at the distance you calculated in (b)?
- (d) What is the amplitude of the electric field at the center of the focus? Assume a refractive index of $n = 1.75$ for sapphire.

Problem 6.3

Recall that the Gaussian ($\text{TEM}_{0,0}$) beam is a solution to the wave equation in the paraxial approximation. Also recall that the Gaussian beam is specified completely in terms of relatively few parameters.

- (a) What are the parameters?
- (b) State (general) conditions on the parameters that guarantee the paraxial approximation is valid.

Problem 6.4

Show that within the paraxial approximation, a thin lens converts a plane wave into a spherical wave. Use the following procedure.

- (a) Use Fermat's principle to show that a thin lens of focal length f introduces an optical delay with an equivalent path length given by

$$\ell_{\text{lens}}(y) = d - \frac{y^2}{2f}, \quad (6.80)$$

where d is an arbitrary length constant and y is as usual the normal distance from the optical axis.

- (b) A spherical wave centered on the optical axis propagating *inward* has the form

$$E^{(+)}(R) = \frac{E_0^{(+)}}{R} \exp(-ikR), \quad (6.81)$$

where $R^2 = y^2 + (z - z_0)^2$ is the radial distance from the center of the wave at $(y, z) = (0, z_0)$. Note that we are suppressing the x -dependence for simplicity, but it isn't difficult to extend the argument to the full three-dimensional case.

Take $z_0 = f$, make the paraxial approximation (expanding to lowest order in y) and show that the field at $z = 0$ can be written

$$E^{(+)}(R) = \frac{E_0^{(+)}}{f} \exp(-ikf) \exp\left(-ik\frac{y^2}{2f}\right). \quad (6.82)$$

(c) What is the phase shift corresponding to $\ell_{\text{lens}}(y)$? Write down a plane wave that matches the spherical wave provided the lens is placed at $z = 0$.

Problem 6.5

A Gaussian beam passes through a thin lens. Show that

$$R_2 = \frac{R_1 f}{f - R_1}, \quad (6.83)$$

where f is the focal length of the lens, R_1 is the radius of curvature of the input beam at the lens, and R_2 is the radius of curvature of the output beam at the lens. (Use the *ABCD* Law.)

Problem 6.6

A Gaussian beam with $w_0 = 1$ cm is focused by a thin lens of focal length $f = 2$ cm. The lens is placed at the focus of the original Gaussian beam. Assume an optical wavelength of $\lambda = 1.0 \mu\text{m}$.

- (a) At what distance from the lens does the new focus occur?
- (b) What is the spot size at the new focus? Give numbers for both an ideal and a realistic lens.
- (c) What is the far-field expansion angle?

Problem 6.7

Verify that the compact form for the Gaussian beam in terms of the complex beam parameter, Eq. (6.41),

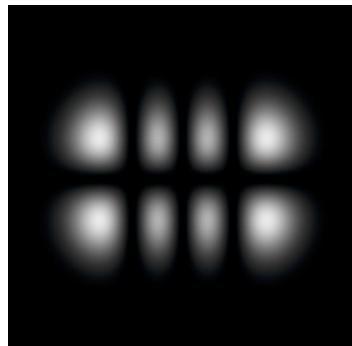
$$E^{(+)}(\mathbf{r}) = E_0^{(+)} \frac{q_0}{q(z)} \exp\left[\frac{ikr^2}{2q(z)}\right] \exp(ikz). \quad (6.84)$$

is equivalent to the standard form for the Gaussian beam, Eq. (6.10),

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} \frac{w_0}{w(z)} \exp\left[-\frac{r^2}{w^2(z)}\right] \exp\left[ikz - i\tan^{-1}\left(\frac{z}{z_0}\right)\right] \exp\left[ik\frac{r^2}{2R(z)}\right]. \quad (6.85)$$

Problem 6.8

- (a) What is the Hermite-Gaussian beam pictured below?



(b) Suppose the horizontal distance between the vertical dark bands is 100 μm at the focus. What is w_0 ?

Problem 6.9

(a) For the TEM_{0,0}, TEM_{0,1}, and TEM_{1,1} modes, show that the fraction of the total power contained in a disc of radius ρ is

$$\frac{P_{0,0}(\rho)}{P_{0,0}(\rho \rightarrow \infty)} = 1 - \exp \left[-2 \left(\frac{\rho}{w_0} \right)^2 \right] \quad (6.86)$$

for the TEM_{0,0} mode,

$$\frac{P_{0,1}(\rho)}{P_{0,1}(\rho \rightarrow \infty)} = 1 - \exp \left[-2 \left(\frac{\rho}{w_0} \right)^2 \right] \left[1 + 2 \left(\frac{\rho}{w_0} \right)^2 \right] \quad (6.87)$$

for the TEM_{0,1} mode, and

$$\frac{P_{1,1}(\rho)}{P_{1,1}(\rho \rightarrow \infty)} = 1 - \exp \left[-2 \left(\frac{\rho}{w_0} \right)^2 \right] \left[1 + 2 \left(\frac{\rho}{w_0} \right)^2 + 2 \left(\frac{\rho}{w_0} \right)^4 \right] \quad (6.88)$$

for the TEM_{1,1} mode. Note: you will need to evaluate an integral of the form

$$\int_0^a x^{2n+1} e^{-bx^2} dx. \quad (6.89)$$

This is easy for $n = 0$, since the integral is just a Gaussian:

$$\int_0^a x e^{-bx^2} dx = \frac{1}{2b} \exp(-bx^2) \Big|_0^a = \frac{1}{2b} [1 - \exp(-ba^2)]. \quad (6.90)$$

Other cases may be evaluated recursively, noting that

$$\int_0^a x^{2n+1} e^{-bx^2} dx = -\frac{\partial}{\partial b} \int_0^a x^{2n-1} e^{-bx^2} dx. \quad (6.91)$$

For example,

$$\int_0^a x^3 e^{-bx^2} dx = -\frac{\partial}{\partial b} \int_0^a x e^{-bx^2} dx = -\frac{\partial}{\partial b} \frac{1}{2b} [1 - \exp(-ba^2)] = \frac{1}{2b^2} [1 - \exp(-ba^2)(1 + ba^2)]. \quad (6.92)$$

(b) For each mode, what radius contains 99% of the total beam power?

Problem 6.10

A laser of wavelength λ emits a Gaussian beam, where the focus of the Gaussian beam occurs at the front of the laser. The laser beam hits a screen, which is a distance d from the front of the laser.

(a) What beam-waist parameter w_0 produces the smallest spot on the screen, and what is the corresponding $1/e^2$ beam radius on the screen?

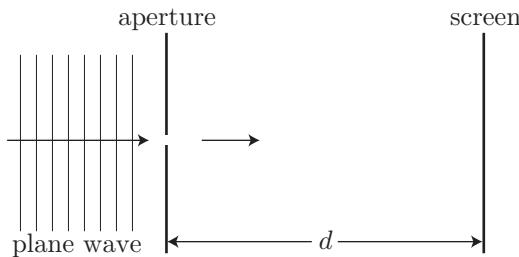
(b) What is the $1/e^2$ beam radius on the screen for $\lambda = 200\pi$ nm and $d = 10$ cm?

Problem 6.11

Here we will consider a simple model for the pinhole camera. The idea is to compute the optimal size of the pinhole to get the best image resolution: the image on the screen of a distant point source must be at least the size of the aperture, but also is blurred by diffraction if the aperture is too small. The optimal diameter can be chosen as a compromise between the two effects; Lord Rayleigh⁴ found an optimal pinhole diameter of $1.9\sqrt{\lambda d}$ for a *uniform* circular aperture.

Consider an aperture illuminated by a plane wave. The light passing through the aperture is viewed on a screen a distance d away.

⁴“Pinhole camera,” Wikipedia entry (http://en.wikipedia.org/wiki/Pinhole_camera) as of 25 February 2006.



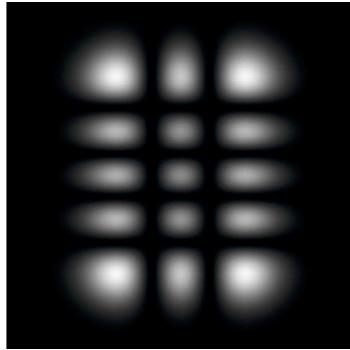
- (a) If the incident light is treated in *geometrical optics*, what is the size of the aperture that gives the smallest spot on the screen? *Explain.* (Hint: what do plane waves correspond to in geometrical optics?)
- (b) Now answer the same question in wave optics. Assume that the aperture has a *Gaussian* transmission coefficient for the electric field of

$$t = \exp\left(-\frac{r^2}{a^2}\right), \quad (6.93)$$

(i.e., $E_{\text{out}} = tE_{\text{in}}$), so that a is the “radius” of the aperture. Now what is the size of the aperture that gives the smallest spot on the screen?

Problem 6.12

- (a) What is the Hermite–Gaussian beam pictured below?



- (b) Suppose this beam has the same fractional power contained in a circular disc of radius a as a Gaussian beam. Compare (qualitatively) the far-field divergence of the two beams.

Problem 6.13

Consider two monochromatic, Gaussian beams: a “red” beam of frequency ω , and a “blue” beam of frequency 2ω . Both have identical transverse intensity profiles at their respective foci. Describe completely how the geometries (both near- and far-field) of the two fields differ.

Problem 6.14

A monochromatic Gaussian beam (with beam waist parameter w_0) is focused onto a thin, *nonlinear* optical element, whose effect on the incident field is

$$E_{\text{out}} = \alpha(E_{\text{in}})^2, \quad (6.94)$$

where α is some constant. Show that if the optical frequency of the input beam is ω , the output field has a component with optical frequency 2ω . Describe completely how the geometry (both near- and far-field) of the 2ω output field differs from the input field.

Hint: The nonlinear optical element only acts at $z = 0$, so consider what the field looks like there.

Problem 6.15

The set of all Hermite–Gaussian modes with the same beam parameter w_0 form a *complete* set. Denoting the $\text{TEM}_{l,m}$ mode by

$$E_{l,m}^{(+)}(\mathbf{r}; w_0) \quad (6.95)$$

to emphasize the dependence on w_0 , this means that an arbitrary electric field profile $E^{(+)}(\mathbf{r})$ at $z = 0$ can be written as a superposition of all the modes at $z = 0$:

$$E^{(+)}(x, y) = \sum_{l,m=0}^{\infty} c_{l,m} E_{l,m}^{(+)}(x, y; w_0). \quad (6.96)$$

(a) Noting that the Hermite–Gaussian modes are all orthogonal, explain how you would compute the coefficients $c_{l,m}$. (Set up any necessary equations but don't refer to the explicit form of $E_{l,m}^{(+)}(\mathbf{r}; w_0)$ or evaluate any complicated integrals.)

(b) Once you know these coefficients, you can compute the field at *any* location along the z -axis by

$$E^{(+)}(x, y, z) = \sum_{l,m=0}^{\infty} c_{l,m} E_{l,m}^{(+)}(x, y, z; w_0). \quad (6.97)$$

Find the flaw in the following argument: each mode $E_{l,m}^{(+)}(\mathbf{r}; w_0)$ has the same far-field divergence angle ($\theta_0 = \tan^{-1}(w_0/z_0)$); thus, *every* field, no matter what its profile (or size) at $z = 0$, has a far-field divergence angle of θ_0 . (For example, you can apply this argument to a Gaussian of size $15w_0$, but this clearly has a different far-field divergence angle than θ_0 for the constituent beams.)

Hint: how is the far-field divergence angle for a Gaussian beam defined? How would modifying the definition give a different answer?

Problem 6.16

Suppose that an optical cavity is described by the round-trip ray matrix

$$\mathbf{M}(z_i) = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad (6.98)$$

which is computed for a ray starting at location z_i . Assume that a Gaussian beam is the characteristic mode of the cavity, and let $q = q(z_i)$ denote the complex beam parameter for the beam at $z = z_i$.

(a) Show that

$$q = \left(\frac{A - D}{2C} \right) - i \sqrt{\frac{1}{C^2} - \left(\frac{A + D}{2C} \right)^2}. \quad (6.99)$$

(b) Show that the radius of curvature at $z = z_i$ is

$$R(z_i) = -\frac{2B}{A - D} \quad (6.100)$$

and the beam radius at $z = z_i$ is

$$w(z_i) = \sqrt{\frac{\lambda|B|}{\pi \sqrt{1 - \left(\frac{A + D}{2} \right)^2}}}. \quad (6.101)$$

Hint 1: what should happen to the q parameter after one round trip if it is to be the resonant mode of a cavity?

Hint 2: recall that $\det(\mathbf{M}) = 1$ for a round-trip matrix.

Problem 6.17

Consider a Gaussian beam propagating along the $+z$ direction in vacuum, with beam waist w_0 and vacuum wavelength λ_0 . Now compare this to the case where the beam shines into a thick dielectric slab of refractive index $n > 1$, extending to $z = +\infty$, and with the (planar) vacuum–dielectric interface occurring a distance d in front of the original focus (i.e., at $z = -d$). The beam is perpendicular to the interface, and the new focus occurs inside the dielectric slab.

- (a) Use the *ABCD* Law to derive an expression for the location of the new focus relative to the interface.
- (b) What is the beam waist in the dielectric case, in terms of the beam waist w_0 in the vacuum case?

Problem 6.18

The Hermite polynomials may be defined as in Eq. (6.76) by the explicit formula (**Rodrigues formula**)

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}, \quad (6.102)$$

but they can also be defined via the **generating function**

$$g(x, t) := e^{2xt - t^2} = \sum_{n=0}^{\infty} H_n(x) \frac{t^n}{n!}. \quad (6.103)$$

That is, the Hermite polynomials are the coefficients in the Taylor expansion of g in t .

- (a) Show that differentiating the generating function with respect to t leads to the recurrence

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x). \quad (6.104)$$

- (b) Show that differentiating the generating function with respect to x leads to the recurrence

$$H'_n(x) = 2nH_{n-1}(x). \quad (6.105)$$

- (c) Show that the Rodrigues formula leads to the same recurrence as in (a).
- (d) Show that the Rodrigues formula leads to the same recurrence as in (b).
- (e) Thus argue that the two definitions for the Hermite polynomials are equivalent.

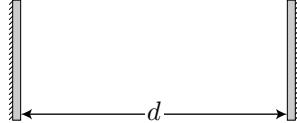
Chapter 7

Fabry–Perot Cavities

Now let's return to the subject of resonators, which we treated earlier in Section 2.7 in the context of ray optics. Now we will perform a better analysis, which includes the wave nature of the light. Interference is a central feature of resonators with waves, and thus resonators are also interferometers. Because of this, the conventional two-mirror resonator is referred to as a **Fabry–Perot interferometer**, **Fabry–Perot cavity**, **Fabry–Perot etalon**, or an **optical spectrum analyzer**.

7.1 Resonance Condition

We will start with the simplest case of a planar cavity of length d , where the mirrors are perfect reflectors.



In order for a plane wave to exist in the cavity, it must return to exactly the same phase after one round trip through the cavity. In this case, it will constructively interfere with itself. Otherwise, the phase will precess on each successive round trip, and eventually lead to destructive interference. Thus, the round-trip accumulated phase must be some integer multiple of 2π :

$$2kd = 2\pi q \quad (q = 0, 1, 2, \dots). \quad (7.1) \quad (\text{cavity resonance condition})$$

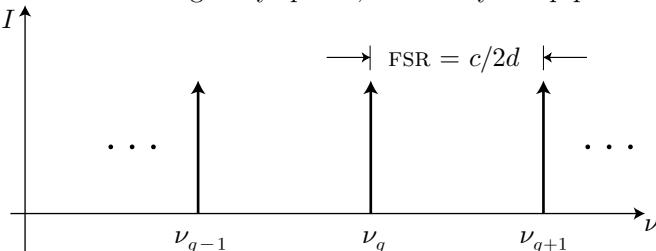
Thus, the allowed stationary waves, or **modes**, of the cavity are plane waves that satisfy

$$k_q = \frac{\pi q}{d}. \quad (7.2) \quad (\text{cavity allowed wave numbers})$$

We can also rewrite this resonance condition in terms of related quantities:

$$\lambda_q = \frac{2d}{q}; \quad \nu_q = \frac{cq}{2d}; \quad \omega_q = \frac{\pi cq}{d}. \quad (7.3) \quad (\text{cavity allowed wavelengths/frequencies})$$

The corresponding spectrum is a set of regularly spaced, arbitrarily sharp peaks.



The frequency spacing is a special quantity called the **free spectral range**, defined as

$$\text{FSR} := \nu_{q+1} - \nu_q = \frac{c}{2d} = \frac{c_0}{2nd}, \quad (7.4)$$

(free spectral range)

where n is the refractive index of the medium filling the cavity.

An alternative to this constructive-interference approach to the resonance condition is to view the reflectors as imposing boundary conditions on the wave. The wave inside the cavity is a standing-wave mode, which is a superposition of left- and right-going waves:

$$E^{(+)}(\mathbf{r}) = E_{01}^{(+)} e^{ikz} + E_{02}^{(+)} e^{-ikz}. \quad (7.5)$$

For ideal mirrors, the amplitudes of the two waves are equal, so that

$$|E_{01}^{(+)}| = |E_{02}^{(+)}| =: E_0^{(+)}. \quad (7.6)$$

Thus, we can write

$$\begin{aligned} E^{(+)}(\mathbf{r}) &= E_0^{(+)} (e^{ikz+i\phi_1} + e^{-ikz-i\phi_2}) \\ &= E_0^{(+)} e^{i(\phi_1-\phi_2)/2} (e^{ikz+i(\phi_1+\phi_2)/2} + e^{-ikz-i(\phi_1+\phi_2)/2}) \\ &= E_0^{(+)} e^{i\Delta\phi/2} \cos(kz + \bar{\phi}), \end{aligned} \quad (7.7)$$

where ϕ_1 and $-\phi_2$ are the phases of the first and second fields, respectively, $\Delta\phi := \phi_1 - \phi_2$, and $\bar{\phi} := (\phi_1 + \phi_2)/2$. The wave must vanish at the mirrors, which we can view as perfect conductors. If the left-hand mirror is at $z = 0$,

$$E^{(+)}(z = 0) = 0 \implies \bar{\phi} = -\pi/2, \quad (7.8)$$

so that the cosine changes to a sine:

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} e^{i\Delta\phi/2} \sin(kz). \quad (7.9)$$

Then the right-hand mirror is at $z = d$:

$$E^{(+)}(z = d) = 0 \implies \sin(kd) = 0 \implies kd = q\pi. \quad (7.10)$$

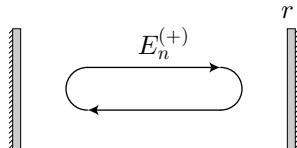
This is the same resonance condition as before.

The term *resonance* here suggests that when the cavity is exposed to light with one of the allowed frequencies, the cavity will respond, but not otherwise. We will discuss this more below, after generalizing what we have done so far to “leaky” cavities.

7.2 Broadening of the Resonances: Cavity Damping

Real cavities involve some loss of the light on each round trip. The losses damp the amplitude of the cavity. Just as in the case of a driven mechanical oscillator, the damping leads to broadening of the resonances.

Consider the circulating field inside a cavity. In particular, we will track the amplitude of the field on its n th round trip in the cavity.



Suppose that the amplitude of the wave is reduced by a factor of r on each round trip:

$$|E_{n+1}^{(+)}| = |rE_n^{(+)}|. \quad (7.11)$$

Here, $E_n^{(+)}$ is the electric field amplitude of the plane wave on the n th round trip through the cavity. The loss due to r could be some transmission of the mirrors (as indicated in the diagram here), losses around the edges of finite-size mirrors, scattering from gases or objects inside the cavity, and so on. On one round trip, the phase of the plane wave changes by

$$\Delta\phi = 2kd = \frac{4\pi d}{\lambda}, \quad (7.12)$$

so that

$$E_{n+1}^{(+)} = re^{i2kd} E_n^{(+)}. \quad (7.13)$$

We may assume that r is real, as any nonzero phase angle can be absorbed into the phase factor as an additional effective length. The total wave is

$$E^{(+)} = E_0^{(+)} + E_1^{(+)} + E_2^{(+)} + \dots = E_0^{(+)} [1 + re^{i2kd} + (re^{i2kd})^2 + \dots] = \frac{E_0^{(+)}}{1 - re^{i2kd}}, \quad (7.14)$$

where we have used $\sum_{n=0}^{\infty} x^n = 1/(1-x)$. The total intensity is

$$\begin{aligned} I &= \frac{I_0}{|1 - re^{i2kd}|^2} \\ &= \frac{I_0}{|1 - r \cos(2kd) - i \sin(2kd)|^2} \\ &= \frac{I_0}{[1 - r \cos(2kd)]^2 + r^2 \sin^2(2kd)} \\ &= \frac{I_0}{1 - 2r \cos(2kd) + r^2 \cos^2(2kd) + r^2 \sin^2(2kd)} \\ &= \frac{I_0}{1 + r^2 - 2r \cos(2kd)} \\ &= \frac{I_0}{(1 - r)^2 + 4r \sin^2(kd)}, \end{aligned} \quad (7.15)$$

where $I_0 = 2|E_0^{(+)}|^2/\eta$.

7.2.1 Standard Form

In standard form, we can write the resonator intensity as

$$I = \frac{I_{\max}}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2(kd)}, \quad (7.16)$$

(cavity circulating intensity)

where the maximum intensity is

$$I_{\max} := \frac{I_0}{(1 - r)^2}, \quad (7.17)$$

(maximum circulating intensity)

and the **finesse**, defined by

$$\mathcal{F} := \frac{\pi\sqrt{r}}{1 - r}, \quad (7.18)$$

(cavity finesse)

is an important parameter that characterizes the lossy nature of the cavity. The only dependence of the intensity on the cavity length d is in the \sin function, so the intensity is periodic in d with period $\lambda/2$. In terms of the free spectral range $c/2d$, we can write the intensity as

$$I = \frac{I_{\max}}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2\left(\frac{\pi\nu}{\text{FSR}}\right)}, \quad (7.19) \quad (\text{cavity circulating intensity})$$

so that in frequency the intensity is also periodic, where the period is the free spectral range.

7.2.2 Maximum and Minimum Intensity

The intensity is maximized when the denominator is smallest. This occurs when the sinusoidal term vanishes, which gives the condition

$$\sin^2\left(\frac{\pi\nu}{\text{FSR}}\right) = 0 \implies \nu = q(\text{FSR}). \quad (7.20)$$

This is just the resonance condition for the lossless cavity. At these points, the cavity attains the maximum intensity I_{\max} . The minimum intensity occurs when the denominator is maximum. Thus the sinusoidal term becomes unity, and

$$I_{\min} = \frac{I_{\max}}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2}. \quad (7.21)$$

The minimum intensity only goes to zero in the limit of large finesse.

7.2.3 Width of the Resonances

As a way to characterize the width of the resonances, we will now find the frequencies such that the intensity falls to $I_{\max}/2$. In this case, the width condition is

$$\sin^2\left(\frac{\pi\nu}{\text{FSR}}\right) = \left(\frac{\pi}{2\mathcal{F}}\right)^2, \quad (7.22)$$

which is satisfied by the frequencies

$$\nu = \pm \frac{\text{FSR}}{\pi} \sin^{-1}\left(\frac{\pi}{2\mathcal{F}}\right), \quad (7.23)$$

up to an added multiple of the free spectral range (again, because of the periodicity of the \sin function). The width of the resonances is really only sensible concept in the limit of large finesse, when the resonances are well resolved. In this limit, we can use the small-angle approximation for the \sin function and write the half-maximum intensity frequencies as

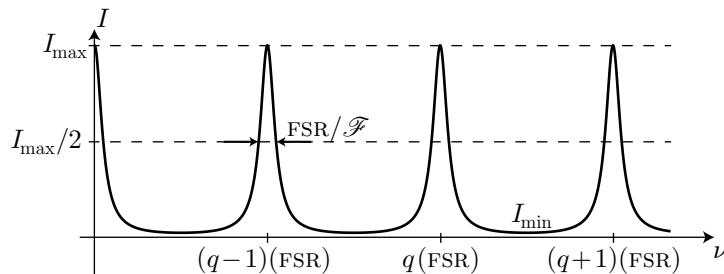
$$\nu \approx \pm \frac{\text{FSR}}{2\mathcal{F}}. \quad (7.24)$$

In this approximation, we can thus write the full width of the resonance at half maximum as

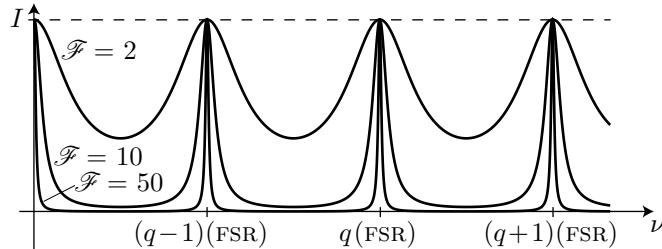
$$\delta\nu_{\text{FWHM}} = \frac{\text{FSR}}{\mathcal{F}}. \quad (\text{cavity resonance full width at half maximum}) \quad (7.25)$$

Notice that in the limit of arbitrarily large finesse, $\delta\nu_{\text{FWHM}} \rightarrow 0$ while I_{\max} diverges—the resonances become δ functions, perfectly defined resonances as in the lossless cavity.

We can summarize the characteristics of the Fabry-Perot cavity in the diagram, recalling that two parameters (FSR and \mathcal{F}) are necessary to completely specify the cavity behavior; all other parameters are determined by these two.



The next diagram compares the cavity intensities for several different finesse, showing the transition from broad resonances to sharply defined peaks.



7.2.4 Survival Probability

Losses in a cavity can come from a number of sources. The mirrors can lead to intensity loss due to partial transmission, losses of beams around the mirror edges, absorption of the light in the reflective coatings, and scatter due to surface roughness. A medium inside the cavity can also absorb or scatter the light. To generalize the treatment of losses, we can define the **survival probability** P_s , which is the fraction of the intensity that stays in the cavity after one round trip. If the two mirrors have **reflectances** (intensity reflection coefficients) $R_{1,2}$, for example, then the survival probability is just

$$P_s = R_1 R_2. \quad (7.26)$$

Additional factors can model other losses due, for example, to other objects in the cavity. In terms of the survival probability, we can write

$$\mathcal{F} = \frac{\pi P_s^{1/4}}{1 - \sqrt{P_s}} \quad (7.27) \quad (\text{finesse in terms of survival probability})$$

as a more convenient expression for the finesse and

$$I_{\max} := \frac{I_0}{(1 - \sqrt{P_s})^2},$$

(maximum cavity intensity in terms of survival probability) (7.28)

for the maximum intensity.

7.2.5 Photon Lifetime

The name “survival probability” more appropriately describes the fate of a photon after one round trip through the cavity. The probability that the photon is lost is simply $1 - P_s$. Thus, a photon lasts for

$$\langle n \rangle = \frac{1}{1 - P_s} \quad (7.29)$$

round trips on average. We can see this as follows. The probability for a photon to exit the cavity on the n th round trip is the product of the probability P_s^{n-1} to survive for $n - 1$ round trips, multiplied by the probability $(1 - P_s)$ for the photon to not survive the n th round trip:

$$P(n) = P_s^{n-1}(1 - P_s). \quad (7.30)$$

The average number of round trips survived is then the mean of this distribution function, noting that the photon can only fail to survive after one round trip:

$$\begin{aligned}
 \langle n \rangle &= \sum_{n=1}^{\infty} nP(n) \\
 &= \sum_{n=0}^{\infty} nP(n) \\
 &= (1 - P_s) \sum_{n=0}^{\infty} nP_s^{n-1} \\
 &= (1 - P_s) \frac{\partial}{\partial P_s} \sum_{n=0}^{\infty} P_s^n \\
 &= (1 - P_s) \frac{\partial}{\partial P_s} \left(\frac{1}{1 - P_s} \right) \\
 &= (1 - P_s) \frac{1}{(1 - P_s)^2} \\
 &= \frac{1}{1 - P_s}.
 \end{aligned} \tag{7.31}$$

Here, we have again used the geometric sum $\sum_{n=0}^{\infty} x^n = 1/(1-x)$, and note the handy trick of using the derivative to transform the sum into a more standard one.

Since the time for one round trip is $\tau_{rt} = 2d/c$, we can more generally write

$$\tau_{rt} = \frac{1}{\text{FSR}}. \tag{7.32}$$

(round-trip time)

Then the “lifetime” of a photon inside the cavity is the product of the round-trip time and the average number of round trips that the photon will survive:

$$\tau_p = \frac{1}{(\text{FSR})(1 - P_s)}. \tag{7.33}$$

(cavity photon lifetime)

For a “good” resonator, P_s is close to unity, so $\sqrt{P_s}$ is also close to unity. In this limit we can expand to lowest order in $(1 - \sqrt{P_s})$, which amounts to setting $P_s^{1/4} \approx 1$ and $(1 - P_s) = (1 - \sqrt{P_s})(1 + \sqrt{P_s}) \approx 2(1 - \sqrt{P_s})$. Thus,

$$\mathcal{F} \approx \frac{\pi}{1 - \sqrt{P_s}} \tag{7.34}$$

and so

$$\tau_p \approx \frac{1}{2(\text{FSR})(1 - \sqrt{P_s})} = \frac{\mathcal{F}}{2\pi(\text{FSR})} = \frac{1}{2\pi\delta\nu_{\text{FWHM}}}. \tag{7.35}$$

Thus we arrive at an “uncertainty relation”

$$\tau_p \delta\nu_{\text{FWHM}} = \frac{1}{2\pi} \tag{7.36}$$

(cavity uncertainty relation)

for optical cavities between frequency and time.

7.2.6 Q Factor

The **Q factor** represents the “quality” of any oscillator. The general definition is

$$Q := 2\pi \cdot \frac{\text{stored energy}}{\text{energy loss per oscillation cycle}}. \tag{7.37}$$

(Q factor for general oscillator)

For a “good” oscillator—that is, one that does not dissipate energy very rapidly—the Q factor is large because the denominator in the definition is small. A cavity is an oscillator, but at optical frequencies, which are of the order of a few hundred THz. Thus, we can already see that optical Q factors are very large. In fact, the finesse plays the same role as Q for optical cavities, but generally produces more manageable numbers.

For an optical cavity, the fraction of energy lost per round trip is $1 - P_s$. Following the same logic as in the last section, the rate at which energy leaves the cavity is $1/\tau_p$, relative to the amount of energy in the cavity. This may not be completely obvious, so let’s quickly show this. In terms of the survival probability, the probability to survive n round trips in the cavity is

$$P_s^n = (P_s)^{t/\tau_{\text{rt}}} = e^{(t/\tau_{\text{rt}}) \log P_s} \approx e^{-(t/\tau_{\text{rt}})(1-P_s)} = e^{-t/\tau_p}, \quad (7.38)$$

where we used $\log(1+x) \approx x$ for small $x = 1 - P_s$, and $\tau_p = \tau_{\text{rt}}/(1 - P_s)$, which follows from Eqs. (7.32) and (7.33). Thus, the energy loss is exponential in time (with the assumption of a good cavity), with time constant τ_p , and hence loss rate $1/\tau_p$.

Continuing with Q , the energy loss per oscillation cycle is just the optical period multiplied by this rate, or $1/\nu_q \tau_p$. Putting this into the above definition, we see that

$$Q = 2\pi\nu_q \tau_p. \quad (7.39) \quad (Q \text{ for optical cavity})$$

We can also write this in terms of the cavity line width, using the uncertainty relation in Eq. (7.36):

$$Q = \frac{\nu_q}{\delta\nu_{\text{FWHM}}}. \quad (7.40) \quad (Q \text{ for optical cavity})$$

We can compare this to the form of the finesse from Eq. (7.25),

$$\mathcal{F} = \frac{\text{FSR}}{\delta\nu_{\text{FWHM}}}, \quad (7.41)$$

and see that Q and \mathcal{F} are the same except that a different (smaller) oscillator frequency is used for the finesse. Since $\nu_q = q(\text{FSR})$, we see that Q and \mathcal{F} differ by a factor of the resonance order q :

$$Q = q\mathcal{F}. \quad (7.42) \quad (Q \text{ related to finesse})$$

The order q is typically a very large integer for optical cavities, since the cavity length often far exceeds the optical wavelength. Thus, the Q factor for optical cavities is typically *huge* compared to the finesse. That’s why the finesse is a preferred quantity in the optical regime—the numbers are much easier to deal with.¹

7.2.7 Example: Finesse and Q

Consider a planar cavity where both mirrors reflect 99% of the incident intensity. Then $P_s = 0.99^2$, and so $\mathcal{F} \approx 300$.

For a cavity length $d = 15$ cm, the round-trip time is $\tau_{\text{rt}} = 2d/c = 1$ ns. (A related handy number to remember for the speed of light is $c_0 \approx 1$ ft/ns.) Then the free spectral range is $\text{FSR} = 1/\tau_{\text{rt}} \approx 1$ GHz.

The photon lifetime is $\tau_p = \mathcal{F}/2\pi(\text{FSR}) \approx 50$ ns. This leads to a cavity line width of $\delta\nu_{\text{FWHM}} = 1/2\pi\tau_p \approx 3$ MHz.

A typical optical frequency is 300 THz, so that $Q = \nu_q/\delta\nu_{\text{FWHM}} = 10^8$, a really large number compared to the finesse.

¹ Nevertheless, the Q factor is a useful measure of the stability of the best oscillators, which now operate at optical frequencies. In 2010 the record for an observed $Q = 4.2 \times 10^{14}$ was set using a clock based on an Al⁺ ion operating at 1.121 PHz, with a FWHM width of 2.7 Hz: C. W. Chou, D. B. Hume, T. Rosenband, and D. J. Wineland, “Optical Clocks and Relativity,” *Science* **329**, 1630 (2010) (doi: 10.1126/science.1192720).

The corresponding wavelength is $\lambda = c_0/\nu_q = 1 \mu\text{m}$, which is in the near-infrared (approximately the wavelength of a Nd:YAG laser). The order is thus $q = \nu_q/(\text{FSR}) = 3 \times 10^5$. Recall that $q = 2d/\lambda$, and it is because typical cavities are long compared to the optical wavelength that the order q is often very large.

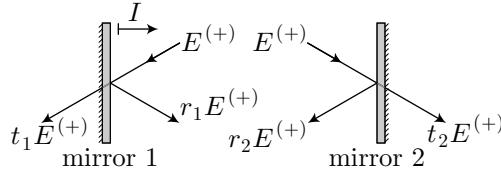
Again, one reason that Q is huge is because the optical frequency is so high. In the microwave, cavity Q factors are more reasonable. For example, in the microwave $\nu_q \sim 1 \text{ GHz}$, and so $\delta\nu_{\text{FWHM}} \sim 1 \text{ MHz}$ would yield a Q of 10^3 . But note that some microwave devices can have extremely narrow line widths, so microwave Q factors can be large as well.

The best “supercavities” in the optical have finesse in the range of 10^5 to 10^6 at present.²

7.3 Cavity Transmission

If you think about it in a certain way, a Fabry–Perot cavity is an amazing device. Imagine a mirror that is nearly perfectly reflecting. A laser beam impinging on the front mirror will mostly bounce off, only transmitting a small fraction of the intensity. But by putting a second mirror *behind* that mirror, just the right distance away, *all the laser power can instead transmit through the mirror*. Makes your brain tingle just thinking about it, doesn’t it? Such is the power of interference. Let’s see how this works out.

We now need to be a bit more precise about the intensities and electric fields in and around the cavity. We will refer to the left- and right-hand mirrors as mirrors 1 and 2, respectively. The total circulating intensity I within the cavity is the intensity propagating to the right of and away from mirror 1. We will also use the conventions in the diagram for the *field* reflection and transmission coefficients (with primed coefficients for the reverses of the directions indicated).



The circulating intensity is related to the intensity of the initial wave before the resonant buildup by Eq. (7.19):

$$I = \frac{I_{\max}}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2\left(\frac{\pi\nu}{\text{FSR}}\right)}, \quad (7.43)$$

where again $I_{\max} = I_0/(1 - \sqrt{P_s})^2$. In terms of the field reflection coefficients from the inside of the cavity, $P_s = r_1^2 r_2^2$.

The output (transmitted) intensity is related to the intracavity intensity by

$$I_{\text{out}} = |t_2|^2 I, \quad (7.44)$$

while the initial intensity I_0 is similarly related to the input intensity:

$$I_0 = |t'_1|^2 I_{\text{in}}. \quad (7.45)$$

Thus we can define the **cavity intensity transmission coefficient** T_{cav} by

$$T_{\text{cav}} := \frac{I_{\text{out}}}{I_{\text{in}}} = |t'_1 t_2|^2 \frac{I}{I_0} = \frac{T_{\text{cav,max}}}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2\left(\frac{\pi\nu}{\text{FSR}}\right)} \quad (\text{cavity intensity transmission coefficient}) \quad (7.46)$$

²A finesse of 1.9×10^6 was reported by G. Rempe, R. J. Thompson, H. J. Kimble, and R. Lalezari, “Measurement of ultralow losses in an optical interferometer,” *Optics Letters* **17**, 363 (1992).

where

$$T_{\text{cav,max}} := \frac{|t'_1 t_2|^2}{(1 - r_1 r_2)^2} \quad (7.47)$$

(maximum cavity transmission)

is the maximum fraction of the input intensity that the cavity can transmit. So here's the statement that we started with: if the two mirrors are identical, then $r_1 = r_2 =: r$ and $t_1 = t_2 =: t$. Then the maximum cavity transmission becomes

$$T_{\text{cav,max}} = \frac{|t' t|^2}{(1 - r^2)^2}. \quad (7.48)$$

But one of the Stokes relations, Eq. (5.17), states that $t' t = 1 - r^2$, so for two identical mirrors,

$$T_{\text{cav,max}} = 1. \quad (7.49)$$

(maximum cavity transmission, identical mirrors)

Thus, when the two mirrors are identical, a *resonant Fabry-Perot cavity transmits all the input light, independent of the reflectivity of the mirrors*, at least in steady state after the circulating field builds up.

7.3.1 Reflected Intensity

So what happened to the input light that supposedly reflects off of the input mirror? It must work out that it disappears via destructive interference, and we can now show this explicitly. Part of the intracavity wave transmits through the front mirror as well. From the way we defined I , the part transmitted through the front mirror is also reduced by the reflection coefficient of the second mirror:

$$I_{\text{out(front)}} = |t_1|^2 |r_2|^2 I = \frac{|r_2|^2 |1 - r_1^2|^2}{|t'_1|^2} I. \quad (7.50)$$

Again, we used the Stokes relation $t' t = 1 - r^2$ to arrive at the last expression. In the case of identical mirrors, then,

$$I_{\text{out(front)}} = \frac{|r|^2 |1 - r^2|^2}{|t'|^2} I, \quad (7.51)$$

and on resonance, $I = I_{\text{max}} = I_0/(1 - r^2)^2$, so

$$I_{\text{out(front)}} = \frac{r^2}{|t'|^2} I_0 = r^2 I_{\text{in}}. \quad (7.52)$$

This is exactly the same intensity as the part of the input beam that reflects off the first mirror. But in terms of the *field*, there is a π phase shift of the reflected beam compared to the transmitted beam, due to the other Stokes relation $r' = -r$. Thus, these two beams *destructively interfere* in steady state. That's how all the power can transmit through the resonant cavity and allow for energy conservation in steady state.

7.3.2 Intracavity Buildup

Notice that these effects require that on resonance, the cavity circulating intensity must be much larger than the input or output intensity in the regime of large finesse. This is most easily seen from Eq. (7.35), which we can rewrite as

$$\frac{\tau_p}{\tau_{\text{rt}}} \approx \frac{\mathcal{F}}{2\pi}. \quad (7.53)$$

Thus, we can conveniently identify $\mathcal{F}/2\pi$ (actually, the correct factor will be \mathcal{F}/π , as we discuss below) as a “buildup factor,” since light entering a resonant cavity makes this many round trips inside the cavity before leaving it. We can also see this directly from the intensities, since the output intensity for a *symmetric* cavity is related to the circulating intensity by $I_{\text{out}} = |t|^2 I = |t|^2 I_{\text{max}}$. From energy conservation, the reflected and

transmitted intensities of a mirror must add to the original intensity, so $|t|^2 = 1 - R$, where R is the intensity reflection coefficient for the mirror, and $1 - R = 1 - \sqrt{P_s}$ for a symmetric cavity. Hence,

$$I_{\max} = \frac{I_{\text{out}}}{1 - \sqrt{P_s}}, \quad (7.54)$$

so that

$$I_{\max} \approx \frac{\mathcal{F}}{\pi} I_{\text{out}} = \frac{\mathcal{F}}{\pi} I_{\text{in}} \quad (\text{symmetric cavity buildup}) \quad (7.55)$$

in the limit of large finesse. Note that this differs from the photon-lifetime estimate by a factor of 2, because the loss for a resonant, symmetric cavity *only* occurs due to the *output* mirror (recall that the output through the *input* mirror destructively interferes with the reflected input light). Also we used that $I_{\text{in}} = I_{\text{out}}$ for a symmetric cavity, so the cavity circulating intensity is a factor of \mathcal{F}/π larger than the input in the good-cavity limit.

For an *asymmetric* cavity, we can combine Eqs. (7.28) and (7.45) to obtain

$$\frac{I_{\max}}{I_{\text{in}}} = \frac{|t'_1|^2}{(1 - \sqrt{P_s})^2}, \quad (7.56)$$

and again, in the limit where Eq. (7.34) applies,

$$\frac{I_{\max}}{I_{\text{in}}} \approx \frac{|t'_1|^2 \mathcal{F}^2}{\pi^2}, \quad (\text{asymmetric cavity buildup}) \quad (7.57)$$

so the buildup factor is slightly more complicated in this case. For a symmetric cavity, again

$$|t'_1|^2 = 1 - R = 1 - \sqrt{P_s} \approx \frac{\pi}{\mathcal{F}}, \quad (7.58)$$

so that we recover the symmetric-cavity expression (7.55).

Note that the way we defined the circulating intensity I , this is only the intensity of the *rightward-propagating* field in the cavity. But the field inside a linear cavity, as pictured at the beginning of this chapter, is a superposition of left- and right-going waves and thus a standing wave. That is, a field probe placed in the cavity would see *more* than I . The intensity I does, however, properly represent the circulating field in a *ring* cavity, where the intensity only circulates in one direction.

For the linear cavity, the left-going wave is reduced in amplitude because of reflection from the right-hand (output) mirror. In the good-cavity limit, this reduction is small, and the total field is an interference pattern of left- and right-going waves, each of intensity I —a standing-wave pattern. The total intracavity intensity I_{cav} for a linear cavity is thus

$$I_{\text{cav}}(z) = \left| \sqrt{I} e^{ikz} - \sqrt{I} e^{-ikz} \right|^2 = 2I[1 - \cos(2kz)] = 4I \sin^2(kz). \quad (\text{circulating intensity inside linear, planar cavity}) \quad (7.59)$$

Thus, the cavity intensity is a sinusoidal pattern with maximum intensity $4I \approx 2\mathcal{F}/\pi$ and period $\lambda/2$. Recall that for consistency with the boundaries that the field must vanish at $z = 0$ and d , so we chose the proper phase of the left-going field.

7.4 Optical Spectrum Analyzer

In the good-cavity regime, a Fabry–Perot cavity transmits light only at well-defined frequencies. Thus the Fabry–Perot cavity is extremely useful in the laboratory as an **optical spectrum analyzer**—a device that can measure the spectrum of an input light source. The width $\delta\nu_{\text{FWHM}}$ is the **resolution** of the analyzer, since it cannot distinguish frequencies within this narrow range. We now also see the reason for the terminology *free spectral range* for the spacing between cavity resonances, since this is the longest range in frequency that the cavity can measure without repetition or “wrapping” of the spectrum.

Very often, a Fabry–Perot spectrum analyzer can be scanned or tuned by changing the cavity length d . This can be accomplished by mounting one of the mirrors on a piezoelectric stack. If the length cavity is changed by an amount δd , the corresponding change in the resonance frequencies is given by

$$\delta\nu_q = \frac{qc}{2(d + \delta d)} - \frac{qc}{2d} \approx -\frac{qc\delta d}{2d^2} = -\nu_q \frac{\delta d}{d}, \quad (7.60)$$

where we have expanded to lowest order in δd . Note that by a similar argument, the free spectral range changes by the much smaller amount $[-(\text{FSR})\delta d/d]$, so the change in free spectral range is a second-order effect (a factor of $\sim 10^5$ smaller for typical optical cavities).

In general, it is very useful to note that small changes in any cavity parameter translates into linear changes in other parameters,

$$\frac{\delta\lambda_q}{\lambda} = \frac{\delta d}{d} \approx -\frac{\delta\nu_q}{\nu_q} = -\frac{\delta k_q}{k_q}, \quad (7.61) \quad (\text{shifts in cavity parameters})$$

at least to lowest order in the perturbation. These relations follow from the fact that these quantities are all proportional or inversely proportional. Most useful spectrum-analyzer-type problems can be solved with similar relations.

This is what a Fabry–Perot optical spectrum analyzer looks like in the laboratory. This is actually a confocal cavity, which we'll get to later, so the parameters work out a little differently than for the planar cavity. The cavity length is $d = 5$ cm giving $\text{FSR} = 1.5$ GHz. The mirrors are coated for 99.2% reflectivity, so that $\mathcal{F} = 200$.



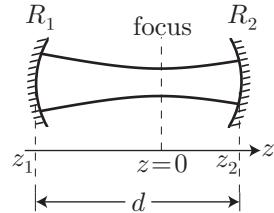
Here is the same cavity disassembled. On the left you can see one of the mirrors mounted on a threaded barrel. The barrel is used to adjust the cavity length to match precisely the mirror curvatures. The other mirror is attached to the right-hand side of the main body on a tubular piezo stack, which gives a total displacement of $\sim 1 \mu\text{m}$ for a 0–15 V control signal. The mirror cover on the right contains a photodiode to monitor the transmitted light.



7.5 Spherical-Mirror Cavities: Gaussian Modes

In the more general case that we analyzed in ray optics, resonators are composed of spherical mirrors. Plane waves are no longer compatible with the mirrors. Instead, we must turn to Gaussian beams. As in the planar case, the waves must vanish at the mirror boundary. Because Gaussian beams have spherical wave fronts, they are a natural choice for cavity modes that match the spherical boundary conditions.

Consider the spherical-mirror cavity shown here. We will try to “fit” a Gaussian beam to it.



As usual the Gaussian beam focus occurs at $z = 0$, and the mirrors of curvature radii R_1 and R_2 are placed at z_1 and z_2 , respectively, and they are separated by a total distance d . Thus, the length condition is

$$z_2 - z_1 = d. \quad (7.62)$$

(As drawn in the diagram, $z_2 > 0$ but $z_1 < 0$.) Matching the wavefront curvature to the curvature of mirror 1,

$$R(z_1) = R_1 \implies R_1 = z_1 + \frac{z_0^2}{z_1}. \quad (7.63)$$

At the second mirror,

$$R(z_2) = -R_2 \implies -R_2 = z_2 + \frac{z_0^2}{z_2}. \quad (7.64)$$

Note the difference in minus sign in the two cases, because the ray-optics convention for the mirror radius of curvature differs from the Gaussian-beam convention for the wave-front curvature $R(z)$: for concave mirrors

as in the diagram, for example, $R_{1,2} < 0$, but the matching wave-front curvature $R(z) \gtrless 0$ for $z \gtrless 0$. These two relations are also consistent even for convex-concave cavities, since the focus turns out to be outside the cavity.

These three equations have the solution

$$\begin{aligned} z_1 &= \frac{-d(R_2 + d)}{R_1 + R_2 + 2d} \\ z_2 &= \frac{d(R_1 + d)}{R_1 + R_2 + 2d} \\ z_0^2 &= \frac{-d(R_1 + d)(R_2 + d)(R_1 + R_2 + d)}{(R_1 + R_2 + 2d)^2}. \end{aligned} \quad (7.65)$$

We have determined the location of $z = 0$ relative to the mirrors along with z_0 , so we have completely determined the Gaussian beam geometry.

7.5.1 Physical Modes

For a physical (bounded) mode, the Rayleigh length z_0 must be real. If z_0 is imaginary, $w^2(z)$ becomes negative at large z , and thus the Gaussian amplitude factor $\exp[-r^2/w^2(z)]$ diverges for large r . Thus, the physical mode condition is

$$z_0^2 \geq 0. \quad (7.66)$$

Inserting the expression from Eqs. (7.65) and performing much algebra, we find that this condition becomes

$$0 \leq g_1 g_2 \leq 1, \quad g_{1,2} = 1 + \frac{d}{R_{1,2}}, \quad (7.67)$$

which is just the same as the stability condition from ray optics.

7.5.2 Symmetric Cavities

In the special case of a symmetric resonator, $-R_1 = -R_2 =: R > 0$ (i.e., both mirrors are concave), the solutions (7.65) simplify greatly. By symmetry we can guess that the focus occurs at the center of the cavity. Also,

$$\begin{aligned} z_1 &= \frac{-d}{2} \\ z_2 &= \frac{d}{2} \\ z_0 &= \frac{d}{2} \sqrt{\frac{2R}{d} - 1}. \end{aligned} \quad (7.68)$$

The corresponding beam waist at the focus is

$$w_0^2 = \frac{\lambda d}{2\pi} \sqrt{\frac{2R}{d} - 1}, \quad (7.69)$$

and the beam waist at the mirrors is

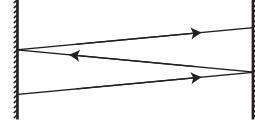
$$w_{1,2}^2 = w^2(z_{1,2}) = \frac{\lambda d/\pi}{\sqrt{\left(\frac{d}{R}\right)\left(2 - \frac{d}{R}\right)}}. \quad (7.70)$$

Note that for $z_0 \in \mathbb{R}$, we must have $R > d/2$, which is the stability condition of a symmetric cavity.

7.5.3 Special Cavities

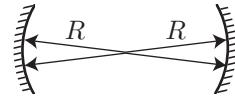
Recall the special resonator cases from Section 2.7.3. Let's look at the symmetric ones again in the context of Gaussian-beam modes.

Case 1. $g_1 = g_2 = 1 \implies R_{1,2} = \infty$ (Planar resonator)



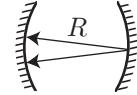
In this case, $z_0 \rightarrow \infty$, $w_0 \rightarrow \infty$, and $w_{1,2} \rightarrow \infty$. This is the plane-wave limit of the Gaussian beam, so we recover our analysis of the planar cavity.

Case 2. $g_{1,2} = -1 \implies R_{1,2} = -d/2$ (Spherical resonator/symmetrical concentric)



Here $z_0 \rightarrow 0$, $w_0 \rightarrow 0$, and $w_{1,2} \rightarrow \infty$. This is the spherical-wave limit of the Gaussian-beam mode, which is what we expect in the limit of a spherical boundary.

Case 3. $g_{1,2} = 0 \implies R_{1,2} = -d$ (Confocal resonator)



For the confocal resonator, $z_0 = d/2$, $w_0 = \sqrt{\lambda d/2\pi}$, and $w_{1,2} = \sqrt{2}w_0$. These relations imply that each mirror is a distance z_0 from the focus. Alternately, the focal length of both mirrors is $f = -R/2 = d/2 = z_0$, so the cavity length is $2z_0$. Again, this is the reason that z_0 is also called the **confocal parameter**.

7.5.4 Resonance Frequencies

Recall from Eq. (6.28) that the total phase of a Gaussian beam is

$$\phi(r, z) = kz - \tan^{-1} \left(\frac{z}{z_0} \right) + k \frac{r^2}{2R(z)}. \quad (7.71)$$

We already matched the shape of the phase fronts to the mirrors, so it suffices to consider the on-axis phase $\phi(0, z)$. Thus, we will only consider the plane-wave and Gouy phases:

$$\phi(0, z) = kz - \tan^{-1} \left(\frac{z}{z_0} \right). \quad (7.72)$$

The round-trip phase is

$$\Delta\phi_{rt} = 2[\phi(0, z_2) - \phi(0, z_1)] = 2k(z_2 - z_1) - \left[\tan^{-1} \left(\frac{z_2}{z_0} \right) - \tan^{-1} \left(\frac{z_1}{z_0} \right) \right]. \quad (7.73)$$

For constructive interference, this phase change must be $\Delta\phi_{rt} = 2\pi q$ for some integer q . Solving for k and using $z_2 - z_1 = d$,

$$k_q = \frac{\pi}{d}q + \frac{1}{d} \left[\tan^{-1} \left(\frac{z_2}{z_0} \right) - \tan^{-1} \left(\frac{z_1}{z_0} \right) \right]. \quad (7.74)$$

Rewriting this in terms of frequency using $k = 2\pi\nu/c$,

$$\nu_q = (\text{FSR}) \left\{ q + \frac{1}{\pi} \left[\tan^{-1} \left(\frac{z_2}{z_0} \right) - \tan^{-1} \left(\frac{z_1}{z_0} \right) \right] \right\}. \quad (7.75)$$

Using Eqs. (7.65), we can write down a mess of stuff that is difficult to simplify because of the signs of the factors under the radical signs. Through a rather long sequence of algebraic steps, the contribution of the Gouy phase reduces to a startlingly simple expression,

$$\nu_q = (\text{FSR}) \left(q + \frac{1}{\pi} \cos^{-1} \sqrt{g_1 g_2} \right), \quad (7.76)$$

where again $g_{1,2} = 1 + d/R_{1,2}$. The Gouy-phase term (with the \cos^{-1}) is again a geometry-dependent term that is independent of q . Since q is very large at optical frequencies, this term leads to an *overall shift* of the spectrum that is small compared to the optical frequency ν_q . However, the frequency shift is not necessarily small on the scale of the free spectral range. The shift vanishes for in the planar and spherical-concentric cavities, but for the confocal limit, the Gouy frequency shift is exactly half of the free spectral range.

7.5.5 Algebraic Digression

How does one go from Eq. (7.75) to Eq. (7.76)? A (very) brief outline of the algebra is as follows. First, write the beam-parameter solutions of Eqs. (7.65) in terms of the stability parameters, with the result

$$\begin{aligned} \frac{z_1}{d} &= \frac{-g_2(1-g_1)}{g_1+g_2-2g_1g_2} \\ \frac{z_2}{d} &= \frac{g_1(1-g_2)}{g_1+g_2-2g_1g_2} \\ \frac{z_0^2}{d^2} &= \frac{g_1g_2(1-g_1g_2)}{(g_1+g_2-2g_1g_2)^2}. \end{aligned} \quad (7.77)$$

In terms of these parameters, the signs are easier to handle since $g_1g_2 > 0$ for a stable cavity. Then it is relatively easy to show that

$$\cos \left[\tan^{-1} \left(\frac{z_2}{z_0} \right) - \tan^{-1} \left(\frac{z_1}{z_0} \right) \right] = \frac{z_0^2 + z_1 z_2}{\sqrt{(z_0^2 + z_1^2)(z_0^2 + z_2^2)}} = \sqrt{g_1 g_2} \quad (7.78)$$

after a bit of algebraic manipulation.

7.6 Spherical-Mirror Cavities: Hermite–Gaussian Modes

The same idea applies to the higher-order $\text{TEM}_{l,m}$ modes of the cavities, since the wave fronts off all the higher-order modes are essentially the same as for the Gaussian. The main difference is that the Gouy term in the on-axis phase becomes

$$\phi_{l,m}(0, z) = kz - (1+l+m) \tan^{-1} \left(\frac{z}{z_0} \right) \quad (7.79)$$

for the $\text{TEM}_{l,m}$ mode, as we can see by examination of Eq. (6.74). Repeating the above argument shows that only the Gouy term is modified, so that

$$\nu_{l,m,q} = (\text{FSR}) \left[q + \frac{1}{\pi} (1+l+m) \cos^{-1} \sqrt{g_1 g_2} \right]. \quad (7.80)$$

Now the Gouy term represents a frequency shift that depends on the cavity geometry *and* the order of the Hermite–Gaussian mode. Thus, the cavity modes must be labeled by three indices as shown. The

terminology is that modes corresponding to different q values are the **longitudinal modes**, which are similar in structure to the plane-wave resonances. The modes corresponding to different values of l and m are the different **transverse modes**, since they have different transverse intensity patterns. Note that there are a number of degeneracies among the modes due to the axial symmetry of the resonator.

7.6.1 Confocal Cavity

Why was the laboratory cavity in the photographs (see p. 123) a confocal cavity? It would seem that a cavity with spherical mirrors would have a more complicated mode structure than the planar cavity and thus more difficult to use as an optical spectrum analyzer. As it turns out, though, the mode structure is simple for this case as well. For a symmetric confocal cavity, $d = -R_1 = -R_2$, so $g_1 = g_2 = 0$. Then $g_1 g_2 = 0$, and since $\cos^{-1}(0) = \pi/2$, the mode frequencies become

$$\nu_{l,m,q} \text{ (confocal)} = (\text{FSR}) \left[q + \frac{1}{2}(1 + l + m) \right]. \quad (7.81)$$

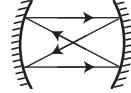
We can write this in terms of an effective index s as

$$\nu_s \text{ (confocal)} = (\text{FSR}_{\text{confocal}}) s, \quad (7.82)$$

where $s = 2q + 1 + l + m$ is a positive integer, and

$$\text{FSR}_{\text{confocal}} := \frac{\text{FSR}}{2} = \frac{c}{4d} = \frac{c_0}{4nd}, \quad (7.83)$$

is an effective free-spectral range for the confocal cavity. Thus, we recover the spectrum for a planar cavity, but with a length of $2d$. Recall the typical trajectory of the confocal cavity that makes 4 passes per round trip.



That's why the cavity is effectively twice as long as for the planar cavity; the off-axis modes travel twice as far before repeating themselves. And that's why laboratory spectrum analyzers are often confocal cavities: the modes are degenerate, in the sense that they all fall into two series of frequency resonances, according to whether s is even or odd (for example, $\nu_{0,0,q} = \nu_{1,1,q-1}$, but there is no equivalent frequency of the form $\nu_{1,0,q'}$ for any q').

Notice, however, that this argument only works if several transverse modes are excited by the input source. If only the $\text{TEM}_{0,0}$ mode is present, then the free spectral range is back to the old value. That's because *longitudinal* modes are still spaced by FSR , not $\text{FSR}_{\text{confocal}}$. The effective confocal free-spectral range only works if modes operate in the ring-trajectory manner shown. Thus, in the laboratory, the input beam to a confocal resonator is usually slightly misaligned to ensure that the “in-between” modes are populated as well, so the relevant free spectral range is $\text{FSR}_{\text{confocal}}$. A confocal cavity operated in this manner is said to be operated in “ring mode.” Notice also that the finesse effectively changes, since the line width is unaffected but the free spectral range is different. Thus,

$$\mathcal{F}_{\text{confocal}} = \frac{\text{FSR}_{\text{confocal}}}{\delta\nu_{\text{FWHM}}} = \frac{\mathcal{F}}{2} \quad (7.84)$$

for a confocal cavity in ring mode.

7.7 Exercises

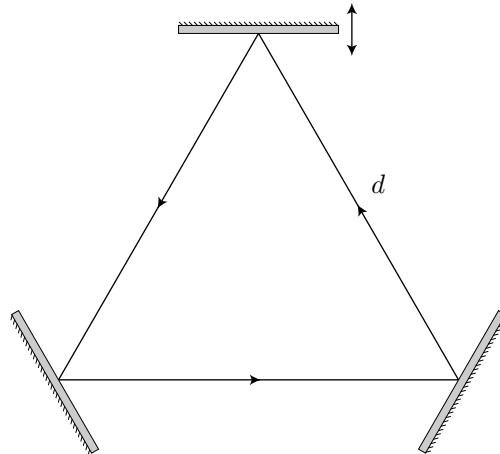
Problem 7.1

Consider a cavity constructed from two planar mirrors separated by 1 cm.

- What is the free spectral range if the cavity is completely empty?
- What is the free spectral range if the cavity is otherwise empty but a 1 mm glass plate ($n = 1.5$) is placed in the cavity? Ignore any reflections from the surfaces of the glass plate.

Problem 7.2

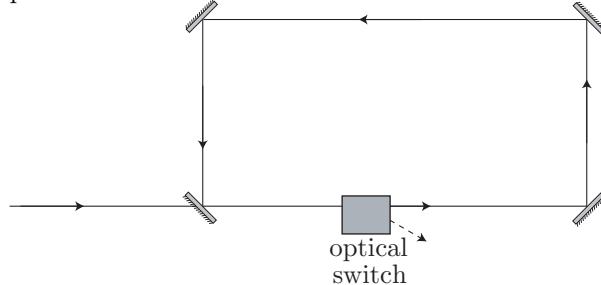
Consider a ring resonator consisting of three planar mirrors, arranged at the vertices of an equilateral triangle of side length d as shown in the diagram.



- What is the free spectral range of the cavity?
- Suppose that the top mirror is moved vertically as shown in the diagram. If the cavity is resonant with light of wavelength λ , by how much should the mirror move to go through one complete interference fringe?

Problem 7.3

One method for boosting the intensity of continuous-wave (cw) light is to use a *cavity dumper*, which is a ring cavity with an optical switch as shown.



The cavity dumper operates as follows: the input light is turned on, and the circulating light builds up in the cavity to a large intensity. Then the optical switch is activated, deflecting the circulating intensity out of the cavity.

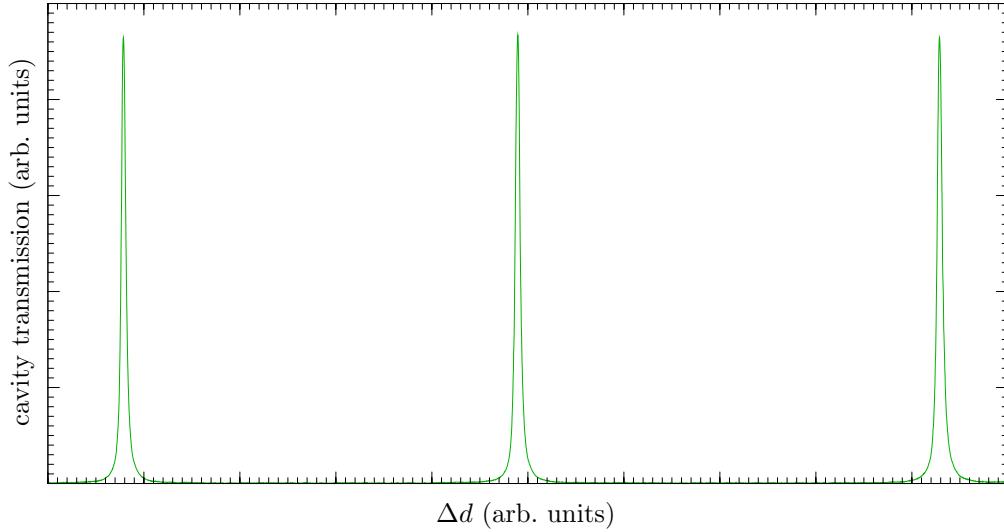
- What are the round-trip time and free spectral range? Assume the beam path is a rectangle of dimensions 12×20 cm. Also model the optical switch as a block of glass ($n = 1.5$) of length 2 cm.
- What are the survival probability, finesse, and photon lifetime of the cavity? Assume an intensity reflection coefficient of 99.8% for all the mirrors and an intensity loss of 0.2% per pass due to the switch.

- (c) Suppose the input power is 1 W and that the circulating power is allowed to build up to its steady-state value. What are the duration and power of the output pulse when the switch is activated? Assume the switch couples all the light out of the cavity.
- (d) What is the maximum repetition rate of the cavity dumper if a 50 W output pulse is desired?

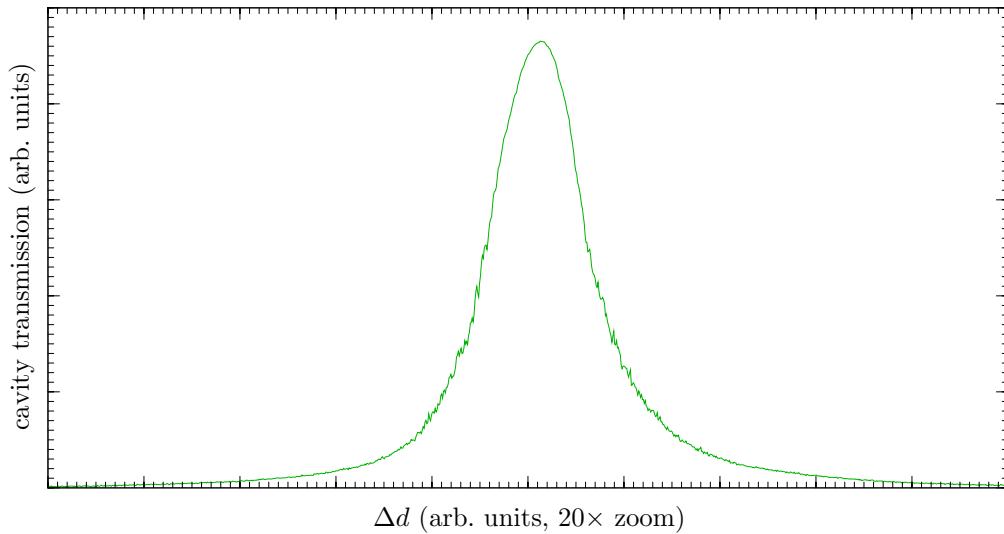
Problem 7.4

Consider a symmetric confocal optical spectrum analyzer operating at a nominal wavelength of 780 nm. The nominal cavity length is 5 cm, which can be changed precisely (on a sub- λ scale) using a piezo stack attached to one of the mirrors. The cavity transmission is monitored as the cavity length d is scanned.

- (a) Here is a measured transmission spectrum,

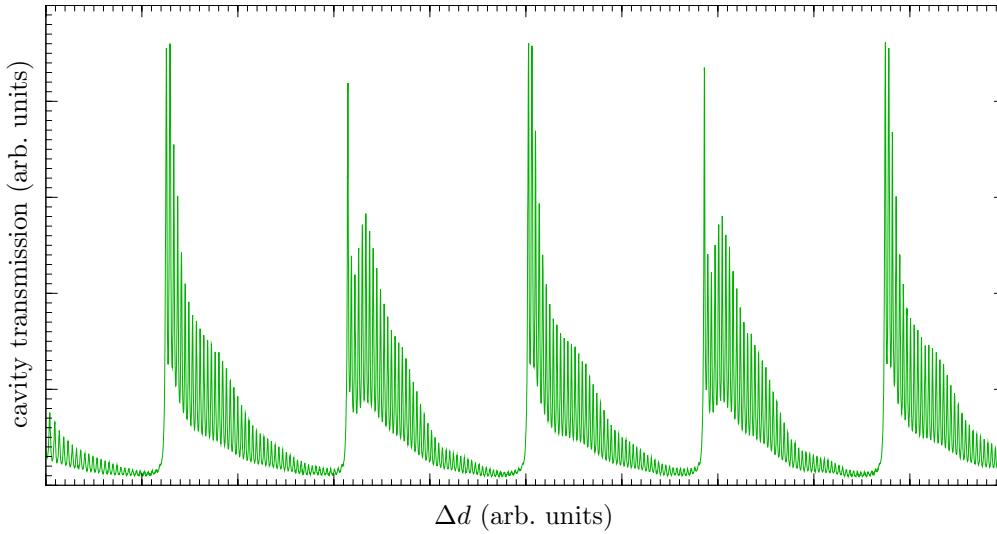


and this is the same set of data but with the horizontal axis zoomed by a factor of 20:

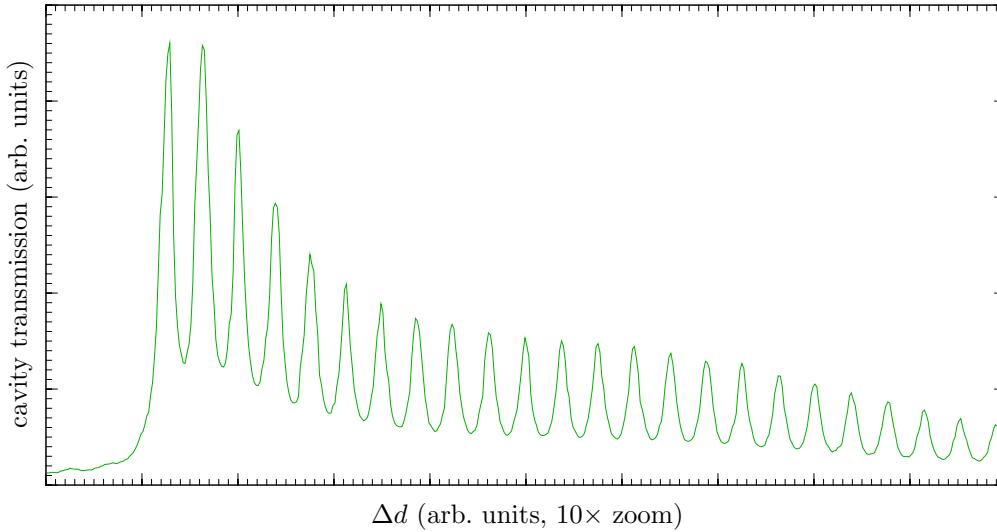


Compute the finesse and resolution $\delta\nu_{\text{FWHM}}$ of the cavity, keeping in mind that this cavity is *confocal*. (Note: in the first scan, the distances between the peaks are slightly different due to the nonlinearity of the piezo stack displacement with voltage; it is best to average the two distances to get an accurate value.)

- (b) The cavity length is changed by a relatively large amount (compared to λ , say by a threaded barrel adjuster, such that the cavity is slightly longer than required to be confocal. The measured spectrum now looks more complicated:



The same data zoomed in by a horizontal factor of 10 shows a structure of many peaks for each longitudinal mode:



Note that the horizontal units here do not correspond to the plots in (a). From the data here, estimate the new cavity length by using the cavity line splittings to calculate the new stability parameter.

Hint 1: The cavity no longer satisfies the confocal condition, so don't treat it as a confocal cavity. Note that the spectrum *almost* repeats after some distance and then *really* repeats after that.

Hint 2: You should calculate the new cavity length by understanding the above spectrum. However, as a sanity check, this new spectrum was taken when the threaded barrel that sets the overall cavity length was rotated by about $1\frac{1}{4}$ turns. The barrel was machined with a nominal pitch of 40 threads/inch.

Problem 7.5

One limitation to the finesse of a Fabry–Perot cavity is the surface roughness of the mirrors. As a simple estimate of this effect, consider a planar resonator of length d_0 with perfectly reflective mirrors.

(a) Recall the normalized form of the normal (Gaussian) distribution with zero mean,

$$f(x; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right). \quad (7.85)$$

Here, σ is the standard deviation or root-mean-square (rms) deviation. Compute the full width at half maximum δx_{FWHM} of this distribution in terms of σ .

(b) Model the effective cavity length as $d = d_0 + \delta d_1 + \delta d_2$, where the $\delta d_{1,2}$ are the (small) deviations due to mirror roughness. Assume that these are uncorrelated random variables over the mirror surfaces. Also assume that they are normally distributed with $\langle \delta d_{1,2} \rangle = 0$ and $\langle \delta d_{1,2}^2 \rangle = \varepsilon^2$ (i.e., the rms roughness is ε), where the angle brackets denote an average over the mirror surfaces. Show that the effective cavity finesse is

$$\mathcal{F} = \frac{\lambda}{8\sqrt{\log 2}\varepsilon} \quad (7.86)$$

due to the surface roughness.

(c) A typical rms surface roughness for a laser-grade mirror substrate is 3 Å, while a typical roughness for a superpolished mirror substrate is 0.5 Å. Assuming that the reflective coating of the mirror has the same roughness as the substrate, compute the roughness-limited finesse for each case at $\lambda = 780$ nm.

Problem 7.6

Show that if

$$\nu_q = (\text{FSR}) \left\{ q + \frac{1}{\pi} \left[\tan^{-1} \left(\frac{z_2}{z_0} \right) - \tan^{-1} \left(\frac{z_1}{z_0} \right) \right] \right\}. \quad (7.87)$$

is the resonance frequency for a Gaussian mode of a stable, spherical-mirror cavity, then the following expression is equivalent:

$$\nu_q = (\text{FSR}) \left(q + \frac{1}{\pi} \cos^{-1} \sqrt{g_1 g_2} \right). \quad (7.88)$$

Problem 7.7

Consider a planar cavity of nominal length $d = 1$ cm. The cavity is resonant with light of $\lambda = 1 \mu\text{m}$.

- (a) Compute the ratio Q/\mathcal{F} for this cavity at this wavelength.
- (b) Suppose that you change the wavelength by $\Delta\lambda$ before you find the next cavity resonance. What is $\Delta\lambda$?
- (c) Suppose that the mirrors have very high reflectivity, so that the cavity is “good.” State whether the following quantities would be larger, smaller, or the same as for a “bad” cavity of the same geometry:

1. survival probability
2. Q factor
3. finesse
4. round-trip time
5. free spectral range
6. resonance width ($\delta\nu_{\text{FWHM}}$)
7. photon lifetime

Problem 7.8

A laser of nominal wavelength $\lambda = 1 \mu\text{m}$ emits at two slightly different wavelengths, where the difference is $\Delta\lambda \ll \lambda$. The transmission of the laser light through a Fabry–Perot spectrum analyzer of nominal length $d = 5$ cm is observed as the analyzer’s length is scanned.

- (a) The transmission of the analyzer is observed to be periodic in the analyzer length with period Δd . What is Δd ?
- (b) Suppose that the analyzer is set so that one wavelength component of the laser is maximally transmitted. You observe that when the analyzer's length is changed by $\Delta d/10$, the *other* laser wavelength is maximally transmitted. What is $\Delta\lambda$?
- (c) The laser itself consists of a Fabry–Perot cavity filled with gas ($n \approx 1$). Suppose that the two output wavelengths correspond to two adjacent resonances of the laser cavity. What is the length of the laser cavity, assuming it is longer than the analyzer?

Problem 7.9

A laser of nominal wavelength $\lambda = 200\pi$ nm is on resonance with a Fabry–Perot spectrum analyzer of nominal length $d = 10$ cm. Suppose that the two mirrors are identical, with *intensity* reflection coefficients $R = 99\%$.

- (a) By how much must you increase d so that the laser excites the next cavity resonance?
- (b) By how much must the laser wavelength increase to go from one resonance to the next?
- (c) By how much must the laser frequency change to go from one resonance to the next?
- (d) Suppose again that the laser is tuned to the analyzer's resonance, where the analyzer is initially evacuated. Now suppose the analyzer is filled with a gas of refractive index n , and over the course of the filling, the laser ends up again on resonance, but two resonances away from the starting point. What is n ?
- (e) Going back to the initial setup, what fraction of the intensity transmits through the analyzer (on resonance)? By how much must the analyzer length change so that the transmitted intensity drops to half this value?

Problem 7.10

We showed [Eq. (7.29)] that the mean number of round trips that a photon survives in a resonator is

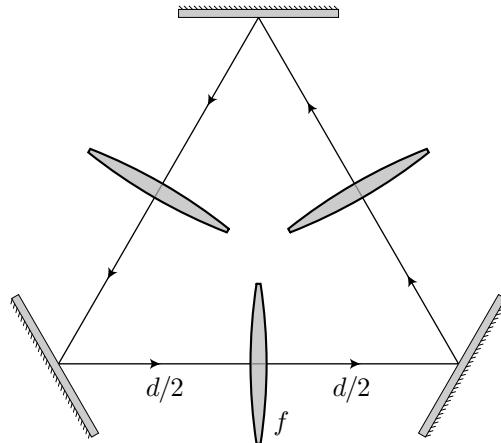
$$\langle n \rangle = \frac{1}{1 - P_s}. \quad (7.89)$$

Derive an expression for the variance $\Delta n^2 := \langle (n - \langle n \rangle)^2 \rangle = \langle n^2 \rangle - \langle n \rangle^2$.

Hint: try working with $\langle n(n + 1) \rangle$.

Problem 7.11

All parts of this problem refer to the resonator below, consisting of three flat mirrors, and a beam path that forms an equilateral triangle, where each side has length d . Each leg of the path has a thin lens of focal length f , centered between the two mirrors; each leg of the resonator is identical to the others. Assume the optics are surrounded by vacuum.



- (a) For what range of f is the resonator stable?
- (b) For any resonant mode of wavelength λ , by how much must d change to go from one resonance to the next?
- (c) Consider the $\text{TEM}_{0,0}$ mode of the resonator. Where do the points of minimum beam size (e.g., diameter) occur? Explain *briefly*.
- (d) Compute the beam-waist parameter w_0 at the locations in (c).
- (e) Suppose each mirror has a reflectance of R , and ignore any absorption losses in the mirrors, and also ignore any losses due to the lenses. Derive an expression for the steady-state circulating power inside the cavity, assuming the cavity is on resonance, and given an input power P_{in} . (Write your answer in terms of R and P_{in} only.)
- (f) Describe quantitatively (in the sense of defining a function you can plot; a plot is also okay if it is labeled well) the time dependence of the output-beam power at the lower-right mirror, given that an input beam of power P_{in} at the lower-left mirror is turned on suddenly, such that the initial turn-on transient hits the input mirror at $t = 0$. Assume the cavity is on resonance, and do not make any assumptions about whether the cavity is “good” or “bad.”
- (g) Suppose the resonator is flooded with a fluid of index n_{bath} . Give a new expression for the stability condition for the resonator, assuming the lenses have a refractive index of $n_{\text{lens}} > n_{\text{bath}}$.

Chapter 8

Polarization

8.1 Vector Plane Waves

Recall that a plane wave in complex notation has the form

$$\mathbf{E}^{(+)}(\mathbf{r}) = \mathbf{E}_0^{(+)} e^{ikz}, \quad (8.1)$$

where the vector amplitude has only transverse components:

$$\mathbf{E}_0^{(+)} = E_{0x}^{(+)} \hat{x} + E_{0y}^{(+)} \hat{y}. \quad (8.2)$$

The corresponding physical field, with the explicit time dependence, is

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}^{(+)}(\mathbf{r}, t) + \text{c.c.} = \mathbf{E}_0^{(+)} e^{i(kz - \omega t)} + \text{c.c.} \quad (8.3)$$

The **polarization** of the wave is related to the *direction* of $\mathbf{E}_0^{(+)}$. However, this direction is time dependent, so we need to work out some details before defining particular polarization states.

8.2 Polarization Ellipse

We will now examine the “trajectory” of the electric-field vector. Writing out the explicit phases of the components,

$$\begin{aligned} E_{0x}^{(+)} &= |E_{0x}^{(+)}| e^{i\phi_x} \\ E_{0y}^{(+)} &= |E_{0y}^{(+)}| e^{i\phi_y}, \end{aligned} \quad (8.4)$$

we can write the real field as

$$\mathbf{E}(\mathbf{r}, t) = E_x(\mathbf{r}, t) \hat{x} + E_y(\mathbf{r}, t) \hat{y}, \quad (8.5)$$

where

$$\begin{aligned} E_x &= E_{0x} \cos(kz - \omega t + \phi_x) \\ E_y &= E_{0y} \cos(kz - \omega t + \phi_y). \end{aligned} \quad (8.6)$$

Recall that for consistency among the real and complex notations, $E_{0x} = 2|E_{0x}^{(+)}|$ and $E_{0y} = 2|E_{0y}^{(+)}|$. These equations are the parametric equations for a general ellipse. With a bit of algebra it is possible to expand the cosines using the sum-angle formula and then eliminate the explicit time dependence to obtain the equation for an ellipse,

$$\left(\frac{E_x}{E_{0x}} \right)^2 + \left(\frac{E_y}{E_{0y}} \right)^2 - 2 \left(\frac{E_x E_y}{E_{0x} E_{0y}} \right) \cos \phi = \sin^2 \phi, \quad (8.7)$$

(polarization ellipse)

where $\phi := \phi_x - \phi_y$ is the relative phase between the components. This form isn't completely transparent, but it turns out that this is the relation for a ellipse *rotated* by an angle α , where

$$\tan 2\alpha = \frac{2E_{0x}E_{0y} \cos \phi}{E_{0x}^2 - E_{0y}^2}. \quad (8.8)$$

(rotation angle of polarization ellipse)

In the special case $\alpha = 0$, we recover the usual form for the unrotated ellipse,

$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 = 1, \quad (8.9)$$

of width $2E_{0x}$ and height $2E_{0y}$.

8.2.1 Simple Cases

We will now consider a few of the simplest polarizations, corresponding to special values of ϕ , E_x , and E_y .

1. **In-phase components** ($\phi = 0$). In this case, Eq. (8.7) reduces to

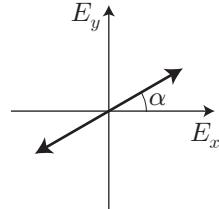
$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 - 2\left(\frac{E_x E_y}{E_{0x} E_{0y}}\right) = 0, \quad (8.10)$$

which is equivalent to

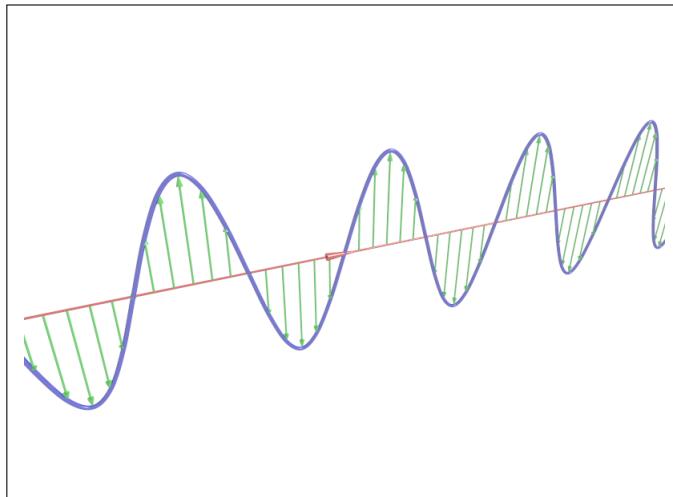
$$\frac{E_x}{E_{0x}} = \frac{E_y}{E_{0y}}. \quad (8.11)$$

Thus, the electric-field vector remains on a straight line in the E_x - E_y plane (say, at $z = 0$) at an angle α from the E_x -axis, where

$$\tan \alpha = \frac{E_{0y}}{E_{0x}} \quad (8.12)$$



Thus, this is the case of **linear polarization**. In the full three-dimensional space, the electric-field vector traces out a sinusoidal waveform, always staying within a single plane:

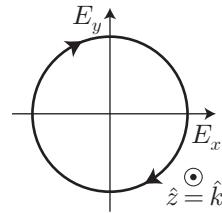


In general, the polarization is linear if $\phi = 0$ or π , or if either of E_{0x} or E_{0y} are zero (in which case ϕ may as well be zero).

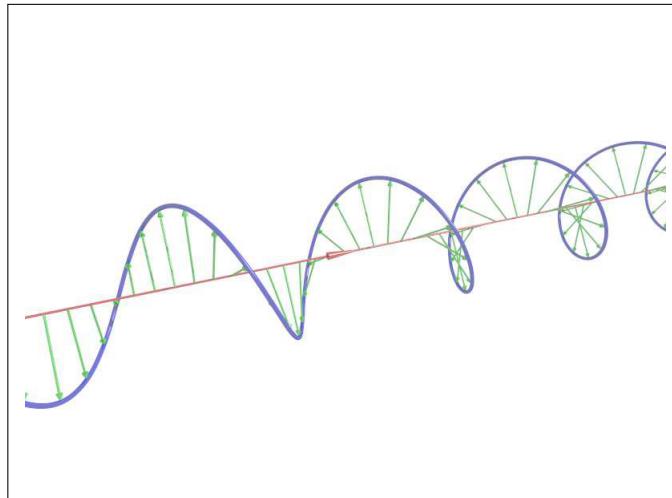
2. **Equal-magnitude components in quadrature** ($\phi = \pi/2$, $E_{0x} = E_{0y} =: E_0$). In this case, Eq. (8.7) for the polarization ellipse reduces to

$$E_x^2 + E_y^2 = E_0^2, \quad (8.13)$$

the equation for a circle of radius E_0 . We can work out the direction of the time-dependent motion of the vector as follows: At $z = 0$, the x -component has the form $\cos(\pi/2 - \omega t)$, while the y -component has the form $\cos(-\omega t)$; at $t = 0$, $(x, y) = (0, 1)$, which changes to $(1, 0)$ at $t = T/4$, where $T = 2\pi/\omega$ is the optical period. Thus, we see that the electric-field vector traces out a circle in the E_x - E_y plane in the *clockwise direction*, viewed such that the *propagation direction is out of the page* (i.e., along the $+z$ -direction):

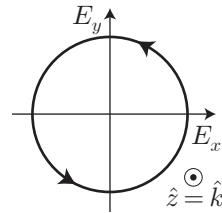


By convention, this polarization is called **right-circular polarization** (RCP). In three-dimensional space, the electric-field vector traces out a right-handed corkscrew, hence the name (curl your fingers in the direction of the electric-field trajectory, and your thumb points along the direction of propagation).

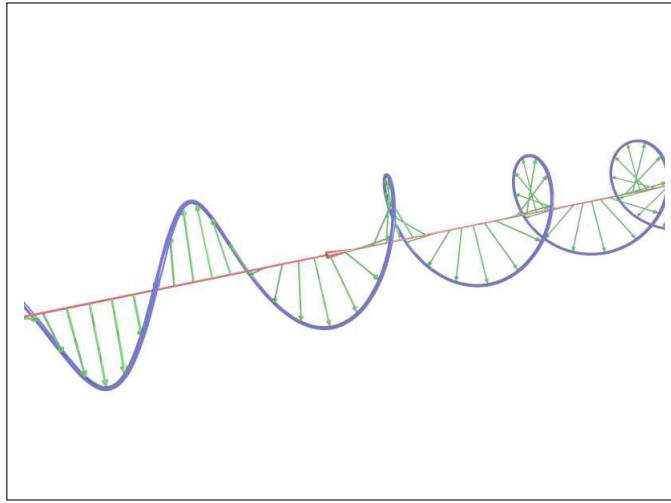


Note that in this picture, the wave propagates to the right.

3. **Equal-magnitude components in opposite quadrature** ($\phi = -\pi/2$, $E_{0x} = E_{0y} =: E_0$). This case is the same as for RCP light, but the electric-field vector travels *counterclockwise*:



Thus, this is **left-circular polarization** (LCP). In three-dimensional space, the electric-field vector traces out a left-handed corkscrew.



One important point to note from this analysis: the *only* difference between linear, RCP, and LCP light is the relative phase between the two component, at least in the proper coordinate system.

In the general case, as we noted before, the electric-field vector traces out some ellipse in the E_x - E_y plane, representing an elliptical polarization somewhere between linear and circular.

8.3 Polarization States: Jones Vectors

To make the notation more explicit, we can write out the electric-field vector $\mathbf{E}_0^{(+)}$ as an explicit two-component vector:

$$\mathbf{E}_0^{(+)} = \begin{bmatrix} E_{0x}^{(+)} \\ E_{0y}^{(+)} \end{bmatrix}. \quad (8.14) \quad (\text{Jones vector})$$

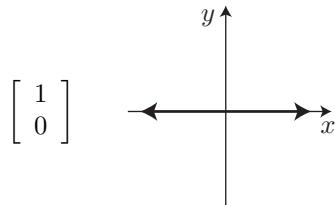
A **Jones vector** is simply this explicit two-component form. Thus, a Jones vector models the polarization state of an electromagnetic wave. Often to represent different polarizations the Jones vectors are written in normalized form, such that

$$|E_{0x}^{(+)}|^2 + |E_{0y}^{(+)}|^2 = 1, \quad (8.15)$$

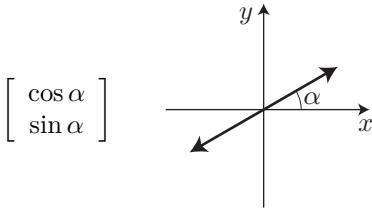
and so that the overall field is written as a separate factor—for polarization purposes, we only care about the magnitudes and phases of the two components *relative* to each other.

Thus, the some of the basic Jones vectors are as follows:

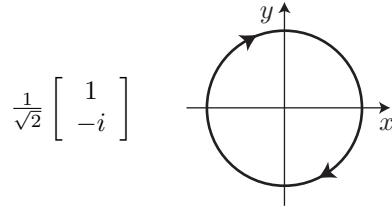
1. **linear polarization** (along the x -direction):



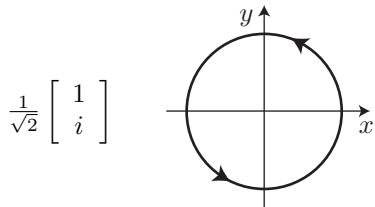
2. linear polarization (angle α):



3. right-circular polarization (RCP):



4. left-circular polarization (LCP):



Notice that the 90° phase shift for the circular polarizations are represented by the complex phase difference between the two components. Also, all of these vectors can have an extra overall phase, which doesn't amount to anything—it's like a translation in time but it doesn't affect the geometry of the way. Not unless, that is, there is interference, in which case the phase is *relative* to some other phase.

8.3.1 Vector Properties

Being vectors, Jones vectors inherit all the usual properties we associate with general vectors. We will just reiterate two of the more useful ones here.

Two Jones vectors are **orthogonal** if their inner product vanishes:

$$\left(\mathbf{E}_{01}^{(+)}\right)^* \cdot \mathbf{E}_{02}^{(+)} = \mathbf{E}_{01}^{(-)} \cdot \mathbf{E}_{02}^{(+)} = 0. \quad (8.16)$$

(orthogonality condition)

Note that since these are complex-valued vectors, the inner product involves complex conjugation of the first vector before the usual dot product.

Thus, by this definition, linear- x and linear- y polarizations are orthogonal, and LCP and RCP are also orthogonal.

Jones vectors also have a magnitude given by

$$\sqrt{\left|E_{0x}^{(+)}\right|^2 + \left|E_{0y}^{(+)}\right|^2}. \quad (8.17)$$

As we mentioned above, a Jones vector is normalized if its magnitude is unity.

8.4 Polarization Devices: Jones Matrices

A *polarization device* is some optical element or system that transforms the polarization state of an optical field. Since we are representing the state of the field by a Jones vector, the most general *linear* transformation of the polarization state is a 2×2 matrix. Thus, we can model any linear polarization device by a 2×2 matrix. Such a matrix is called a **Jones matrix**. Mathematically,

$$\mathbf{E}_{02}^{(+)} = \mathbf{T}\mathbf{E}_{01}^{(+)}, \quad (8.18)$$

(Jones-matrix transformation)

where \mathbf{T} is the Jones matrix modeling the system, $\mathbf{E}_{01}^{(+)}$ is the input field, and $\mathbf{E}_{02}^{(+)}$ is the output field.

8.4.1 Linear Polarizer

A *linear polarizer* (usually referred to simply as a “polarizer”) is a device that selectively annihilates the electric field along one direction. The corresponding Jones matrix for a polarizer that blocks the y -component of the field is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad (8.19)$$

(linear polarizer along x)

which induces the transformation

$$\begin{bmatrix} E_{0x}^{(+)} \\ E_{0y}^{(+)} \end{bmatrix} \longrightarrow \begin{bmatrix} E_{0x}^{(+)} \\ 0 \end{bmatrix}. \quad (8.20)$$

Thus, independent of the initial state of the field, the final state is linear- x . Thus, this is a polarizer with its **transmission axis** along the x -direction. Obviously, the Jones matrix for a polarizer that transmits along the y -axis is

$$\mathbf{T} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \quad (8.21)$$

(linear polarizer along y)

8.4.2 Wave Retarder

A **wave retarder** is a device that adds a phase shift to one component of the field. The Jones matrix for a wave retarder that retards the phase of the y -direction by $\Delta\phi$ is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & e^{i\Delta\phi} \end{bmatrix}. \quad (8.22)$$

(wave retarder along y)

This is a *retardation* of phase in *time*, since our phase convention here is $\exp(-i\omega t)$ for the time dependence. For this retarder, the x -axis is the “fast axis,” and the y -axis is the “slow axis.”

Let’s look at two special cases more closely.

$$\begin{aligned} \Delta\phi = \pi/2 &\longrightarrow \text{“quarter-wave plate” } (\lambda/4\text{-plate}) \\ \Delta\phi = \pi &\longrightarrow \text{“half-wave plate” } (\lambda/2\text{-plate}) \end{aligned} \quad (8.23)$$

(special wave retarders)

These two cases are especially important in the optics laboratory.

The quarter-wave plate induces the transformation

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 \\ i \end{bmatrix}, \quad (8.24)$$

and thus converts linear polarization (with the right orientation) to LCP (or RCP). This is one common laboratory method for producing circular polarization.

The half-wave plate induces the transformation

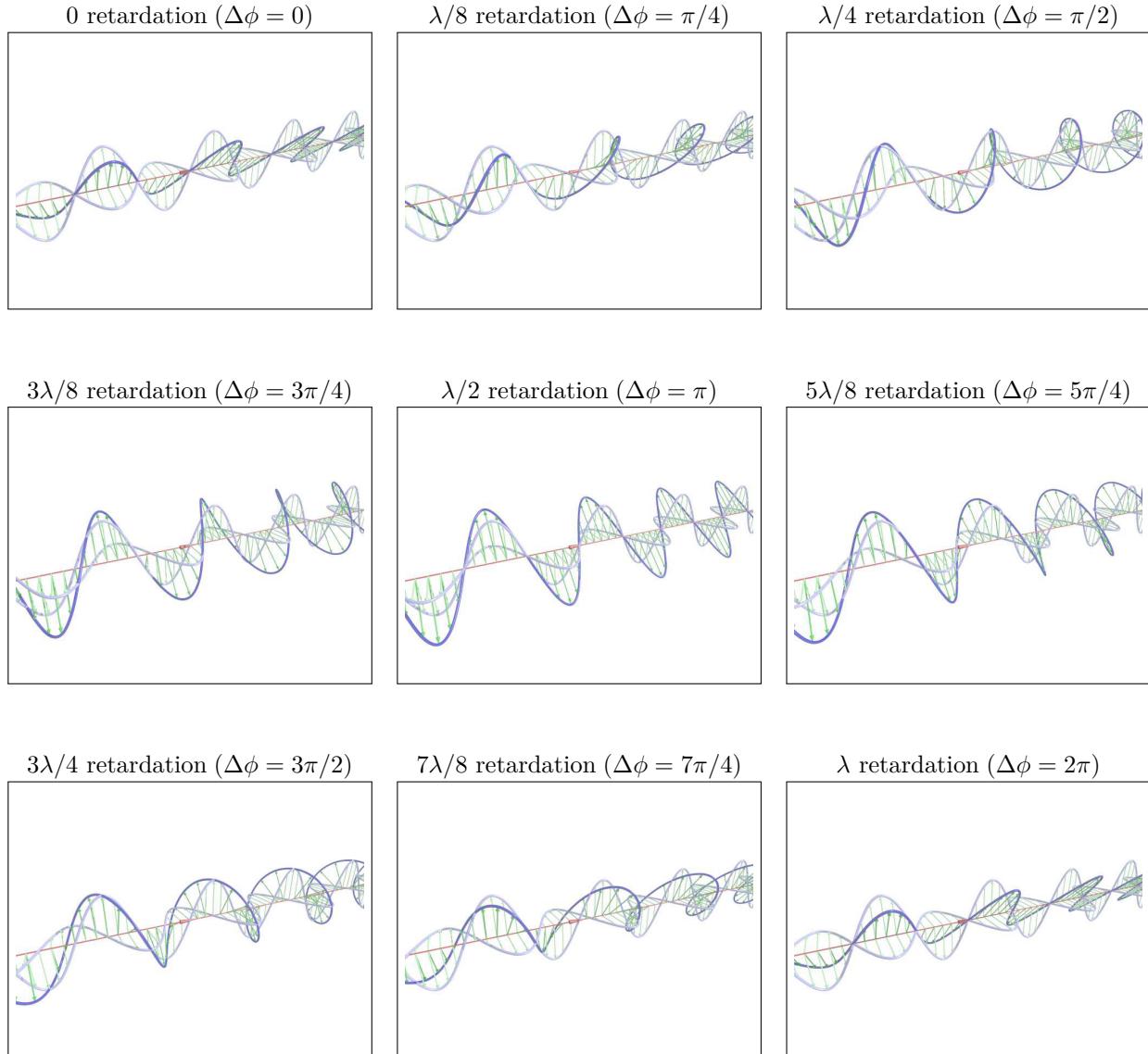
$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad (8.25)$$

and thus rotates linearly polarized light by 90° , at least if it has the right initial orientation (of 45° from the slow axis). More generally, the half-wave plate flips the polarization about the fast (x) axis.

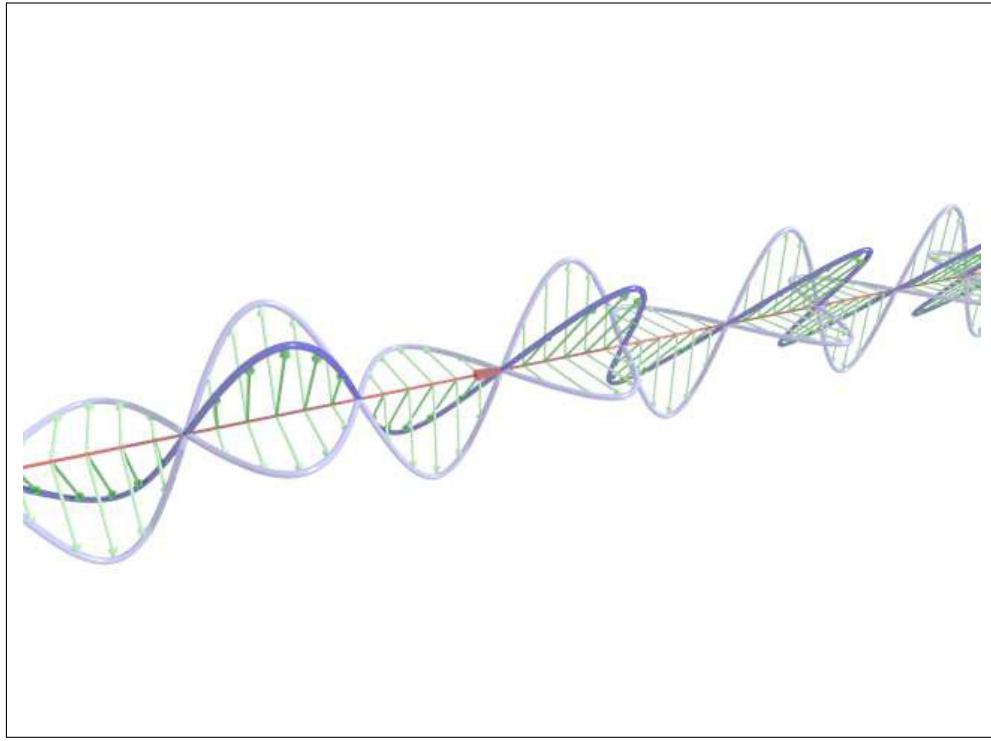
Now we can visualize how a wave retarder changes the polarization state of light. The wave propagation is to the right (along $+z$); the x -direction is vertical, and the y -direction is horizontal. The initial polarization state is linear at 45° , with Jones vector

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (8.26)$$

The wave is shown as the dark wave, and the x - and y - components are shown in a lighter color. As the y -component is retarded, the polarization changes to LCP ($\lambda/4$ retardation) to linear with a 90° rotation ($\lambda/2$ retardation), to RCP ($3\lambda/4$ retardation), and finally back to the original state (λ retardation).



The same evolution of the polarization is shown here as an animation in the electronic version of this document (currently this only works under Acrobat).



8.4.3 Polarization Rotator

The Jones matrix for a polarization rotator with rotation angle θ is

$$\mathbf{T}(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}. \quad (8.27)$$

(polarization rotator)

This matrix induces the transformation

$$\begin{bmatrix} \cos \theta_1 \\ \sin \theta_1 \end{bmatrix} \rightarrow \begin{bmatrix} \cos \theta_2 \\ \sin \theta_2 \end{bmatrix} \quad (8.28)$$

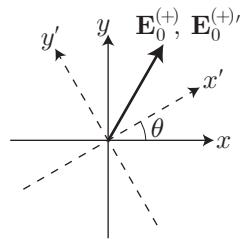
provided $\theta_2 = \theta_1 + \theta$.

8.4.4 Cascaded Systems

Of course, the power of this matrix formalism is that this works for cascaded systems via matrix multiplication, just like in ray optics (although there are now new matrices). Just as in ray optics, the order of matrix multiplication is critical: the first polarization device that the field encounters is the rightmost matrix in the product.

8.5 Coordinate Transformations

Consider a coordinate change from (x, y) coordinates to a new set of coordinates (x', y') that are rotated by an angle θ with respect to the original coordinates.



If $\mathbf{E}_0^{(+)}$ is the electric-field vector in the original coordinates, then the vector $\mathbf{E}_0'^{(+)}$ in the new coordinates is given by

$$\mathbf{E}_0'^{(+)} = \mathbf{R}(\theta)\mathbf{E}_0^{(+)}. \quad (8.29)$$

(effect of coordinate-system rotation)

Here, $\mathbf{R}(\theta)$ is the **rotation matrix**, defined by

$$\mathbf{R}(\theta) := \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}. \quad (8.30)$$

(rotation matrix)

Notice that the rotation of the *coordinates* by an angle θ is the same as a rotation of the *vector* by an angle $(-\theta)$, or $\mathbf{R}(\theta) = \mathbf{T}(-\theta)$. This is evident from the drawing, since the $\mathbf{E}_0'^{(+)}$ vector is closer to the x' -axis than the $\mathbf{E}_0^{(+)}$ vector is to the x -axis. Mathematically, you can see this by comparing the rotation matrix, Eq. (8.30), to the Jones matrix for the polarization rotator, Eq. (8.28)—that is, the Jones matrix for a polarization rotator of angle θ is $\mathbf{R}(-\theta)$.

What about rotations of polarization devices? That is, given a Jones matrix \mathbf{T} , what is the Jones matrix \mathbf{T}_θ for the same device rotated by an angle θ ? The intuitive way to handle this is to consider an equivalent picture where the field goes through the *unrotated* device. To compensate, the input field $\mathbf{E}_0^{(+)}$ should be rotated by an angle $(-\theta)$, so that the input to the rotated device is $\mathbf{R}(\theta)\mathbf{E}_0^{(+)}$. After passing through the device, we should restore things to the original coordinates by undoing the rotation, rotating the output vector by an angle θ . Thus, we have argued that

$$\mathbf{T}_\theta\mathbf{E}_0^{(+)} = \mathbf{R}(-\theta)\mathbf{T}\mathbf{R}(\theta)\mathbf{E}_0^{(+)}. \quad (8.31)$$

This should be true for any input polarization, so

$$\mathbf{T}_\theta = \mathbf{R}(-\theta)\mathbf{T}\mathbf{R}(\theta) \quad (8.32)$$

(Jones matrix for rotated optical device)

is the transformation law for rotating the polarization device by θ .

We can also derive this transformation more formally. The inner product of two vectors

$$\left(\mathbf{E}_{02}^{(+)}\right)^* \cdot \mathbf{E}_{01}^{(+)} \quad (8.33)$$

is a *scalar*, and thus is independent of coordinate rotations. Thus, we can consider the scalar quantity

$$\left(\mathbf{E}_{02}^{(+)}\right)^* \cdot \left(\mathbf{T}\mathbf{E}_{01}^{(+)}\right). \quad (8.34)$$

Rotating both the device and the vectors by an angle θ , we have

$$\left(\mathbf{R}(-\theta)\mathbf{E}_{02}^{(+)}\right)^* \cdot \left(\mathbf{T}_\theta\mathbf{R}(-\theta)\mathbf{E}_{01}^{(+)}\right). \quad (8.35)$$

These combined rotations are equivalent to a coordinate rotation by angle $(-\theta)$, so the two scalars are the same. Using the fact that the dot product is a matrix product with a transposition, $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b}$, the transpose property $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$, and the fact that $[\mathbf{R}(\theta)]^T = [\mathbf{R}(\theta)]^{-1} = \mathbf{R}(-\theta)$, we can rewrite the expression (8.35) as

$$\left(\mathbf{E}_{02}^{(+)}\right)^* \cdot \left(\mathbf{R}(\theta)\mathbf{T}_\theta\mathbf{R}(-\theta)\mathbf{E}_{01}^{(+)}\right). \quad (8.36)$$

For consistency of this expression with (8.34), we must have

$$\mathbf{T} = \mathbf{R}(\theta)\mathbf{T}_\theta\mathbf{R}(-\theta), \quad (8.37)$$

which we can rewrite as

$$\mathbf{T}_\theta = \mathbf{R}(-\theta)\mathbf{T}\mathbf{R}(\theta), \quad (8.38)$$

which is the same as Eq. (8.32).

As with the polarization vector, the effect of rotating the *coordinates* is the opposite:

$$\mathbf{T}' = \mathbf{R}(\theta)\mathbf{T}\mathbf{R}(-\theta).$$

(Jones matrix after coordinate rotation by θ) (8.39)

Keeping the signs straight is important, but it can be a bit tricky.

8.6 Normal Modes

Polarization devices are represented by matrices, which obviously have eigenvectors and eigenvalues. Recall that the eigenvectors (eigenpolarizations) are defined by the relation

$$\mathbf{T}\mathbf{E}_0^{(+)} = \lambda\mathbf{E}_0^{(+)}, \quad (8.40)$$

(normal-mode condition)

where λ (the eigenvalue) is a constant. There are at most two such vectors for a 2×2 matrix that satisfy this equation, and we can call them $\mathbf{E}_{0,\lambda_1}^{(+)}$ and $\mathbf{E}_{0,\lambda_2}^{(+)}$, for the eigenvectors corresponding to λ_1 and λ_2 , respectively. These eigenvectors are called the **normal modes** of a system.

The normal modes are easy to propagate through the system, since the effect of the Jones matrix is the same as multiplication by a constant factor. A vector can be decomposed into a superposition of the eigenvectors, and then easily propagated through the device:

$$\mathbf{T}\mathbf{E}_0^{(+)} = \lambda_1\alpha_1\mathbf{E}_{0,\lambda_1}^{(+)} + \lambda_2\alpha_2\mathbf{E}_{0,\lambda_2}^{(+)}. \quad (8.41)$$

(propagation via normal modes)

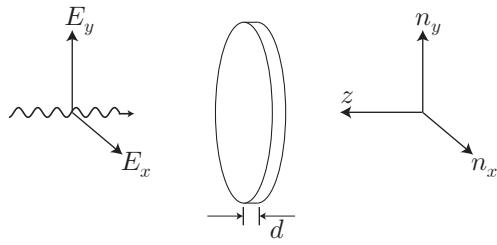
Note that in general, the eigenvalues could have a zero value (e.g., for a polarizer). Polarization devices can be lossy, so there is no general expression for the determinant of a Jones matrix, in contrast to the ray matrix case where the Hamiltonian structure ensures a nonvanishing determinant.

8.7 Polarization Materials

8.7.1 Birefringence

Now that we have covered the mathematical formalism of polarization, we will discuss some of the materials used to make polarization devices. One important class of optical materials are **birefringent** materials, where the material is *anisotropic*: the wave sees a different index of refraction for different polarizations. Materials could be birefringent, for example, due to anisotropic crystal structure or mechanical stress. Another example of material birefringence is the **electro-optic effect**, where certain crystals become birefringent when placed in a dc, uniform electric field. Examples of electro-optic crystals are lithium tantalate (LTA), potassium dihydrogen phosphate (KDP), potassium dideuterium phosphate (KD*P), and ammonium dihydrogen phosphate (ADP).

Consider a wave incident on a birefringent plate as shown.



Suppose that $n_y > n_x$. Then in the notation above, the y -direction is the “slow axis,” and the x -direction is the “fast axis.” Then the optical path length for x -polarization is

$$k_x d = \frac{2\pi n_x d}{\lambda_0}, \quad (8.42)$$

and the optical path length for the y -polarization is

$$k_y d = \frac{2\pi n_y d}{\lambda_0}. \quad (8.43)$$

Thus, the relative phase shift between the two components is $2\pi(n_y - n_x)d/\lambda_0$. So for example, we can make a $\lambda/4$ -plate by choosing

$$(n_y - n_x)d = \frac{\lambda_0}{4}. \quad (8.44) \quad (\lambda/4\text{-plate})$$

This gives a $\pi/2$ phase retarder along the y -direction, with Jones matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & e^{i\pi/2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}, \quad (8.45)$$

as we saw above.

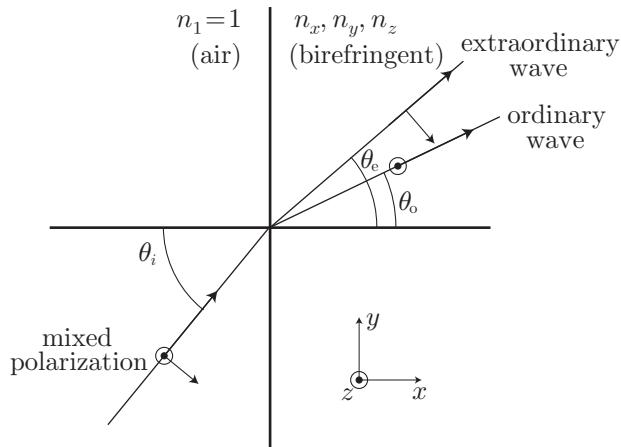
On the practical side, making a wave plate based on the criterion (8.44) can lead to impractically thin—and thus fragile—wave plates (which would only be a few optical wavelengths thick). More common is a *multiple-order wave plate*, which adds a large multiple of 2π in phase retardation. Thus, the criterion becomes something like

$$(n_y - n_x)d = \left(m + \frac{1}{4}\right)\lambda_0, \quad (8.46) \quad (\text{multiple-order } \lambda/4\text{-plate})$$

for a $\lambda/4$ -plate, where m is an integer that makes the resulting optic \sim mm thick. The wave plate represented by (8.44) is a **zero-order wave plate**, since it corresponds to $m = 0$. Multiple-order wave plates are relatively economical, but they suffer from a problem: wave plates are designed to achieve a certain retardation at a certain wavelength, but give different retardations at other wavelengths (since $d \propto \lambda_0$). The refractive indices also vary slightly with wavelength, which can further compound the problem. While such errors may be tolerable for a zero-order wave plate, any such errors are magnified by a factor of order m for a multiple-order wave plate. Zero-order wave plates are similarly much less sensitive to temperature variations, which cause the thickness of the plate to change. In practice, zero-order wave plates are constructed by combining two multiple-order wave plates of order with retardations of $(m + 1/4)\lambda_0$ and $m\lambda_0$ (in the case of a $\lambda/4$ -plate), with fast axes oriented at 90° with respect to each other. Thus, all the excess retardation cancels, giving the equivalent of a thin $\lambda/4$ -plate.

8.7.1.1 Multiple Refraction

Birefringent materials also have more complicated behavior for rays at non-normal incidence. The basic effect is **multiple refraction**, where different polarizations are refracted differently. Consider a wave with mixed polarization incident from free space into a birefringent medium.



Let's define a bit of notation that we will return to when introducing the Fresnel relations. The incident ray lies in the x - y plane, which we will call the **plane of incidence**. The polarization that lies inside the plane of incidence is **P-polarization** (for “parallel” to the plane of incidence), whereas the orthogonal polarization, normal to the plane of incidence, is **S-polarization** (for “senkrecht,” the German word for “perpendicular”).

The S-polarized component gives rise to the **ordinary wave**, which sees only the index n_z in the material. Thus, Snell's law gives

$$\sin \theta_i = n_z \sin \theta_o. \quad (8.47)$$

(ordinary-wave refraction)

The P-polarized component gives rise to the **extraordinary wave**, which is more complicated, because it sees an angle-dependent index:

$$n(\theta_e) = n_y \cos \theta_e + n_x \sin \theta_e. \quad (8.48)$$

(refractive index for extraordinary wave)

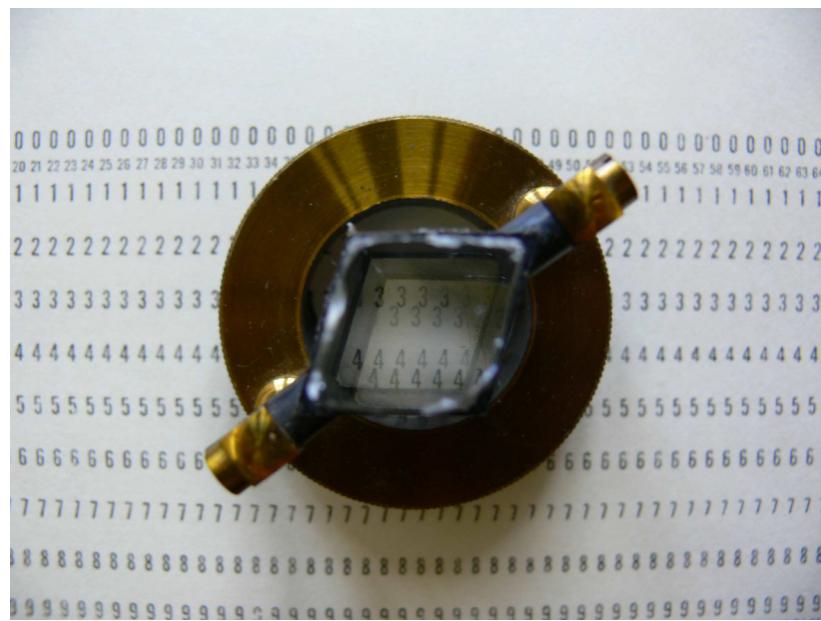
Thus, Snell's law cannot be solved analytically for θ_e :

$$\sin \theta_i = n(\theta_e) \sin \theta_e. \quad (8.49)$$

(extraordinary-wave refraction)

First, this relation leads to a different refraction angle for the extraordinary wave, compared to the ordinary wave. In some cases, this is a useful effect. For example, the separation of polarizations is useful in constructing polarizers (polarizing beam splitters). But multiple refraction is even a bit more subtle. It turns out that for the extraordinary waves, the geometrical rays are not in general orthogonal to the wave fronts (hence the name “extraordinary”). For this reason, it is even possible to get separation of the polarizations for normal incidence.

Multiple refraction due to a birefringent calcite crystal is shown here in this photograph.



8.7.2 Optical Activity

Whereas a birefringent material has different refractive indices for different *linear* optical polarizations, an **optically active** material has different indices for the two *circular* polarizations. In mathematical notation, we can denote the two indices by n_+ and n_- for RCP and LCP light, respectively.

Optical activity is characteristic of media with helical molecules. The helicity of the medium breaks the symmetry between the two polarizations. Examples of such media include quartz and sugar in the solid phase, or sugar solution and turpentine in the liquid phase. Note that given a material with a particular helicity, the same material with the opposite helicity exists in principle as well—it would just be the mirror image of the original material. For example, quartz can have $n_+ > n_-$ or $n_+ < n_-$, depending on the crystal structure. Glucose and fructose are two different sugars that give rise to opposite senses of optical activity.

One other important manifestation of optical activity is the **Faraday effect**. Some crystals, such as YIG (yttrium-indium-garnet), exhibit optical activity when they are placed in a uniform magnetic field.

Optical activity causes **optical rotation** of input linear polarization. To see this, note that linearly polarized light (say, linear- x) is a linear combination of RCP and LCP:

$$\mathbf{E}_{0,\text{in}}^{(+)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 \\ i \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ -i \end{bmatrix} \quad (8.50)$$

After passing through an optically active medium of thickness d , the Jones vector becomes

$$\mathbf{E}_{0,\text{out}}^{(+)} = \frac{1}{2} \begin{bmatrix} 1 \\ i \end{bmatrix} e^{i\phi_+} + \frac{1}{2} \begin{bmatrix} 1 \\ -i \end{bmatrix} e^{i\phi_-}, \quad (8.51)$$

where $\phi_{\pm} = k_{\pm}d = 2\pi n_{\pm}d/\lambda_0$. We can rewrite this in the form

$$\mathbf{E}_{0,\text{out}}^{(+)} = e^{i\phi_0} \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix}, \quad (8.52)$$

where $\phi = (\phi_- - \phi_+)/2$ and $\phi_0 = (\phi_- + \phi_+)/2$. Thus, linear polarization is rotated by the angle ϕ . It is more useful to specify a rotation per unit length,

$$\frac{d\phi}{dz} = \frac{\pi(n_- - n_+)}{\lambda_0}. \quad (8.53) \quad (\text{polarization rotation per unit length})$$

For the Faraday effect, it is conventional to specify the rotation per unit length as

$$\frac{d\phi}{dz} = VB, \quad (8.54)$$

(Faraday effect)

where the constant V is the **Verdet constant**. The rotation is typically proportional to the magnetic field in regimes of interest for laboratory polarization rotators.

8.8 Exercises

Problem 8.1

Write down the Jones matrix for an ideal polarizer that transmits x -polarized light. A realistic polarizer along the x -direction blocks some fraction α (of the *electric field*) of the x polarization, and transmits some fraction β of the y polarization. (Both α and β should be small for a decent polarizer.) Write down the Jones matrix for this realistic polarizer.

Problem 8.2

(a) Consider a linear polarizer and a wave linearly polarized at an angle θ with respect to the polarizer's transmission axis. Show that the *intensity* of the wave is reduced by $\cos^2 \theta$ after passing through the polarizer (this is called the *Law of Malus*).

(b) Consider a system of N cascaded polarizers. The polarizers have their transmission axes at angles $\pi/2N, 2\pi/2N, 3\pi/2N, \dots, \pi/2$ from the x -axis, in the order that an input wave sees them. That is, the last polarizer is oriented along the y -direction. Suppose that an input wave is polarized in the x -direction. Compute the intensity transmission coefficient for the system. Show that the transmission coefficient approaches unity as $N \rightarrow \infty$. This is a simple realization of the *quantum Zeno effect*, where each polarizer acts as a “measurement” of the polarization state—the polarization is “dragged” by the measurements as long as they are sufficiently frequent.

Note that it *isn't* sufficient to merely argue that $\cos(\pi/2N) \rightarrow 0$ as $N \rightarrow \infty$, because the interplay of the cosine with the exponent is nontrivial. In particular, it is critical that the first-order term in $1/N$ vanishes for the cosine, while the exponent scales as N . For example, if the exponent scales more quickly with N ,

$$\lim_{N \rightarrow \infty} \left[\cos \left(\frac{\pi}{2N} \right) \right]^{N^4} = 0, \quad (8.55)$$

then we can have convergence to other values.

Problem 8.3

One important optical polarization system is the *optical isolator*, which prevents laser light from being reflected back into the laser (which could cause instability or even damage). An optical isolator typically consists of a polarizer (say, along x), followed by a 45° polarization rotator, followed by a polarizer at 45° . Note that if \mathbf{T} is the Jones matrix for an optic, the Jones matrix for the optic rotated by an angle θ is $\mathbf{T}_\theta = \mathbf{R}(-\theta) \mathbf{T} \mathbf{R}(\theta)$, where $\mathbf{R}(\theta)$ is the rotation matrix.

(a) Derive the Jones matrices for forward and backward propagation through the isolator. Assume ideal polarizers and note that the rotator produces the same rotation independent of the light's direction. Show that x -polarized light passes forward through the system unattenuated, but that *any* light traveling backwards through the isolator will be completely extinguished. (Realistic isolators attenuate the reverse beam by about 40 dB.)

(b) A cheaper isolator is a polarizer (say, along x) followed by a quarter-wave retarder oriented at 45° . Show that if a laser passes through this isolator, the isolator will block any light returning from a direct mirror reflection, but will not block arbitrary return polarizations.

Problem 8.4

A beam of light can be described by its position and direction as a geometrical ray as well as by its polarization Jones vector. Suppose the position/direction vector and polarization vector transform respectively according to

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \mathbf{M} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}; \quad \begin{bmatrix} A_{x,2} \\ A_{y,2} \end{bmatrix} = \mathbf{T} \begin{bmatrix} A_{x,1} \\ A_{y,1} \end{bmatrix} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \begin{bmatrix} A_{x,1} \\ A_{y,1} \end{bmatrix}. \quad (8.56)$$

Write down an expression for the matrix Ω that models both aspects of the optical system; that is, the matrix that gives the transformation

$$\begin{bmatrix} y_2 \\ \theta_2 \\ A_{x,2} \\ A_{y,2} \end{bmatrix} = \Omega \begin{bmatrix} y_1 \\ \theta_1 \\ A_{x,1} \\ A_{y,1} \end{bmatrix}. \quad (8.57)$$

Problem 8.5

One method of Q -switching a laser is to include a polarizer and a switchable retarder (Pockels cell) oriented at 45° in the cavity. If we ignore losses other than due to the polarization rotation, the cavity finesse is unchanged. To “ Q -spoil” the cavity, we can switch on some phase retardation to introduce extra loss. Calculate the cavity finesse for the ring cavity of Problem 1 without the gain medium, for retardations of $\Delta\phi = 0$, $\Delta\phi = \pi/4$, and $\Delta\phi = \pi/2$. (Obviously you should not make any low-loss approximations here.)

Problem 8.6

Suppose you have a laser beam parallel to the optical table, and the light polarization is perpendicular to the table. You need light that is polarized parallel to the table. How can you do this with only two mirrors? (And without turning the laser on its side!)

Problem 8.7

Consider a planar cavity. Suppose that an ideal (linear) polarizer and a polarization rotator (which rotates the polarization by $\Delta\theta$) are placed inside the cavity. Assuming the mirrors are perfect reflectors, what is the finesse of the cavity?

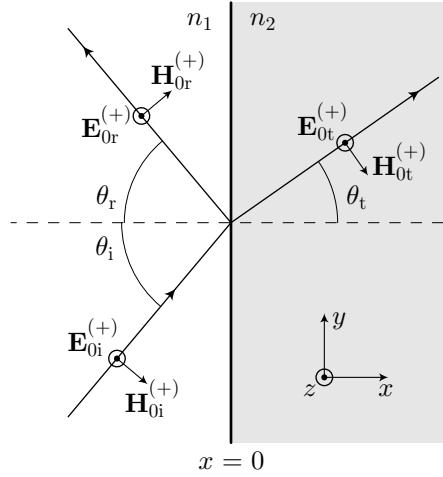
Chapter 9

Fresnel Relations

9.1 Optical Waves at a Dielectric Interface

One situation where the vector nature of light is crucial is in the reflection and refraction at a dielectric interface. In Chapter 2, we considered reflection and refraction in ray optics as a consequence of Fermat's principle. But that only fixed the reflection and refraction *angles*. We need wave optics to establish the *amplitude* of the reflected and refracted components.

Consider the setup in the diagram. We are assuming *a priori* that there is an incident, reflected, and transmitted wave.



We will assume that the waves are monochromatic, so that the three fields have the form

$$\begin{aligned} \mathbf{E}_i^{(+)} &= \mathbf{E}_{0i}^{(+)} e^{i\mathbf{k}_i \cdot \mathbf{r}} \\ \mathbf{E}_r^{(+)} &= \mathbf{E}_{0r}^{(+)} e^{i\mathbf{k}_r \cdot \mathbf{r}} \\ \mathbf{E}_t^{(+)} &= \mathbf{E}_{0t}^{(+)} e^{i\mathbf{k}_t \cdot \mathbf{r}}, \end{aligned} \quad (9.1)$$

where the subscripts “i,” “r,” and “t” refer to the incident, reflected, and transmitted waves, respectively. Note that we are making some pretty specific assumptions about the field, since we are anticipating the result. We can make whatever assumptions we want, so long as the solution turns out to be self-consistent—the self-consistency justifies the assumptions. The thing to worry about is the generality of the solution, but we will return to this later.

These three waves are not independent: the electromagnetic *boundary conditions* at the dielectric interface impose constraints among the three waves. The electromagnetic boundary conditions are

1. The *tangential* components (to the dielectric interface) of \mathbf{E} and \mathbf{H} are continuous across the boundary.

2. The *normal* components of \mathbf{D} and \mathbf{B} are continuous across the boundary.

These boundary conditions follow directly from the Maxwell equations, most straightforwardly in integral form. It turns out that the first boundary condition is sufficient to completely constrain the fields.

Let's now apply the boundary conditions. First, we'll use the fact that $\mathbf{E} \cdot \hat{z}$ (the tangential component) is continuous across the boundary:

$$(\mathbf{E}_{0i}^{(+)} \cdot \hat{z}) e^{ik_i y \sin \theta_i} + (\mathbf{E}_{0r}^{(+)} \cdot \hat{z}) e^{ik_r y \sin \theta_r} = (\mathbf{E}_{0t}^{(+)} \cdot \hat{z}) e^{ik_t y \sin \theta_t}. \quad (9.2)$$

The left-hand side is the sum of the incident and reflected waves, while the right-hand side is the transmitted wave. We have defined our coordinate system such that the interface is located at $x = 0$.

For this to be true for all y , the exponents must *all* be identically equal: the phases of the waves must agree everywhere along the boundary. For the phases of the incident and reflected waves to be equal, we require

$$k_i \sin \theta_i = k_r \sin \theta_r. \quad (9.3)$$

But $k_i = k_r$, since both waves see the same refractive index, and $k = nk_0$, where k_0 is the vacuum wave number. Thus, we recover the law of reflection:

$$\theta_i = \theta_r. \quad (9.4)$$

(Law of Reflection)

Equating the incident and transmitted phases,

$$k_i \sin \theta_i = k_t \sin \theta_t. \quad (9.5)$$

This simplifies to give Snell's Law.

$$n_1 \sin \theta_i = n_2 \sin \theta_t. \quad (9.6)$$

(Snell's Law)

Thus, we have recovered everything that we derived from ray optics. The *geometrical* properties of reflection and refraction are purely a consequence of the wave geometry.

Now to consider the amplitudes. The continuity of the electric field becomes

$$\mathbf{E}_{0i}^{(+)} + \mathbf{E}_{0r}^{(+)} = \mathbf{E}_{0t}^{(+)}, \quad (9.7)$$

because the electric fields lie entirely along the z -direction. We can rewrite this relation as

$$\frac{E_{0i}^{(+)} + E_{0r}^{(+)}}{E_{0t}^{(+)}} = 1, \quad (9.8)$$

which will be more convenient for later manipulations.

The continuity of \mathbf{H} is only slightly more complicated. The tangential component is the component in the y -direction, so the continuity condition becomes

$$\mathbf{H}_{0i}^{(+)} \cdot \hat{y} + \mathbf{H}_{0r}^{(+)} \cdot \hat{y} = \mathbf{H}_{0t}^{(+)} \cdot \hat{y}, \quad (9.9)$$

where we have already used the equality of the phases. Using the fact that the magnetic and electric fields are related by $H_0 = E_0/\eta = nE_0/\eta_0$, where η is the wave impedance and η_0 is the vacuum wave impedance, this boundary condition becomes

$$-\frac{n_1 E_{0i}^{(+)}}{\eta_0} \cos \theta_i + \frac{n_1 E_{0r}^{(+)}}{\eta_0} \cos \theta_r = -\frac{n_2 E_{0t}^{(+)}}{\eta_0} \cos \theta_t. \quad (9.10)$$

Again, we can rewrite this as

$$\frac{E_{0i}^{(+)} - E_{0r}^{(+)}}{E_{0t}^{(+)}} = \frac{n_2 \cos \theta_t}{n_1 \cos \theta_i}. \quad (9.11)$$

Adding Eqs. (9.8) and (9.11),

$$2 \frac{E_{0i}^{(+)}}{E_{0t}^{(+)}} = 1 + \frac{n_2 \cos \theta_t}{n_1 \cos \theta_i} = \frac{n_1 \cos \theta_i + n_2 \cos \theta_t}{n_1 \cos \theta_i}. \quad (9.12)$$

Now we can define the **field transmission coefficient** t_s as the ratio of the transmitted to incident field:

$$t_s := \frac{E_{0t}^{(+)}}{E_{0i}^{(+)}} = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_i + n_2 \cos \theta_t}. \quad (9.13)$$

We can rewrite Eq. (9.8) as

$$\frac{E_{0r}^{(+)}}{E_{0i}^{(+)}} = \frac{E_{0t}^{(+)}}{E_{0i}^{(+)}} - 1, \quad (9.14)$$

so if we define the **field reflection coefficient** r_s as the ratio of the reflected to incident field amplitude,

$$r_s := \frac{E_{0r}^{(+)}}{E_{0i}^{(+)}} = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t}. \quad (9.15)$$

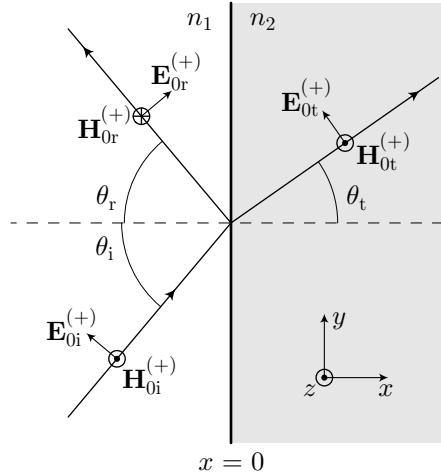
Note that we have assumed a particular input polarization for the light. The notation is that the polarization is **transverse electric** (TE), since the electric field is tangential to the interface. A more common notation in optics is **S-polarization** ("S" is for "senkrecht," which means "perpendicular" in German), because the electric-field vector is perpendicular to the **plane of incidence** (the x - y plane in our coordinates here). Thus, Eqs. (9.15) and (9.13), rewritten here,

$$\begin{aligned} r_s &= \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t} \\ t_s &= \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_i + n_2 \cos \theta_t}. \end{aligned} \quad (9.16)$$

(Fresnel relations, S-polarization)

are the **Fresnel relations for S-polarization**.

We can repeat this derivation for the other polarization, parallel to the plane of incidence, called **transverse magnetic** (TM) or **P-polarization**. The setup is shown here in the following diagram.



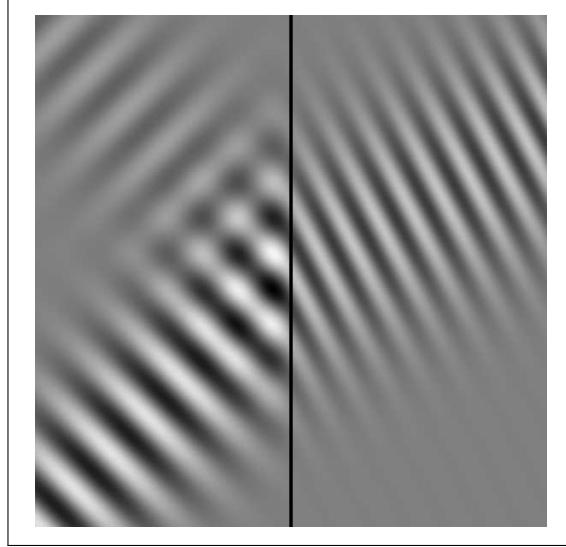
For the vector convention shown, the Fresnel relations turn out to be

$$\begin{aligned} r_p &= \frac{n_1 \cos \theta_t - n_2 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i} \\ t_p &= \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i}. \end{aligned} \quad (9.17)$$

(Fresnel relations, P-polarization)

These are the **Fresnel relations for P-polarization**.

Here is a graphical representation of incident, reflected, and transmitted fields at an air-glass interface that matches our setup here (air on the left, glass on the right). This plot is animated in the electronic version.



9.1.1 Phase Changes and the Brewster Angle

The phase factors of all the waves must be equal, so the relative phases are in some sense fixed. The reflection and transmission coefficients are also real, so there is no phase shift of the components, except possibly for an overall minus sign. Thus, there is the possibility of a π phase shift of the wave after it encounters the boundary.

Transmission. By examination of Eqs. (9.16) and (9.17) it is evident that both $t_s > 0$ and $t_p > 0$. Thus, for either polarization, the *transmitted* wave is in phase with the incident wave.

Reflection: S-polarization. The reflection coefficient is more subtle. For S-polarization, the sign of the reflection coefficient depends on the sign of $(n_1 \cos \theta_i - n_2 \cos \theta_t)$. If $n_1 > n_2$, then from Snell's law, we know that $\theta_t > \theta_i$, and so $\cos \theta_t < \cos \theta_i$. Thus, $r_s > 0$, and the reflected wave is in phase with the incident wave.

On the other hand, if $n_1 < n_2$, then $\theta_t < \theta_i$, and so $\cos \theta_t > \cos \theta_i$. In this case, $r_s < 0$, and the reflected wave has a π phase shift. *Thus, the reflected wave only picks up a π phase shift when the light is incident on a more dense medium* (e.g., from air onto glass).

Reflection: P-polarization. This case is even a bit more complex. From Eqs. (9.17), the sign of r_p depends on the sign of $n_1 \cos \theta_t - n_2 \cos \theta_i$. Suppose that this quantity is positive. Since $n_1/n_2 = \sin \theta_t / \sin \theta_i$, the condition

$$n_1 \cos \theta_t - n_2 \cos \theta_i > 0 \quad (9.18)$$

is equivalent to

$$\sin \theta_t \cos \theta_t - \sin \theta_i \cos \theta_i > 0. \quad (9.19)$$

This, in turn, is equivalent to the condition

$$\sin(\theta_t - \theta_i) \cos(\theta_t + \theta_i) > 0, \quad (9.20)$$

which follows from the general trigonometric identity $\cos(A + B) \sin(A - B) = \sin A \cos A - \sin B \cos B$. Eq. (9.20) is true if the two factors are either both positive or both negative. Let's summarize these two cases as follows:

1. $n_2 < n_1$: both factors are positive when $\theta_t > \theta_i$ and $\theta_t + \theta_i < \pi/2$

2. $n_2 > n_1$: both factors are negative when $\theta_t < \theta_i$ and $\theta_t + \theta_i > \pi/2$

There is a π phase change on reflection if neither set of conditions holds. The borderline case occurs when $\theta_i + \theta_t = \pi/2$, which means that the reflected ray is perpendicular to the refracted ray. Also, this is the case where the numerator in the expression for r_p in Eq. (9.17) vanishes, so the borderline case corresponds to $r_p = 0$, so that *the reflected wave vanishes*. From Snell's law, the this borderline case corresponds to an incidence angle $\theta_{i,B}$ determined by

$$n_1 \sin \theta_{i,B} = n_2 \sin \theta_t = n_2 \sin \left(\frac{\pi}{2} - \theta_{i,B} \right) = n_2 \cos \theta_{i,B}. \quad (9.21)$$

Thus, the incidence angle for the borderline case is

$$\theta_{i,B} = \tan^{-1} \frac{n_2}{n_1}, \quad (9.22)$$

(Brewster's angle)

which is called **Brewster's angle**. Note that the reflection coefficient only vanishes for P-polarization. There is no analogous behavior for S-polarization. Thus, Brewster's angle is also called the **polarizing angle** because light incident at this angle is *polarized on reflection* from a dielectric interface—only S-polarized light reflects, while part of the S-polarization and *all* of the P-polarization transmit.

9.2 Reflectance and Transmittance

Now we must characterize the energy transport due to the fields at the dielectric interface. The energy fluxes at the surface due to the the incident, reflected, and transmitted waves are given in terms of the respective Poynting vectors, $\langle \mathbf{S}_i \rangle \cdot \hat{x}$, $\langle \mathbf{S}_r \rangle \cdot \hat{x}$, and $\langle \mathbf{S}_t \rangle \cdot \hat{x}$. Note that since these are plane waves, the incident and reflected waves overlap. Thus, it seems that we should consider interference cross terms when calculating the Poynting vectors. However, we can treat them separately by realizing that plane waves are an idealization: physical beams have finite width, so we can go far away from the interface until the incident and reflected beams no longer overlap.

To consider the energy balance, we note that (see Problem 4.4) the Poynting vector can be written in terms of the electromagnetic energy density as

$$\langle \mathbf{S} \rangle = c \langle w \rangle \hat{k}. \quad (9.23)$$

Energy balance requires that the energy flux towards the interface is the same as the energy flux away:

$$\langle \mathbf{S}_i \rangle \cdot \hat{x} = \langle \mathbf{S}_t \rangle \cdot \hat{x} + \langle \mathbf{S}_r \rangle \cdot (-\hat{x}). \quad (9.24)$$

Writing this out in terms of the energy densities,

$$\frac{c_0}{n_1} \langle w_i \rangle \cos \theta_i = \frac{c_0}{n_2} \langle w_t \rangle \cos \theta_t - \frac{c_0}{n_1} \langle w_r \rangle \cos \theta_i. \quad (9.25)$$

We can now define the **reflectance** as the ratio of reflected to incident energy fluxes:

$$R := \frac{-\langle \mathbf{S}_r \rangle \cdot \hat{x}}{\langle \mathbf{S}_i \rangle \cdot \hat{x}} = \frac{\frac{c_0}{n_1} \langle w_r \rangle \cos \theta_i}{\frac{c_0}{n_1} \langle w_i \rangle \cos \theta_i} = \frac{\langle w_r \rangle}{\langle w_i \rangle}. \quad (9.26)$$

(reflectance)

Similarly, the **transmittance** as the ratio of transmitted to incident energy fluxes:

$$T = \frac{\langle \mathbf{S}_t \rangle \cdot \hat{x}}{\langle \mathbf{S}_i \rangle \cdot \hat{x}} = \frac{n_1 \langle w_t \rangle \cos \theta_t}{n_2 \langle w_i \rangle \cos \theta_i}. \quad (9.27)$$

(transmittance)

Now we can use the explicit forms for the energy densities in terms of the fields:

$$\langle w_i \rangle = 2n_1^2 \epsilon_0 |E_i^{(+)}|^2; \quad \langle w_r \rangle = 2n_1^2 \epsilon_0 |E_r^{(+)}|^2; \quad \langle w_t \rangle = 2n_2^2 \epsilon_0 |E_t^{(+)}|^2. \quad (9.28)$$

Thus, the reflectance is the relative intensity,

$$R = |r|^2, \quad (9.29) \quad (\text{reflectance in terms of reflection coefficient})$$

but the transmittance is *not*, in general, the obvious expression in terms of the field transmission coefficient ($T \neq |t|^2$). Rather,

$$T = \frac{n_2 \cos \theta_t}{n_1 \cos \theta_i} |t|^2. \quad (\text{transmittance in terms of transmission coefficient}) \quad (9.30)$$

In terms of the reflectance and transmittance, Eq. (9.25) becomes

$$R + T = 1. \quad (9.31) \quad (\text{conservation of energy})$$

Note that the expression (9.30) may seem counterintuitively complicated. In practice, we usually care about reflection at a dielectric interface that is at one end of, say, a slab of glass (such as a glass window used as a beam splitter). In this case, the incident and transmitted energies are measured *outside* the dielectric medium. In particular, they are measured in the *same* surrounding medium (say, air), in which case the transmittance reduces to $|t|^2$, the relative transmitted intensity. Thus, for reflection from the surface of a glass window, the reflected and transmitted intensities (detected outside the window) must add up to the incident intensity.

9.3 Internal Reflection

Recall that Snell's law reads

$$n_1 \sin \theta_i = n_2 \sin \theta_t. \quad (9.32)$$

As in the geometrical case, if $n_1 > n_2$ there is a critical incident angle given by

$$\sin \theta_{i,c} = \frac{n_2}{n_1}, \quad (9.33) \quad (\text{critical angle})$$

which comes from the condition $\theta_t = \pi/2$, the maximum possible refraction angle. Beyond this incident angle, the transmitted angle becomes a complex number. The simple interpretation is that there is no transmitted plane wave, but we will explore this phenomenon in more detail.

The transmitted wave has the form

$$\mathbf{E}_t^{(+)} = \mathbf{E}_{0t}^{(+)} \exp \{ik_t [(\cos \theta_t)x + (\sin \theta_t)y]\} \quad (9.34)$$

for either polarization. Supposing that $\theta_i > \theta_{i,c}$, we have

$$\sin \theta_t = \frac{n_1}{n_2} \sin \theta_i > 1, \quad (9.35)$$

so that

$$\cos \theta_t = \pm \sqrt{1 - \sin^2 \theta_t} = \pm i \sqrt{\left(\frac{n_1}{n_2}\right)^2 \sin^2 \theta_i - 1}, \quad (9.36)$$

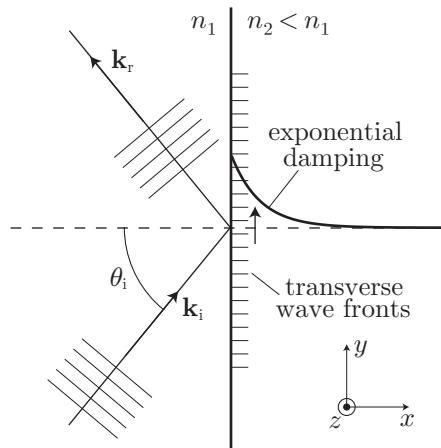
where the final form shows explicitly that the cosine is an imaginary number. Thus, the transmitted wave becomes

$$\mathbf{E}_t^{(+)} = \mathbf{E}_{0t}^{(+)} \exp \left\{ ik_t \left[\left(\pm i \sqrt{\left(\frac{n_1}{n_2}\right)^2 \sin^2 \theta_i - 1} \right) x + (\sin \theta_t) y \right] \right\}. \quad (9.37)$$

For a bounded solution, we must choose the $+i$ branch of the square root in Eq. (9.36). Then we see that the transmitted wave,

$$\mathbf{E}_t^{(+)} = \mathbf{E}_{0t}^{(+)} \exp \left[-k_t \sqrt{\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1} x \right] \exp \left[i k_t \frac{n_1}{n_2} (\sin \theta_i) y \right], \quad (\text{evanescent wave}) \quad (9.38)$$

which is an exponentially damped harmonic wave: the left exponential factor represents exponential damping along the x -direction, while the right exponential factor has the form of a traveling plane wave along the y -direction. This damped field is called the **evanescent field**.



From the damping factor in Eq. (9.38), the wave decays as

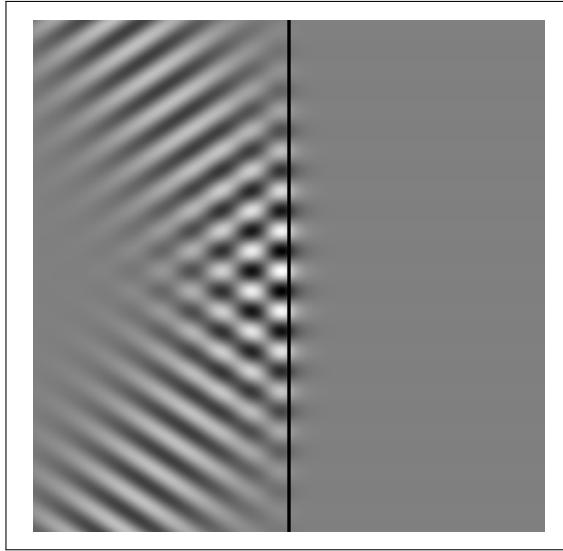
$$\mathbf{E}_t^{(+)} \sim e^{-x/\delta}, \quad (9.39)$$

where the **skin depth** δ , given by

$$\delta = \left[k_t \sqrt{\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1} \right]^{-1}, \quad (\text{skin depth}) \quad (9.40)$$

is the characteristic distance over which the evanescent wave damps away. Since the (second) propagating-wave factor in Eq. (9.38) only involves y , the wave fronts of the evanescent wave propagate *along* the interface, as shown. This is one way to see that the evanescent wave carries no energy away from the interface. The other, which we will see shortly, is that $R = 1$. Thus, this situation ($\theta > \theta_{i,c}$) is called **total internal reflection**.

Here is a visualization of internal reflection at a glass-air interface. The incoming wave is incident at an angle just slightly past the critical angle, so that the skin depth is relatively large. You can clearly see the propagation of the wave fronts parallel to the interface.



Actually, what I said above about the evanescent wave isn't quite true. It *can* carry away some energy, provided that the evanescent wave is interrupted. For example, for internal reflection at a glass-air interface, the wave is damped in the air. But if *another* glass interface is placed very close to the first, so that the air is present only in a thin ($\sim\lambda$) gap, the damped wave will begin to propagate again in the second glass region. This phenomenon is called **frustrated total internal reflection** and is the analog of quantum tunneling. This effect can be used, for example, to make a variable-ratio beam splitter by adjusting the size of a small air gap between two prisms¹.

9.3.1 Phase Shifts

The reflection coefficients from Eqs. (9.16) and (9.17) are

$$\begin{aligned} r_s &= \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t} \\ r_p &= \frac{n_1 \cos \theta_t - n_2 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i}. \end{aligned} \quad (9.41)$$

From Eq. (9.36), we have (upon choosing the positive branch of the square root)

$$\cos \theta_t = i \sqrt{\left(\frac{n_1}{n_2}\right)^2 \sin^2 \theta_i - 1}, \quad (9.42)$$

so that we can eliminate θ_t in the reflection coefficients:

$$\begin{aligned} r_s &= \frac{\cos \theta_i - i \sqrt{\sin^2 \theta_i - (n_2/n_1)^2}}{\cos \theta_i + i \sqrt{\sin^2 \theta_i - (n_2/n_1)^2}} \\ r_p &= - \frac{(n_2/n_1)^2 \cos \theta_i - i \sqrt{\sin^2 \theta_i - (n_2/n_1)^2}}{(n_2/n_1)^2 \cos \theta_i + i \sqrt{\sin^2 \theta_i - (n_2/n_1)^2}}. \end{aligned} \quad (9.43)$$

Notice that both of these reflection coefficients are of the form

$$\pm \frac{\alpha - i\beta}{\alpha + i\beta}, \quad (9.44)$$

¹D. Bertani, M. Cetica, and R. Polloni, "A simple variable-ratio beam splitter for holography," *Journal of Physics E: Scientific Instruments* **16**, 602 (1983).

so that the reflectance for both polarizations is unity:

$$R_s = R_p = 1. \quad (9.45)$$

(total internal reflection)

However, there is a phase shift on reflection, since the reflection coefficients themselves are not unity in general. The phase angle in each case is $\exp(-2\phi)$, (up to a minus sign) where $\tan \phi = \beta/\alpha$, so we can write

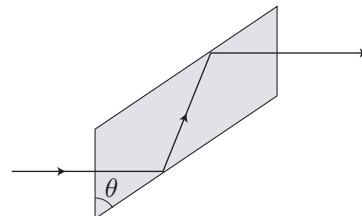
$$r_s = \exp \left[-i2 \tan^{-1} \left(\frac{\sqrt{\sin^2 \theta_i - (n_2/n_1)^2}}{\cos \theta_i} \right) \right]$$

$$r_p = \exp \left[i\pi - i2 \tan^{-1} \left(\frac{\sqrt{\sin^2 \theta_i - (n_2/n_1)^2}}{(n_2/n_1)^2 \cos \theta_i} \right) \right]$$

(phase shifts for internal reflection) (9.46)

for the explicit phase shifts of the two polarizations. Apart from an overall minus sign (π phase shift), there is a refractive index ratio in the expression for r_p that is absent in the expression for r_s . Hence, if both polarization components are present, there is a *relative* phase shift between them.

This phase shift can be very useful. In Section 8.7.1, we discussed how a birefringent material can make a wave retarder (wave plate). But the phase shifts due to internal reflection obviously also act as wave retarders. One optical element, the **Fresnel rhomb**, uses two internal reflections to produce a $\pi/2$ ($\lambda/4$) phase shift.



For glass with $n = 1.52$, the correct incidence angle turns out to be $\theta = 55.5^\circ$. The rhomb has the sometimes inconvenient property of translating the beam, but unlike typical wave plates, the retardation of a rhomb is relatively insensitive to the optical wavelength used.

9.4 Air-Glass Interface: Sample Numbers

To get a feel for the theory presented here, let's work out specific numbers for an interface between air ($n_1 = 1$) and glass ($n_2 = 1.52$). At normal incidence, $\cos \theta_i = \cos \theta_t = 1$, so

$$r = r_s = r_p = \frac{n_1 - n_2}{n_1 + n_2} = -0.21 \quad (9.47)$$

and

$$R = |r|^2 = 4.3\%. \quad (9.48)$$

Thus, a glass window attenuates light due to reflection near normal incidence by about twice this amount, around 8%.

At 45° incidence, the reflectances for the two polarizations are quite different. First, the refraction angle is

$$\theta_t = \sin^{-1} \left(\frac{n_1}{n_2} \sin \theta_i \right) = 27.7^\circ. \quad (9.49)$$

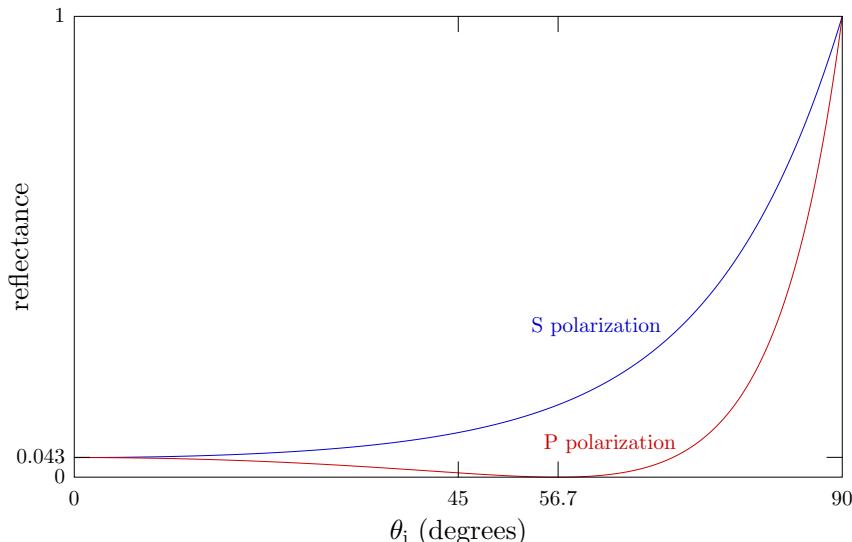
Thus, the reflectances are

$$\begin{aligned} R_S &= |r_s|^2 = \left| \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t} \right|^2 = 9.7\% \\ R_P &= |r_p|^2 = \left| \frac{n_1 \cos \theta_t - n_2 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i} \right|^2 = 0.94\%, \end{aligned} \quad (9.50)$$

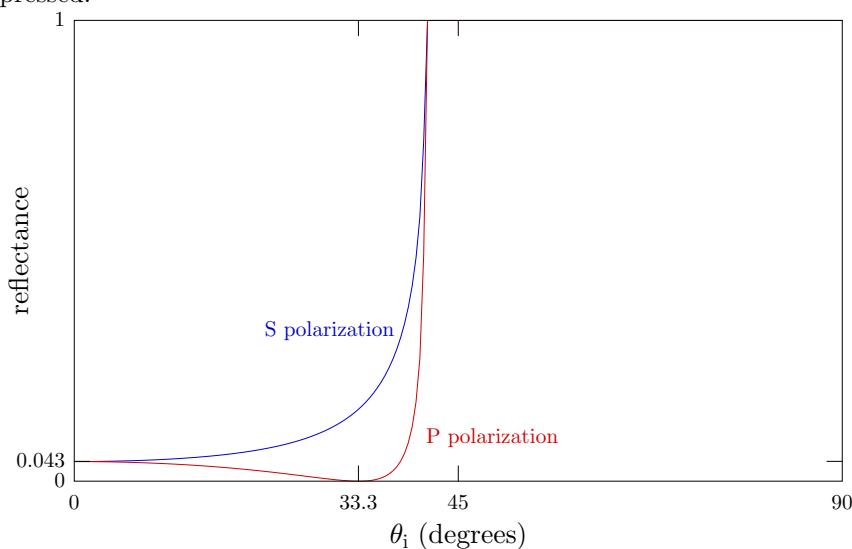
or about 10% and 1%, respectively. (These numbers are useful to remember in the laboratory for making simple 10% beam splitters.) The reflectance for P-polarization is much smaller because this incidence angle is fairly close to the Brewster angle of

$$\theta_{i,B} = \tan^{-1} \frac{n_2}{n_1} = 56.7^\circ. \quad (9.51)$$

The reflectances for the air-glass interface for both polarizations are shown here as a function of the incident angle.



On the other hand, for a glass-air interface, the reflectance curves look the same except that they are horizontally compressed.



The reflectances are unity beyond the critical incident angle

$$\theta_{i,c} = \sin^{-1} \frac{n_{\text{air}}}{n_{\text{glass}}} = 41.1^\circ, \quad (9.52)$$

as is evident from the plot.

9.5 Reflection at a Dielectric-Conductor Interface

One useful generalization of reflection at a dielectric interface is to treat the case where one of the materials is a conductor. This covers the important case of metallic reflectors, which are some of the most common in everyday experience as well as the laboratory. We will not deal explicitly with the case of reflection at the interface of two conductors, as this is more subtle. But to deal with reflection from a *single* conductor, we will need to treat wave propagation in a conductive medium.

9.5.1 Propagation in a Conducting Medium

In Chapter 4, we considered only Maxwell's equations in dielectric media without sources. *With* field sources, the Maxwell equations become

$$\begin{aligned} \nabla \cdot \mathbf{D} &= \rho \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}, \end{aligned} \tag{9.53}$$

(Maxwell equations coupled to sources)

where ρ is the charge density and \mathbf{J} is the current density.

For an ohmic material, the current density is proportional to the local electric field,

$$\mathbf{J} = \sigma \mathbf{E}, \tag{9.54}$$

(Ohm's Law)

where σ is the **conductivity**. We can think of the conductivity as a real constant, but in reality at optical frequencies it is best modeled by a complex number, due to the lag of the response of the conduction electrons to the electric field (see Section 17.2.2 for a model of the conductivity).

Charge and current are related by the **continuity equation**

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{J}, \tag{9.55}$$

(continuity equation)

which is essentially equivalent to the statement that charge is conserved. If the current density at some point is increasing, then there must be a convergence of current density to transport charge to that point. We can use the relation $\mathbf{D} = \epsilon \mathbf{E}$ and Eq. (9.54) to write

$$\frac{\partial \rho}{\partial t} = -\sigma(\nabla \cdot \mathbf{E}) = -\frac{\sigma}{\epsilon} \rho. \tag{9.56}$$

This means that $\rho \rightarrow 0$ exponentially with a time constant ϵ/σ . For a good conductor (large σ), the damping time is very short, so to good approximation we can set $\rho = 0$. On longer time scales of optical interest, the charge density relaxes to nothing, and this is why we can justify ignoring it. This is a simple example of an *adiabatic approximation*. For our purposes, this means that the first Maxwell equation in Eq. (9.53) becomes

$$\nabla \cdot \mathbf{D} = 0, \tag{9.57}$$

(first Maxwell equation for conductor)

so that the only source is the current density, which we can write in terms of \mathbf{E} .

We can now repeat the derivation of the wave equation from Chapter 4, with the result

$$\nabla^2 \mathbf{E} - \mu_0 \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} - \mu_0 \sigma \frac{\partial \mathbf{E}}{\partial t} = 0, \quad (9.58)$$

(wave equation for conductor)

with the same result for \mathbf{H} . Note that we are ignoring magnetic properties of the conductor by assuming $\mu = \mu_0$. We can again try the plane-wave *ansatz*

$$\mathbf{E}^{(+)}(\mathbf{r}, t) = \mathbf{E}_0^{(+)} e^{i(kz - \omega t)}. \quad (9.59)$$

Substitution shows that the plane wave is a solution provided that

$$k^2 = \mu_0 \epsilon \omega^2 + i \mu_0 \sigma \omega = \omega^2 \mu_0 \left(\epsilon + i \frac{\sigma}{\omega} \right). \quad (9.60)$$

Note that the frequency dependence in this equation is deceptively simple, as in general σ and ϵ are themselves functions of frequency. If we define a **complex permittivity**

$$\tilde{\epsilon} = \epsilon + i \frac{\sigma}{\omega}, \quad (9.61)$$

(complex permittivity)

then the dispersion relation (9.60) takes the same form as for an ordinary dielectric,

$$k^2 = \omega^2 \mu_0 \tilde{\epsilon}, \quad (9.62)$$

(dispersion relation for conductor)

with the replacement $\epsilon \rightarrow \tilde{\epsilon}$.

Similarly, since the refractive index for a dielectric is the square root of the dielectric constant,

$$n^2 = \frac{\epsilon}{\epsilon_0}, \quad (9.63)$$

we can define the **complex refractive index** \tilde{n} as the square root of the *complex* dielectric constant:

$$\tilde{n}^2 = \frac{\tilde{\epsilon}}{\epsilon_0}. \quad (9.64)$$

(complex refractive index)

We can also write the complex refractive index itself in terms of real and imaginary parts as

$$\tilde{n} = n(1 + i\kappa), \quad (9.65)$$

(complex refractive index)

where κ is the **attenuation index** or **extinction coefficient**. This is the central point of electromagnetic wave propagation in conductors: *everything is the same as for an ordinary dielectric, so long as you use a complex refractive index*.

Let's compute explicitly the real and imaginary parts of the refractive index. First, computing \tilde{n}^2 from both Eqs. (9.64) and (9.65),

$$\tilde{n}^2 = n^2(1 - \kappa^2) + i2n^2\kappa = \frac{\epsilon}{\epsilon_0} + i\frac{\sigma}{\omega\epsilon_0}. \quad (9.66)$$

Matching the real and imaginary parts, we have the two equations

$$n^2(1 - \kappa^2) = \frac{\epsilon}{\epsilon_0}, \quad 2n^2\kappa = \frac{\sigma}{\omega\epsilon_0}. \quad (9.67)$$

These equations have the solutions

$$\begin{aligned} n^2 &= (\operatorname{Re}[\tilde{n}])^2 = \frac{\epsilon}{2\epsilon_0} \left[\sqrt{1 + \left(\frac{\sigma}{\epsilon\omega} \right)^2} + 1 \right] \\ n^2\kappa^2 &= (\operatorname{Im}[\tilde{n}])^2 = \frac{\epsilon}{2\epsilon_0} \left[\sqrt{1 + \left(\frac{\sigma}{\epsilon\omega} \right)^2} - 1 \right]. \end{aligned} \quad (9.68)$$

(real/imaginary parts of complex refractive index)

Similarly, we can compute the wave vector in terms of the refractive index as

$$k = \omega \sqrt{\mu_0 \tilde{\epsilon}} = \omega \sqrt{\mu_0 \epsilon_0} \tilde{n} = k_+ + ik_-, \quad (9.69)$$

where

$$k_{\pm}^2 = \frac{\omega^2 \mu_0 \epsilon}{2} \left[\sqrt{1 + \left(\frac{\sigma}{\epsilon \omega} \right)^2} \pm 1 \right]. \quad (\text{real/imaginary parts of wave number}) \quad (9.70)$$

Thus, we can rewrite the plane-wave solution as

$$\mathbf{E}^{(+)}(\mathbf{r}) = \mathbf{E}_0^{(+)} e^{ikz} = \mathbf{E}_0^{(+)} e^{-k_- z} e^{ik_+ z}. \quad (9.71) \quad (\text{damped plane wave})$$

As we saw for internal reflection, the plane wave breaks up into propagating and damping factors—the imaginary part of the refractive index leads to damping of the wave, while the real part represents the propagating aspect of the wave. We can again define a **skin depth** δ for the damped wave by

$$e^{-k_- z} = e^{-z/\delta}, \quad (9.72)$$

so that

$$\delta = \frac{1}{k_-} = \frac{1}{\omega \sqrt{\mu_0 \epsilon_0} n \kappa} = \frac{\lambda_0}{2\pi n \kappa} = \frac{\lambda_0}{2\pi \text{Im}[\tilde{n}]}. \quad (9.73) \quad (\text{skin depth in conductor})$$

As before, this is the length scale over which the wave damps exponentially away. Typically, for a good conductor, $n\kappa$ is large, so that $\delta \ll \lambda_0$. This is called the **skin effect**: for good conductors, an electromagnetic wave penetrates the surface by only a small fraction of the vacuum wavelength.

9.5.2 Inductive Heating

Since the transmitted wave gets damped, something must be taking up that dissipated energy. Of course, this goes into the metal as heat. While being undesirable from the point of view of making mirrors for high-power lasers, this effect can be useful in the machine shop or foundry for heat-treating or melting metals. An **induction heater** drives a high-voltage, radio-frequency signal through a coil that surrounds the conductive object to be heated. The electromagnetic wave generated by the coil couples into the conductive object, and generates heat as it is absorbed. Of course, the equivalent picture is that the radiation induces currents in the object, causing Joule heating—from our analysis above, we saw that it is precisely these induced currents that damp the electromagnetic wave after it penetrates the surface.

The following photographic sequence shows a quarter being heated and then melted by an inductive heater. Roughly speaking, you can think of the coil as being a solenoid, which ideally produces a uniform, oscillating magnetic field along the solenoid axis. The corresponding electric field is roughly a cylindrical wave collapsing onto the quarter, with the wave polarization being in the plane of the quarter. The wavelength for the 200 kHz radiation is quite large, around 1.5 km. However, the copper inside the quarter is an extremely good conductor; assuming a conductivity of $60 \times 10^6 \text{ (m} \cdot \Omega)^{-1}$, the skin depth turns out to be only 0.15 mm. Thus, due to the skin effect, the radiation is absorbed in a thin shell on the outside of the quarter. You can see this in the photo series, because the outside of the quarter is visibly much hotter than the middle, since

the middle is only being heated by thermal conduction.



Here is a movie of the heating process, where you can see the heat flowing in from the outside of the quarter to the inside.



9.5.3 Fresnel Relations

It is worth reiterating: with the conductor present, everything is the same as in the dielectric case, except that the refractive index is now complex. In particular, the Fresnel reflection coefficients still hold with the complex refractive index:

$$\begin{aligned} r_s &= \frac{n_1 \cos \theta_i - \tilde{n}_2 \cos \theta_t}{n_1 \cos \theta_i + \tilde{n}_2 \cos \theta_t} \\ r_p &= \frac{n_1 \cos \theta_t - \tilde{n}_2 \cos \theta_i}{n_1 \cos \theta_t + \tilde{n}_2 \cos \theta_i}. \end{aligned} \quad (9.74)$$

(Fresnel relations with complex index)

Here, n_1 is the refractive index of the dielectric (from which the wave is incident), and \tilde{n}_2 is the complex refractive index of the conductor. Note that this is even a bit more complicated than before, though, since from Snell's law, θ_t is a complex number too! We won't go into much detail regarding the wave in the medium beyond what we've already done, since we are usually only interested in this wave if it is then transmitted back into free space where it can be detected. This situation is better treated with the matrix formalism for multilayer films that we will get to shortly.

Let's plug in some numbers for the simple example case of a polished silver mirror at normal incidence. The reflectance is

$$R = \left| \frac{n_1 - \tilde{n}_2}{n_1 + \tilde{n}_2} \right|^2. \quad (9.75)$$

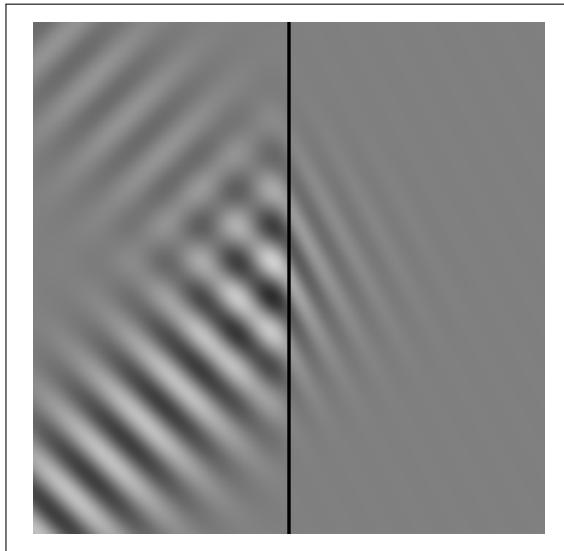
For silver at 780 nm,² $\tilde{n} = 0.27 + 4.47i$, which give $R = 95.0\%$. Thus silver is a good reflector, but some of the incident energy is lost in the metal. This is because the wave penetrates a short distance into the metal. Putting in numbers, we find that the skin depth is

$$\delta = \frac{\lambda_0}{2\pi n \kappa} = \frac{\lambda_0}{28.1} = 27.8 \text{ nm} \quad (9.76)$$

²J. H. Weaver and H. P. R. Frederikse, "Optical Properties of Selected Elements," in *CRC Handbook of Chemistry and Physics*, 82nd ed., David R. Lide, Ed. (CRC Press, Boca Raton, 2001), p. 12-133.

at 780 nm.

Here is a visualization of wave reflection from a dielectric-conductor interface. The specific indices used are vacuum ($n_1 = 1$) on the left and slightly conductive glass ($\tilde{n}_2 = 1.52 + 0.152i$) on the right. Note that the wave damping is similar to the case of internal reflection, but now the transmitted wave propagates away from the interface rather than along it.



9.6 Exercises

Problem 9.1

We derived the Fresnel relations for S-polarized light giving the reflection coefficient

$$r_S = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t} \quad (9.77)$$

and transmission coefficient

$$t_S = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_i + n_2 \cos \theta_t} \quad (9.78)$$

of the electric field at a dielectric interface. Use the setup from the text to derive the Fresnel relations for the reflection coefficient,

$$r_P = \frac{n_1 \cos \theta_t - n_2 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i} \quad (9.79)$$

and transmission coefficient

$$t_P = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i} \quad (9.80)$$

for P-polarized light.

Problem 9.2

When we worked out the Fresnel relations for the reflection and coefficients at a planar, dielectric interface, we ignored two boundary conditions, namely that the components of \mathbf{D} and \mathbf{B} *normal* to the interface are continuous across the boundary. Write out both of these boundary conditions in terms of the Fresnel coefficients r and t for both polarizations.

Problem 9.3

One way of making a cheap but good polarizer (except for wavefront distortion) is to use a stack of microscope slides. Assuming an index of refraction $n = 1.52$ and that the slides are all oriented at Brewster's angle with respect to a randomly polarized laser beam, how many slides does it take to attenuate one polarization by a factor of 10^{-4} in intensity compared to the other? Keep in mind that *both* sides of each slide are at Brewster's angle. Ignore multiple reflections in your calculation.

Problem 9.4

Diode lasers in the near infrared (say, around 780 nm) are made from GaAlAs crystals. For low-power diodes (in the 5-15 mW range), the crystal ends are cleaved to form flat surfaces that act as the two flat reflectors of the laser cavity. Model a typical diode laser as a crystal ($n = 3.5$) of length 300 μm , surrounded by air ($n = 1$).

(a) For this resonator, calculate the round-trip time, the free spectral range, the finesse, the photon lifetime, and the frequency width (FWHM) of the modes. (Note that for higher power diodes, the output facet is typically *antireflection* coated to couple out more of the light in the cavity.)

(b) Assuming that the diode laser frequency is that of a single longitudinal mode, calculate the change in laser frequency per degree Celsius for a temperature change. Assume a nominal operating (vacuum) wavelength of 780 nm and a coefficient of thermal expansion ($\Delta L/L$) for GaAlAs of $7 \times 10^{-6}/^\circ\text{C}$. (In practice this tuning rate is only valid for small temperature changes; for larger changes, the change in the band gap is more important, causing much larger average tuning rates as the laser "hops" between longitudinal modes.)

Problem 9.5

Plot the intensity reflection coefficients as a function of incident angle for light incident from air ($n = 1$) onto crown glass ($n = 1.52$) for both S- and P-polarized light. Repeat for light incident from within the glass.

Problem 9.6

For total internal reflection from a glass–air interface ($n_1 = 1.52$ for glass, $n_2 = 1$ for air), plot the skin depth normalized to the vacuum wavelength (δ/λ_0) as a function of incident angle.

Problem 9.7

Derive the wave equation for the magnetic field \mathbf{H} in a homogeneous medium of permittivity ϵ and conductivity σ .

Problem 9.8

To describe the propagation of electromagnetic waves in a conductor, we derived a wave equation for \mathbf{E} , coupled to the current density induced by the field in the conductor. An alternate approach is to consider the wave equation to be coupled to the *polarization* \mathbf{P} of the medium.

As the first step in this method, recall that the electric displacement, electric field, and polarization density are related by

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}, \quad (9.81)$$

and ignoring magnetic effects, we can assume

$$\mathbf{B} = \mu_0 \mathbf{H}. \quad (9.82)$$

Then use Maxwell's equations in a source-free, dielectric medium,

$$\begin{aligned} \nabla \cdot \mathbf{D} &= 0 \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\partial_t \mathbf{B} \\ \nabla \times \mathbf{H} &= \partial_t \mathbf{D}, \end{aligned} \quad (9.83)$$

to derive a wave equation driven by \mathbf{P} . You may assume the polarization field to be *transverse*, i.e., $\nabla \cdot \mathbf{P} = 0$.

Problem 9.9

Compute the reflectance and skin depth for light at normal incidence from vacuum at wavelength 780 nm for polished surfaces of gold ($\tilde{n} = 0.08 + 4.60i$) and copper ($\tilde{n} = 0.24 + 4.80i$).

Problem 9.10

- (a) Show that in the limit of low frequency, the skin depth for a good (nonmagnetic) conductor can be written

$$\delta \approx \sqrt{\frac{2}{\mu_0 \omega \sigma}}. \quad (9.84)$$

- (b) Evaluate the skin depth at 200 kHz for titanium ($\sigma = 2 \times 10^6 \text{ (m} \cdot \Omega)^{-1}$). Recall that $\mu_0 = 4\pi \times 10^{-7} \text{ N/A}^2$.

- (c) Suppose you have a disc of titanium the size and shape of a quarter. The disc is in the center of a solenoidal coil, which is driven at 200 kHz, and it lies in the plane orthogonal to the coil axis. Assume that the solenoid length is much larger than all other length scales in this setup. What is the (qualitative) distribution of power absorption in the disc? *Explain*.

Problem 9.11

It is conventional to represent the mode functions of a *compact* region in space by **normalized mode functions** $\mathbf{f}_{\mathbf{k},\zeta}(\mathbf{r})$, where \mathbf{k} is the wave vector (which takes on discrete values), and ζ is an index marking the two possible polarizations. These mode functions are normalized in the sense that

$$\int_V d^3r |\mathbf{f}_{\mathbf{k},\zeta}(\mathbf{r})|^2 = 1, \quad (9.85)$$

where V is the volume over which the mode exists (i.e., the cavity size). More generally, the orthonormality relation is

$$\int_V d^3r \mathbf{f}_{\mathbf{k},\zeta}(\mathbf{r}) \cdot \mathbf{f}_{\mathbf{k}',\zeta'}^*(\mathbf{r}) = \delta_{\mathbf{k},\mathbf{k}'}^3 \delta_{\zeta,\zeta'}. \quad (9.86)$$

However, for mode functions defined over *all* space, the integration volume is unbounded and \mathbf{k} takes on any value in \mathbb{R}^3 , and this orthonormality relation generalizes to

$$\int d^3r \mathbf{f}_{\mathbf{k},\zeta}(\mathbf{r}) \cdot \mathbf{f}_{\mathbf{k}',\zeta'}^*(\mathbf{r}) = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{\zeta,\zeta'}, \quad (9.87)$$

assuming

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(k) e^{ikx} dk, \quad \tilde{f}(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} dx, \quad (9.88)$$

as the Fourier-transform convention in each dimension.

In the presence of a vacuum–dielectric interface, the mode function for TE polarization can be written

$$\mathbf{f}_{\mathbf{k},\text{TE}} = \hat{\varepsilon}_{\mathbf{k},\text{TE}} [e^{i\mathbf{k}\cdot\mathbf{r}} \Theta(z) + r_{\text{TE}} e^{i\mathbf{k}_r \cdot \mathbf{r}} \Theta(z) + t_{\text{TE}} e^{i\mathbf{k}_t \cdot \mathbf{r}} \Theta(-z)], \quad (9.89)$$

where the dielectric (of permittivity ϵ , and index n) occupies the half-space $z \leq 0$, r_{TE} and t_{TE} are the Fresnel reflection and transmission coefficients for TE (S) polarization, and $\Theta(z)$ is the Heaviside step function. Show that this mode function satisfies the orthonormality relation

$$\int d^3r \frac{\epsilon(\mathbf{r})}{\epsilon_0} \mathbf{f}_{\mathbf{k},\text{TE}}^*(\mathbf{r}) \cdot \mathbf{f}_{\mathbf{k}',\text{TE}}(\mathbf{r}) = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'), \quad (9.90)$$

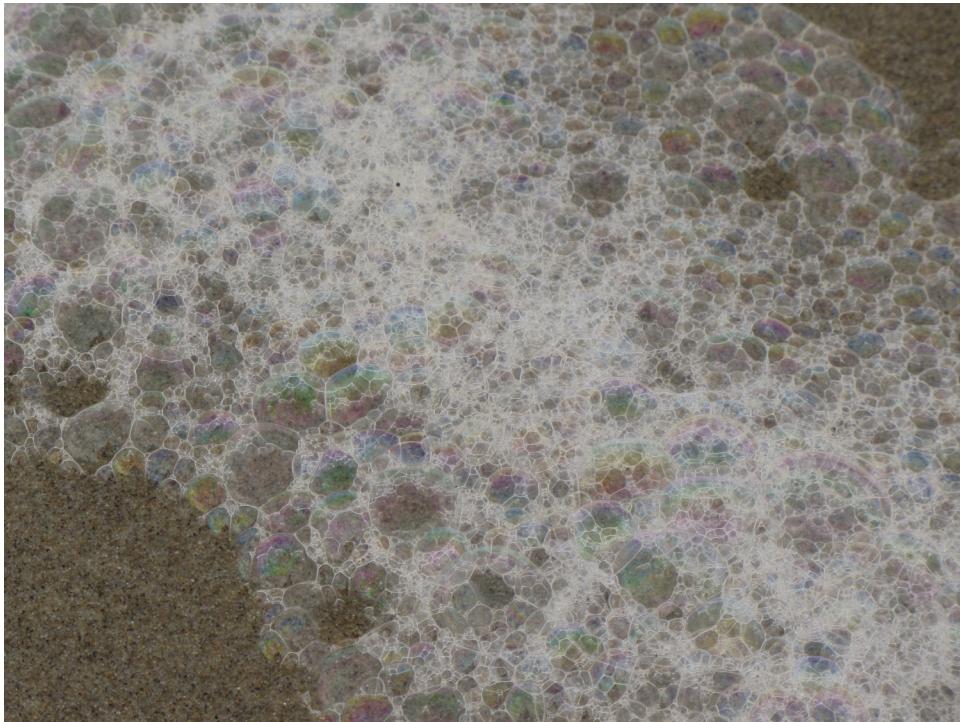
which reduces to the free-space relation as $\epsilon(\mathbf{r}) \rightarrow \epsilon_0$. For simplicity, suppose that $\mathbf{f}_{\mathbf{k},\text{TE}}(\mathbf{r})$ and $\mathbf{f}_{\mathbf{k}',\text{TE}}(\mathbf{r})$ are both incident from the vacuum side.

Note: there is a bit of a trick to getting the delta function to come out right on the transmitting side. You might want to refer to the article that first proved this for help: C. K. Carniglia and L. Mandel, “Quantization of Evanescent Electromagnetic Waves,” *Physical Review D* **3**, 280 (1971) (doi: 10.1103/PhysRevD.3.280).

Chapter 10

Thin Films

Now with reflections from single interfaces under our belt, we can tackle multiple interfaces in the form of thin films. The most obvious everyday examples of thin optical films are the swirling colors in films of oil on puddles of water, or similar colors in bubbles of soap or sea foam, as shown in the photo below.

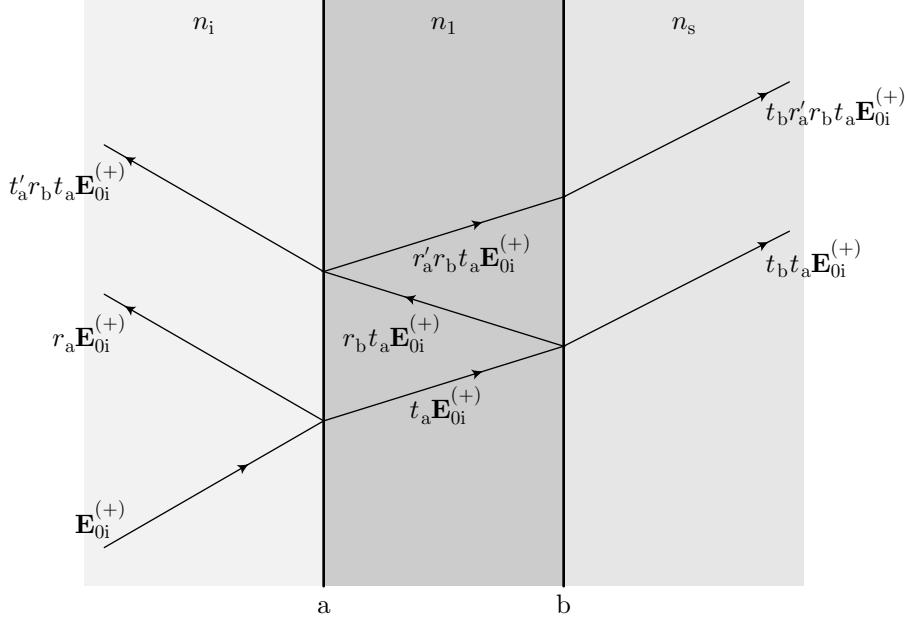


However, optical thin films have widespread applications in modifying the reflection properties of surfaces, including antireflection coatings (e.g., of eyeglasses and camera lenses), high-reflection coatings for laser optics, wavelength filters for spectroscopy and color separation in projectors and photography, polarizers, and so on.

We will start out with a simple treatment of a single-layer film in the style of a Fabry–Perot cavity, and then rederive the result in a much more powerful matrix formalism. We will be implicitly assuming dielectric thin films, but of course we can also handle conductive films simply by using complex refractive indices.

10.1 Reflection-Summation Model

We will now consider the multiple reflections due to an input wave $\mathbf{E}_{0i}^{(+)}$ at a single thin film. We are implicitly considering only a single input polarization (S or P), but the analysis works for either one separately.

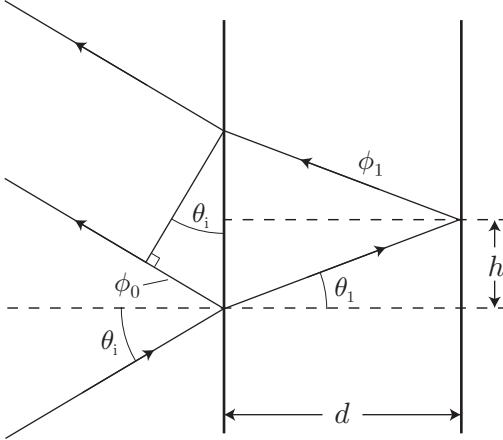


The convention is that r_a , t_a , r_b , and t_b are the reflection and transmission coefficients at interfaces a and b *for left-to-right incidence*. For right-to-left incidence, the coefficients are r'_a , t'_a , r'_b , and t'_b . The interfaces a and b divide the media with refractive indices n_i (for incident medium), n_1 (the coating), and n_s (the substrate or transmission medium); we will use the same subscripts to denote waves inside these media (e.g., \mathbf{k}_1 is the wave vector in medium of index n_1).

Then the total reflected field is the sum of all the individual reflected waves. After the first term, a pattern emerges because each successive reflection is the same as the previous reflection but with one more round trip inside the thin film.

$$\mathbf{E}_{0r}^{(+)} = r_a \mathbf{E}_{0i}^{(+)} + t'_a r_b t_a \mathbf{E}_{0i}^{(+)} e^{i\phi} + t'_a r_b r'_a r_b t_a \mathbf{E}_{0i}^{(+)} e^{i2\phi} + \dots \quad (10.1)$$

Here, ϕ is the propagation phase accumulated between successive reflections *as measured along the same wave front*. This is a somewhat subtle point that requires a careful geometric construction.



As shown in the diagram, the correct form for ϕ follows from the following argument. The phase difference between successive reflections is the phase of the ray after one round trip inside the thin film as it exits

(call this phase ϕ_1), but we need to *subtract* the propagation of the directly reflected ray to the same perpendicular position along the ray (call this second phase ϕ_0). That's how displaced plane waves interfere: the relative phase of the parts that align along the same wave front (normal to the ray) are what determine the interference. Thus,

$$\phi = \phi_1 - \phi_0, \quad (10.2)$$

where

$$\phi_1 = 2\mathbf{k}_1 \cdot \mathbf{r}_1 = 2k_1 \sin \theta_1 h + 2k_1 \cos \theta_1 d, \quad (10.3)$$

and

$$\phi_0 = \mathbf{k}_0 \cdot \mathbf{r}_0 = 2k_i \sin \theta_i h. \quad (10.4)$$

Thus,

$$\phi = \phi_1 - \phi_0 = 2h(k_1 \sin \theta_1 - k_i \sin \theta_i) + 2k_1 d \cos \theta_1. \quad (10.5)$$

The first term vanishes by Snell's law, so we are left with

$$\phi = 2k_1 d \cos \theta_1. \quad (10.6) \quad (\text{round-trip phase in thin film})$$

Now we can write out the explicit sum for the reflected wave by completing the pattern from Eq. (10.1):

$$\mathbf{E}_{0r}^{(+)} = \mathbf{E}_{0i}^{(+)} \left[r_a + t'_a r_b t_a e^{i\phi} \sum_{n=0}^{\infty} (r_b r'_a e^{i\phi})^n \right]. \quad (10.7)$$

We can evaluate this using the formula

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad (10.8)$$

which is valid provided $|x| < 1$, with the result

$$r_{\text{film}} := \frac{\mathbf{E}_{0r}^{(+)}}{\mathbf{E}_{0i}^{(+)}} = r_a + \frac{t'_a r_b t_a e^{i\phi}}{1 - r_b r'_a e^{i\phi}}. \quad (10.9)$$

Using the Stokes relations $t'_a t_a = 1 - r_a^2$ and $r_a = -r'_a$, we can eliminate the transmission coefficients, with the result

$$r_{\text{film}} = r_a + \frac{(1 - r_a^2) r_b e^{i\phi}}{1 + r_a r_b e^{i\phi}}. \quad (10.10)$$

Rewriting all this as a single fraction, we finally come to

$$r_{\text{film}} = \frac{r_a + r_b e^{i\phi}}{1 + r_a r_b e^{i\phi}}, \quad (10.11) \quad (\text{thin film reflection coefficient})$$

which is the main result of this section.

10.1.1 Example: Single Glass Plate as a Fabry–Perot Etalon

To get a better feeling for the meaning behind Eq. (10.11), let's look at the special case of $n_i = n_s$. For example, this could be a glass plate surrounded by air. In this case, $r_a = -r_b$, so that

$$r_{\text{film}} = \frac{r_a (1 - e^{i\phi})}{1 - r_a^2 e^{i\phi}}, \quad (10.12)$$

and the reflectance becomes

$$R_{\text{film}} = |r_{\text{film}}|^2 = \frac{2r_a^2 (1 - \cos \phi)}{1 + r_a^4 - 2r_a^2 \cos \phi}, \quad (10.13) \quad (\text{etalon reflectance})$$

assuming that r_a is real, as for an air–glass interface. Notice that R_{film} drops to zero when $\cos \phi = 1$. This happens when $\phi = 2\pi q$ for some integer q , or

$$k_1 \cos \theta_1 = \frac{\pi q}{d}. \quad (10.14)$$

Comparing this to the resonance condition for a planar-mirror Fabry-Perot cavity [Eq. (7.2)],

$$k_q = \frac{\pi q}{d}, \quad (10.15)$$

we see that we have just rederived the Fabry-Perot resonance condition, but now generalized for non-normal incidence. As we mentioned in Chapter 7, a glass window can act as a planar Fabry-Perot etalon. Even though the window thickness is nominally fixed (neglecting thermal expansion and mechanical stress), we can tune the etalon by changing its orientation with respect to the incident wave.

The transmission of the plate is then

$$\begin{aligned} T_{\text{film}} &= 1 - R_{\text{film}} \\ &= \frac{(1 - r_a^2)^2}{1 + r_a^4 - 2r_a^2 \cos \phi} \\ &= \frac{(1 - r_a^2)^2}{(1 - r_a^2)^2 + 4r_a^2 \sin^2(\phi/2)} \\ &= \frac{1}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2(\phi/2)}, \end{aligned} \quad (10.16)$$

where

$$\mathcal{F} = \frac{\pi r_a}{1 - r_a^2} \quad (10.17) \quad (\text{etalon finesse})$$

is the finesse of the etalon. The transmission angle is, by two applications of Snell's law, just the same as the incident angle.

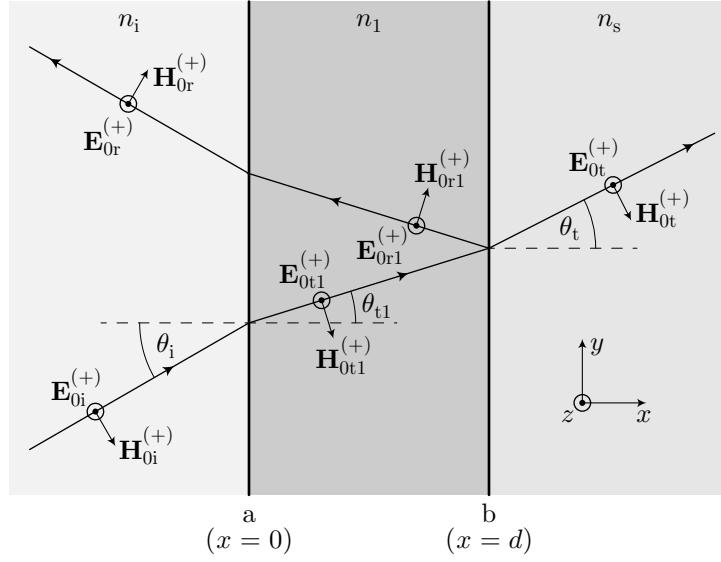
We can think of an arbitrary thin film as an etalon, with finesse

$$\mathcal{F} = \frac{\pi \sqrt{r_a r_b}}{1 - r_a r_b}. \quad (10.18) \quad (\text{finesse of single film})$$

Of course, the finesse depends implicitly on the incident angle via the angle-dependence of r_a . The transmitted angle is just given in terms of the incident angle by $n_i \sin \theta_i = n_s \sin \theta_t$.

10.2 Thin Films: Matrix Formalism

Now we will rederive the results from the last section by matching boundary conditions, as in Chapter 9. We will assume S-polarization, and later modify the results to handle P-polarization. The setup is shown in the diagram.



We are assuming fields as follows:

$$\begin{aligned}
 & \text{Incident Region: } \mathbf{E}_{0i}^{(+)} e^{i\mathbf{k}_i \cdot \mathbf{r}} + \mathbf{E}_{0r}^{(+)} e^{i\mathbf{k}_r \cdot \mathbf{r}} \\
 & \text{Thin-Film Region 1: } \mathbf{E}_{0t1}^{(+)} e^{i\mathbf{k}_{t1} \cdot \mathbf{r}} + \mathbf{E}_{0r1}^{(+)} e^{i\mathbf{k}_{r1} \cdot \mathbf{r}} \\
 & \text{Substrate/Transmission Region: } \mathbf{E}_{0t}^{(+)} e^{i\mathbf{k}_t \cdot (\mathbf{r} - d\hat{x})}.
 \end{aligned} \tag{10.19}$$

This is the same setup as for reflection at a *single* interface, except for the addition of left- and right-propagating fields inside the film. Notice also the phase offset in the transmitted field, so that the transmitted field has zero phase at boundary b.

As before, we apply boundary conditions. The transverse component of the electric field ($\mathbf{E} \cdot \hat{z}$) must be continuous across each boundary:

$$\begin{aligned}
 E_{0i}^{(+)} + E_{0r}^{(+)} &= E_{0t1}^{(+)} + E_{0r1}^{(+)} =: E_a^{(+)} \quad (\text{interface a}) \\
 E_{0t1}^{(+)} e^{i\delta} + E_{0r}^{(+)} e^{-i\delta} &= E_b^{(+)} =: E_b^{(+)} \quad (\text{interface b}).
 \end{aligned} \tag{10.20}$$

Here,

$$\delta := (\mathbf{k}_{t1} \cdot \hat{x})d = \frac{2\pi}{\lambda_0} n_1 d \cos \theta_{t1} \tag{10.21} \quad (\text{film propagation phase})$$

is the phase change traveling directly across the interface. Note that $\delta = \phi/2$, where we justified ϕ before in Eq. (10.6). In this case, we can think of matching phases everywhere, but considering the phases at *constant* y to ensure they are comparable, hence the “straight across” phase. Note that a tilted wave in the film region has an effective wavelength of $\lambda_0 / \cos \theta_{t1}$, along a horizontal slice of the wave.

Similarly, the transverse component of the magnetic field ($\mathbf{H} \cdot \hat{y}$) must be continuous across the boundary:

$$\begin{aligned}
 H_{0i}^{(+)} \cos \theta_i - H_{0r}^{(+)} \cos \theta_i &= H_{0t1}^{(+)} \cos \theta_{t1} - H_{0r1}^{(+)} \cos \theta_{t1} =: H_a^{(+)} \quad (\text{interface a}) \\
 H_{0t1}^{(+)} \cos \theta_{t1} e^{i\delta} - H_{0r1}^{(+)} \cos \theta_{t1} e^{-i\delta} &= H_b^{(+)} \cos \theta_t =: H_b^{(+)} \quad (\text{interface b}).
 \end{aligned} \tag{10.22}$$

We can rewrite the magnetic fields as

$$\begin{aligned}
 \alpha_i(E_{0i}^{(+)} - E_{0r}^{(+)}) &= \alpha_1(E_{0t1}^{(+)} - E_{0r1}^{(+)}) = H_a^{(+)} \\
 \alpha_1(E_{0t1}^{(+)} e^{i\delta} - E_{0r1}^{(+)} e^{-i\delta}) &= \alpha_s E_{0t}^{(+)} = H_b^{(+)},
 \end{aligned} \tag{10.23}$$

where the media are characterized by the α constants. The constants for the incident, film, and substrate media are

$$\begin{aligned}\alpha_i &:= \frac{n_i}{\eta_0} \cos \theta_i \\ \alpha_1 &:= \frac{n_1}{\eta_0} \cos \theta_{t1} \\ \alpha_s &:= \frac{n_s}{\eta_0} \cos \theta_t,\end{aligned}\tag{10.24}$$

(geometry coefficients)

where as usual, η_0 is the vacuum wave impedance.

Our goal is now to find a relation between the fields at interfaces a and b. Adding and subtracting the expressions for $E_b^{(+)}$ and $H_b^{(+)}$,

$$\begin{aligned}E_{0t1}^{(+)} &= \frac{\alpha_1 E_b^{(+)} + H_b^{(+)}}{2\alpha_1} e^{-i\delta} \\ E_{0r1}^{(+)} &= \frac{\alpha_1 E_b^{(+)} - H_b^{(+)}}{2\alpha_1} e^{i\delta}.\end{aligned}\tag{10.25}$$

Using the definitions of $E_a^{(+)}$ and $H_a^{(+)}$ in terms of the intra-film fields,

$$\begin{aligned}E_a^{(+)} &= E_{0t1}^{(+)} + E_{0r1}^{(+)} = \cos \delta E_b^{(+)} - \frac{i \sin \delta}{\alpha_1} H_b^{(+)} \\ H_a^{(+)} &= \alpha_1 (E_{0t1}^{(+)} - E_{0r1}^{(+)}) = -i \alpha_1 \sin \delta E_b^{(+)} + \cos \delta H_b^{(+)}. \end{aligned}\tag{10.26}$$

We can conveniently rewrite this in matrix form:

$$\begin{bmatrix} E_a^{(+)} \\ H_a^{(+)} \end{bmatrix} = \begin{bmatrix} \cos \delta & \frac{-i \sin \delta}{\alpha_1} \\ -i \alpha_1 \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} E_b^{(+)} \\ H_b^{(+)} \end{bmatrix} =: \mathbf{F} \begin{bmatrix} E_b^{(+)} \\ H_b^{(+)} \end{bmatrix}.\tag{thin-film transfer matrix} \quad (10.27)$$

The matrix \mathbf{F} is the **thin-film transfer matrix**, and has the general form

$$\mathbf{F} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.\tag{10.28}$$

The matrix “transfers” the fields from interface a to interface b.

As usual for matrix formalisms, this works for multiple films via matrix multiplication:

$$\begin{bmatrix} E_0^{(+)} \\ H_0^{(+)} \end{bmatrix} = \mathbf{F}_1 \mathbf{F}_2 \mathbf{F}_3 \cdots \mathbf{F}_N \begin{bmatrix} E_N^{(+)} \\ H_N^{(+)} \end{bmatrix}.\tag{composition of thin-film matrices} \quad (10.29)$$

Here,

$$\mathbf{F}_j = \begin{bmatrix} \cos \delta_j & \frac{-i \sin \delta_j}{\alpha_j} \\ -i \alpha_j \sin \delta_j & \cos \delta_j \end{bmatrix},\tag{thin-film matrices} \quad (10.30)$$

where the phase shifts are

$$\delta_j = \frac{2\pi}{\lambda_0} n_j d_j \cos \theta_{tj}\tag{10.31}$$

(thin-film phase shifts)

(d_j is the thickness of the j th layer), the geometry coefficients are

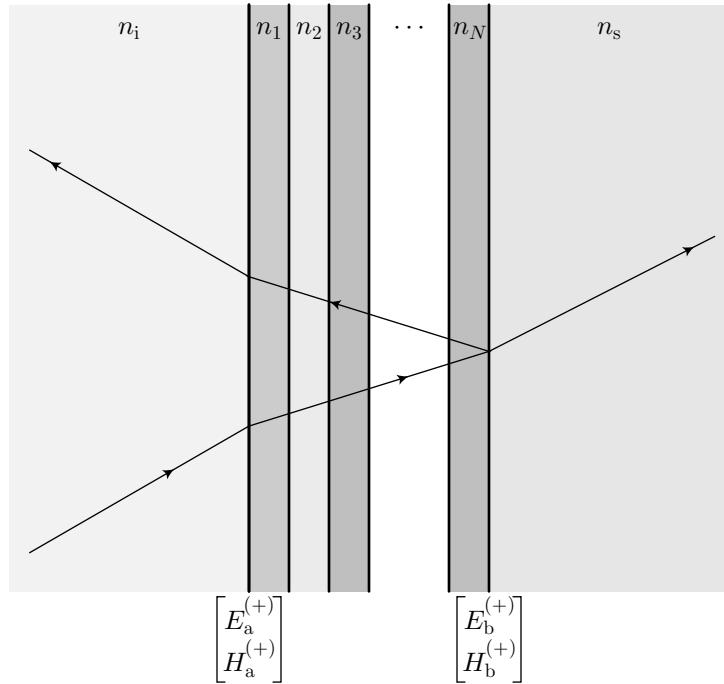
$$\alpha_j = \frac{n_j}{\eta_0} \cos \theta_{tj}, \quad (10.32)$$

(geometry coefficients)

and the transmission angles are given by

$$\theta_{tj} = \sin^{-1} \left(\frac{n_i}{n_j} \sin \theta_i \right). \quad (10.33)$$

(transmission angles)



Now we can solve for the reflected and transmitted fields. Using the definitions

$$\begin{aligned} E_a^{(+)} &= E_{0i}^{(+)} + E_{0r}^{(+)} \\ E_b^{(+)} &= E_{0t}^{(+)} \\ H_a^{(+)} &= \alpha_i (E_{0i}^{(+)} - E_{0r}^{(+)}) \\ H_b^{(+)} &= \alpha_s E_{0t}^{(+)} \end{aligned} \quad (10.34)$$

in the matrix equation (10.27), we find

$$\begin{bmatrix} E_{0i}^{(+)} + E_{0r}^{(+)} \\ \alpha_i (E_{0i}^{(+)} - E_{0r}^{(+)}) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E_{0t}^{(+)} \\ \alpha_s E_{0t}^{(+)} \end{bmatrix}. \quad (10.35)$$

This is equivalent to the system of equations

$$\begin{aligned} 1 + r &= At + B\alpha_s t \\ \alpha_i(1 - r) &= Ct + D\alpha_s t, \end{aligned} \quad (10.36)$$

where as usual $r := E_{0r}/E_{0i}$ and $t := E_{0t}/E_{0i}$. Solving these equations, we find the film reflection and transmission coefficients

$$\begin{aligned} r_{\text{film}} &= \frac{\alpha_i A + \alpha_i \alpha_s B - C - \alpha_s D}{\alpha_i A + \alpha_i \alpha_s B + C + \alpha_s D} \\ t_{\text{film}} &= \frac{2\alpha_i}{\alpha_i A + \alpha_i \alpha_s B + C + \alpha_s D} \end{aligned} \quad (\text{film reflection/transmission coefficients}) \quad (10.37)$$

in terms of the transfer-matrix elements.

The above analysis assumed S-polarization. It turns out that for P-polarization, the coefficients in Eqs. (10.37) are still valid, with a slight reinterpretation. The only modification is that in these relations and in the transfer matrix in Eq. (10.27), we must calculate the geometric coefficients α_j differently. For the two polarizations, we have

$$\begin{aligned} \alpha_j &= \frac{n_j}{\eta_0} \cos \theta_{tj} \quad (\text{S-polarization/TE polarization}) \\ \alpha_j &= \frac{n_j}{\eta_0 \cos \theta_{tj}} \quad (\text{P-polarization/TM polarization}). \end{aligned} \quad (\text{recipe for both polarizations}) \quad (10.38)$$

(The same modifications apply to the incident and substrate coefficients α_i and α_s .) With these changes, the expression for the reflection coefficient is still valid, though the expression for the transmission coefficient is more complicated. For most purposes, it is sufficient to note that the film reflection coefficient

$$r_{\text{film}} = \frac{\alpha_i A + \alpha_i \alpha_s B - C - \alpha_s D}{\alpha_i A + \alpha_i \alpha_s B + C + \alpha_s D} \quad (\text{film reflection coefficient, both polarizations}) \quad (10.39)$$

has the same form for both polarizations, and the transmittance can be computed from $T_{\text{film}} = 1 - R_{\text{film}} = 1 - |r_{\text{film}}|^2$.

Even though the formulae get a bit complicated, the general idea is simple: compute the α_j , compute the transfer matrix for the whole film, and then compute the reflection coefficient, and reflectance and transmittance if desired.

10.3 Optical Coating Design

Now we'll consider a few of the simplest optical coating designs in common use.

10.3.1 Single-Layer Antireflection Coating

Consider a single-layer film at normal incidence. The transfer matrix is just the single-layer matrix from Eq. (10.30)

$$\mathbf{F} = \begin{bmatrix} \cos \delta & \frac{-i \sin \delta}{\alpha_1} \\ -i \alpha_1 \sin \delta & \cos \delta \end{bmatrix}. \quad (10.40)$$

Putting the matrix elements into the expression from Eqs. (10.37) for the film reflection coefficient, we find

$$\begin{aligned} r &= \frac{\alpha_i \cos \delta - i \frac{\alpha_i \alpha_s}{\alpha_1} \sin \delta + i \alpha_1 \sin \delta - \alpha_s \cos \delta}{\alpha_i \cos \delta - i \frac{\alpha_i \alpha_s}{\alpha_1} \sin \delta - i \alpha_1 \sin \delta + \alpha_s \cos \delta} \\ &= \frac{n_1(n_i - n_s) \cos \delta - i(n_i n_s - n_1^2) \sin \delta}{n_1(n_i + n_s) \cos \delta - i(n_i n_s + n_1^2) \sin \delta}, \end{aligned} \quad (10.41)$$

where for normal incidence,

$$\alpha_i = \frac{n_i}{\eta_0}; \quad \alpha_1 = \frac{n_1}{\eta_0}; \quad \alpha_s = \frac{n_s}{\eta_0}, \quad (10.42)$$

and n_i is the refractive index of the incidence material (say, air), n_1 is the index of the coating and n_s is the index of the substrate (say, glass).

We can then calculate the reflectance (intensity reflection coefficient) of the film as

$$R = |r|^2 = \frac{n_1^2(n_i - n_s)^2 \cos^2 \delta + (n_i n_s - n_1^2)^2 \sin^2 \delta}{n_1^2(n_i + n_s)^2 \cos^2 \delta + (n_i n_s + n_1^2)^2 \sin^2 \delta}. \quad (\text{reflectance, single-layer film}) \quad (10.43)$$

This is still a bit messy, but we'll now make it simpler.

An important but simple case is the quarter-wave film (still at normal incidence). The thickness is

$$d = \frac{\lambda}{4} = \frac{\lambda_0}{4n_1}, \quad (10.44)$$

where it is important to note that the layer is a quarter wavelength for the wavelength *inside the medium*. This thickness leads to a phase shift of

$$\delta = \frac{2\pi}{\lambda_0} n_1 d \cos \theta_{t1} = \frac{\pi}{2}, \quad (10.45)$$

which means that $\cos \delta = 0$ and $\sin \delta = 1$. This leads to a reflectance of

$$R = \left(\frac{n_i n_s - n_1^2}{n_i n_s + n_1^2} \right)^2. \quad (\text{reflectance, single quarter-wave film}) \quad (10.46)$$

Notice that the reflectance vanishes if the film index is chosen such that

$$n_1 = \sqrt{n_i n_s}. \quad (\text{single-layer antireflection condition}) \quad (10.47)$$

As an aside, let's quickly compare this to the Fabry–Perot etalon discussion from the first part of this chapter. The reflectance vanishes even though there is only $\lambda/2$ of propagation over one full round trip. Normally, a Fabry–Perot cavity requires an integer multiple of λ of propagation distance per round trip, but now we have to be careful about phase changes on reflection. For example, if $n_i < n_1$, then the antireflection condition (10.47) implies that $n_1 < n_s$ (and $n_1 > n_s$ in the case that $n_i > n_1$). Thus, exactly one of the reflections inside the film will produce a π phase shift, playing the role of an extra $\lambda/2$ of propagation distance. This is different from the etalon that we considered early on in this chapter, which was for the case $n_i = n_s$, which requires a thickness of $\lambda/2$ for R to vanish.

Let's now consider the example of a real, single-layer antireflection (AR) coating. For an air–glass interface, we have $n_i = 1$ and $n_s = 1.52$ for optical crown glass. Ideally, $n_1 = 1.23$ for an AR coating (where R vanishes). The closest material suitable for making an optical coating (with good optical properties and durability) is MgF_2 , which has $n = 1.38$ (for visible wavelengths). Putting this into the formula (10.46), we find $R = 1.3\%$, which is much better than the uncoated reflectance of $R = 4.3\%$ but not really that close to zero. The cheapest optical coatings in the visible are single-layer MgF_2 .

10.3.2 Two-Layer Antireflection Coating

The basic problem with single-layer antireflection coatings is that there is a limited choice of suitable coating materials, and so it is difficult to match the optimum index condition very precisely. One solution to this problem is to add a second layer, in the hope that an extra free parameter will make the choice of materials more flexible.

Let's assume a stack of two quarter-wave films at normal incidence. The matrix for a single quarter-wave layer is

$$\mathbf{F}_1 = \begin{bmatrix} \cos \delta & -i \sin \delta \\ -i\alpha_1 \sin \delta & \cos \delta \end{bmatrix} = \begin{bmatrix} 0 & -\frac{i}{\alpha_1} \\ -i\alpha_1 & 0 \end{bmatrix}. \quad (10.48)$$

The transfer matrix for the two-layer stack is the product of two of these matrices:

$$\mathbf{F} = \mathbf{F}_1 \mathbf{F}_2 = \begin{bmatrix} 0 & -\frac{i}{\alpha_1} \\ -i\alpha_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -\frac{i}{\alpha_2} \\ -i\alpha_2 & 0 \end{bmatrix} = \begin{bmatrix} -\frac{\alpha_2}{\alpha_1} & 0 \\ 0 & -\frac{\alpha_1}{\alpha_2} \end{bmatrix}. \quad (10.49)$$

Again, putting the matrix elements into the expression from Eqs. (10.37) for the film reflection coefficient, we find

$$r = \frac{-\frac{\alpha_i \alpha_2}{\alpha_1} + \frac{\alpha_1 \alpha_s}{\alpha_2}}{-\frac{\alpha_i \alpha_2}{\alpha_1} - \frac{\alpha_1 \alpha_s}{\alpha_2}} = \frac{n_2^2 n_i - n_1^2 n_s}{n_2^2 n_i + n_1^2 n_s}. \quad (10.50)$$

The reflectance is thus

$$R = \left(\frac{n_2^2 n_i - n_1^2 n_s}{n_2^2 n_i + n_1^2 n_s} \right)^2. \quad (10.51)$$

(two-layer reflectance)

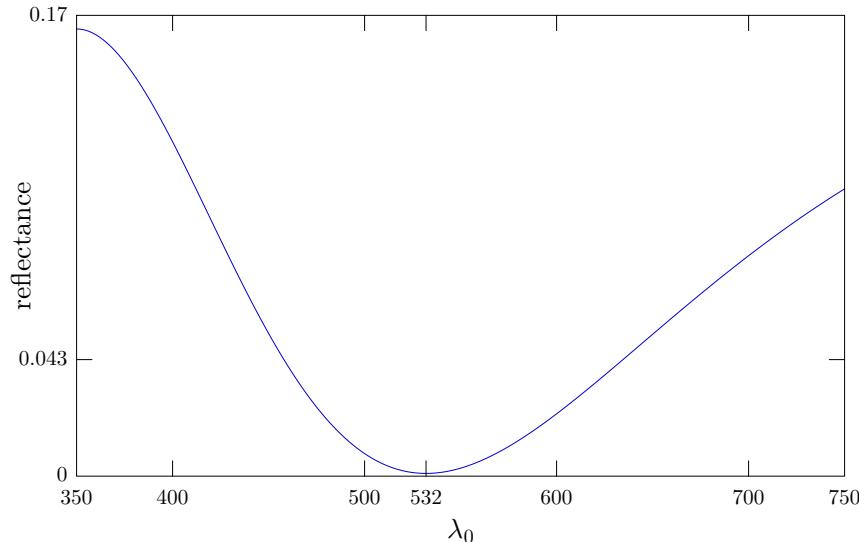
The reflectance vanishes if $n_2^2 n_i = n_1^2 n_s$, which gives the AR condition

$$\frac{n_2}{n_1} = \sqrt{\frac{n_s}{n_i}}, \quad (10.52)$$

(two-layer antireflection condition)

allowing us to pick the *ratio* of two refractive indices rather than a single absolute index.

As an example, again for crown glass ($n_i = 1$ and $n_s = 1.52$), the ideal AR ratio is $n_2/n_1 = 1.23$. We can choose ZrO_2 ($n_2 = 2.1$) and CeF_3 ($n_1 = 1.65$), which gives $n_2/n_1 = 1.27$, a value that is pretty close to the ideal value. Working out the reflectance, we find $R = 0.1\%$ for this coating (about as good as you can expect given fabrication tolerances). This kind of coating is called a “vee” coating due to the shape of the reflectance plot as a function of wavelength: there is a relatively narrow range (say, 10 nm or so) of good antireflection for a coating in the visible. In the example plotted here, the design wavelength is $\lambda_0 = 532$ nm, which is marked in the plot, and the uncoated air-glass reflectance is also marked in the plot.



It is possible to make a more broadband AR coating by adding a third layer, but we won't go into that here.

10.3.3 High Reflector: Quarter-Wave Stack

From the analysis of the two-layer AR coating, we note that if the ratio n_2/n_1 is very different from $\sqrt{n_s/n_i}$, then the reflectance is relatively large. Thus, the idea behind making a *high* reflectance (HR) coating is to essentially reverse the order of the layers in the antireflection coating: n_2 must have a larger index than n_1 for good reflection.

As an example, let's take the same materials as for the AR coating [ZrO_2 ($n_2 = 2.1$) and CeF_3 ($n_1 = 1.65$)], but reverse the order of the layers. In this case, the reflectance turns out to be $R = 18\%$. This is a much higher reflectance than the AR coating, although it is not as high as, say, a good metallic coating.

Of course, it is possible to get higher reflectances by choosing other materials with a greater index contrast. But these improvements will be marginal. How can we get a reflectance approaching unity? The general idea is to stack many of these double layers to form a **quarter-wave stack**.

Consider N double layers. The transfer matrix for one of the double layers at normal incidence is

$$\mathbf{F} = \begin{bmatrix} -\frac{n_2}{n_1} & 0 \\ 0 & -\frac{n_1}{n_2} \end{bmatrix}. \quad (10.53)$$

Thus, the transfer matrix for the whole stack is

$$\mathbf{F}^N = \begin{bmatrix} \left(-\frac{n_2}{n_1}\right)^N & 0 \\ 0 & \left(-\frac{n_1}{n_2}\right)^N \end{bmatrix}. \quad (10.54)$$

The reflection coefficient is

$$r = \frac{n_i \left(-\frac{n_2}{n_1}\right)^N - n_s \left(-\frac{n_1}{n_2}\right)^N}{n_i \left(-\frac{n_1}{n_2}\right)^N - n_s \left(-\frac{n_2}{n_1}\right)^N} = \frac{\left(\frac{n_i}{n_s}\right) \left(-\frac{n_2}{n_1}\right)^{2N} - 1}{\left(\frac{n_i}{n_s}\right) \left(-\frac{n_2}{n_1}\right)^{2N} + 1}. \quad (10.55)$$

Thus, the reflectance is

$$R = \left[\frac{\left(\frac{n_i}{n_s}\right) \left(-\frac{n_2}{n_1}\right)^{2N} - 1}{\left(\frac{n_i}{n_s}\right) \left(-\frac{n_2}{n_1}\right)^{2N} + 1} \right]^2. \quad (10.56)$$

(reflectance of dielectric stack, normal incidence)

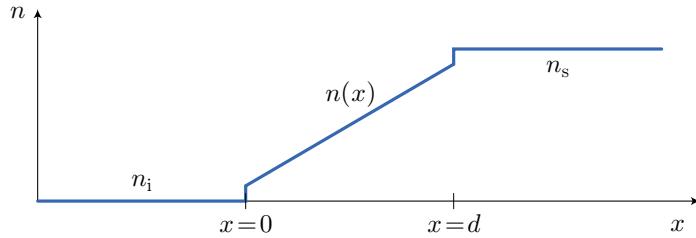
For the same materials as above, $n_i/n_s = 1/1.52$ and $n_2/n_1 = 1/1.27$, so for

$$\begin{aligned} N = 2 : R &= 36\% \\ N = 3 : R &= 53\% \\ N = 4 : R &= 68\% \\ N = 8 : R &= 94\% \\ N = 12 : R &= 99\%. \end{aligned} \quad (10.57)$$

Thus, we see that as the number of double layers increases, the reflectance rapidly (exponentially) converges to unity. We could use fewer layers with a better choice of materials (higher index contrast), although we should note that the intermediate reflectances shown here are also useful as beam splitters. The periodic structure of these quarter-wave stacks gives a simple example of a **photonic crystal** or **Bragg reflector**.

10.4 Gradient-Index Layers

So far we have only considered propagation and reflection due to layers of homogeneous media, and interfaces between them. But what if we have a layer where the refractive index varies *continuously*? We will again consider a planar geometry with a refractive index that varies only in the x -direction. Specifically, we will consider the case of a homogeneous “incident” medium of refractive index n_i occupying the region $x < 0$, a “substrate” medium of index n_s occupying $x > d$, and an arbitrarily and continuously varying index profile $n(x)$ in the region $0 \leq x \leq d$.



Note that the index profile need not be continuous at $x = 0$ or $x = d$, and $n(x)$ itself need not be continuous, though it should at least be piecewise continuous, and the motivation for this calculation comes from the continuous case.

The main motivation for such **gradient-index (GRIN)** layers is as an alternative to thin interference films to produce antireflection coatings for optical surfaces. The idea stems from the observation that abrupt changes in the refractive index cause reflections—the Fresnel reflection coefficient at normal incidence is $(n_1 - n_2)/(n_1 + n_2)$, so larger index steps cause larger reflections. One possible strategy is to try to avoid the abrupt index change at an optical surface by introducing a layer that smoothly interpolates the refractive index from air to glass. This could in principle be accomplished by many layers of gradually increasing index, but in practice the materials aren’t available to do this well. More practically, it is possible to shape a surface with subwavelength structures, such as the pattern shown here.



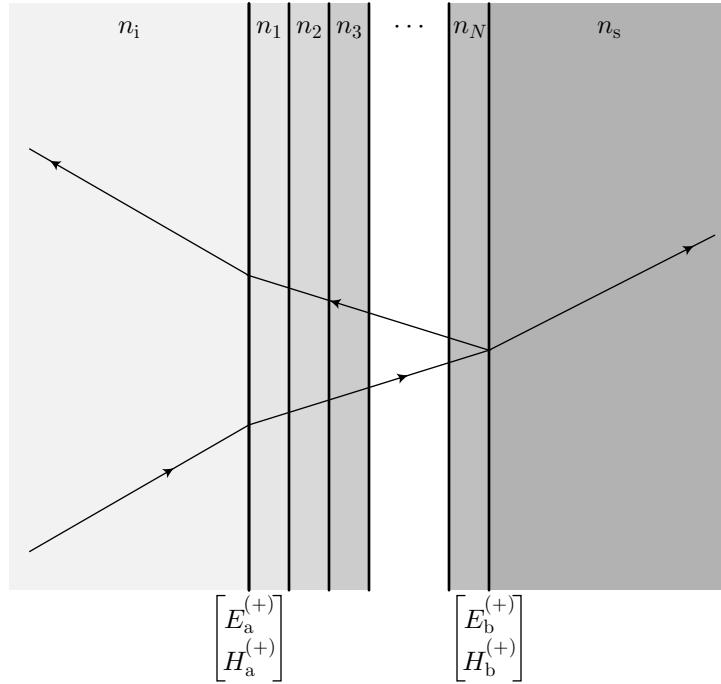
At first glance, this surface doesn’t seem too promising as an antireflection surface, as it is basically a diffraction grating. But if the period of the grating is sufficiently small (smaller than the optical wavelength), the light does not “see” the surface as a diffraction grating (this occurs when the angle of the first diffracted orders is so large as to be nonsensical, so that no orders except for the zeroth order propagate). In this case, we can coarse-grain the permittivity, or equivalently, the square of the refractive index, at each position x normal to the surface plane, by averaging the density of the material in the other directions at fixed x . (Note that the susceptibility and thus the permittivity give the direct polarization response of the medium to the incident field, and this is why the permittivity or squared index, not the index, should be coarse-grained.) If the index in the coarse-grained sense then changes sufficiently slowly in space, we might then expect reflections to be suppressed, in a way that should be robust to angle and wavelength changes, as compared to interference-based dielectric coatings. Such coatings are already available in commercial SLR camera lenses.¹

To work out the reflection coefficient for the GRIN layer, we will apply the thin-film matrix formalism by dividing up the GRIN layer into N sublayers of width

$$\delta x = \frac{d}{N}. \quad (10.58)$$

If N is large, then the sublayers are thin, and the refractive index is approximately constant over each layer, so that the matrix formalism applies.

¹For example, see <http://www.canon.com/news/2008/sep17e.html>, where the subwavelength structures seem to be cone-shaped.



We take advantage of the fact that δx is small to simplify the matrices, and we will also explicitly take the continuum limit $N \rightarrow \infty$ to develop a modified formalism to handle the continuous layer.

As in the matrix-formalism development, we will assume S-polarization, and we will indicate the changes necessary to handle P-polarization at the end.

10.4.1 Finite Stack of Very Thin Films

In a stack of thin films, the transfer matrix of the j th film is given by

$$\mathbf{F}_j = \begin{bmatrix} \cos \delta_j & \frac{-i \sin \delta_j}{\alpha_j} \\ -i \alpha_j \sin \delta_j & \cos \delta_j \end{bmatrix}. \quad (10.59)$$

Assuming the layer is very thin, of thickness δx , then the phase shift is similarly small,

$$\delta_j = \frac{2\pi}{\lambda_0} n_j(\delta x) \cos \theta_{tj} = k_0 n_j(\delta x) \cos \theta_{tj}, \quad (10.60)$$

and we can expand the transfer matrix to lowest order in δ_j :

$$\mathbf{F}_j \approx \begin{bmatrix} 1 & \frac{-i \delta_j}{\alpha_j} \\ -i \alpha_j \delta_j & 1 \end{bmatrix}. \quad (10.61)$$

Then with

$$\alpha_j = \frac{n_j}{\eta_0} \cos \theta_{tj}, \quad (10.62)$$

we can write this matrix directly in terms of refractive indices and angles as

$$\mathbf{F}_j \approx \begin{bmatrix} 1 & -i\eta_0 k_0 \delta x \\ -i\frac{k_0}{\eta_0} n_j^2 \cos^2 \theta_{tj} \delta x & 1 \end{bmatrix}. \quad (10.63)$$

But using Snell's law, we can rewrite the angle factor as

$$n_j^2 \cos^2 \theta_{tj} = n_j^2 - n_j^2 \sin^2 \theta_{tj} = n_j^2 - n_i^2 \sin^2 \theta_i, \quad (10.64)$$

and thus we have

$$\mathbf{F}_j \approx \begin{bmatrix} 1 & -i\eta_0 k_0 \delta x \\ -i\frac{k_0}{\eta_0} (n_j^2 - n_i^2 \sin^2 \theta_i) \delta x & 1 \end{bmatrix} \quad (10.65)$$

as the transfer matrix in the limit of very thin layers.

10.4.2 Continuous Medium

Rather than multiply the N matrices together as in the normal matrix formalism, we will take the continuum limit first, and then handle the propagation of the fields by converting the matrix relation into a system of coupled differential equations. Recall from Eq. (10.27) that the matrix \mathbf{F}_j “transfers” the *parallel components* of $\mathbf{E}^{(+)}$ and $\mathbf{H}^{(+)}$ from the output edge to the input edge of the j th layer. If we put this into continuous notation, where $\delta x \rightarrow dx$, $n_j \rightarrow n(x)$, and we regard the j th layer of the medium as occupying the region $x - dx$ to x , we have

$$\begin{bmatrix} E(x - dx) \\ H(x - dx) \end{bmatrix} = \begin{bmatrix} 1 & -i\eta_0 k_0 dx \\ -i\frac{k_0}{\eta_0} [n^2(x) - n_i^2 \sin^2 \theta_i] dx & 1 \end{bmatrix} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix}, \quad (10.66)$$

where again E and H refer *only* to the components of the fields parallel to the film surface (i.e., E_z and H_y in our coordinate system, for S-polarization), and note that we are dropping the $(+)$ superscripts to simplify the notation.

Expanding the left-hand side to first order in dx ,

$$\begin{bmatrix} E(x) - \frac{dE}{dx} dx \\ H(x) - \frac{dH}{dx} dx \end{bmatrix} = \begin{bmatrix} 1 & -i\eta_0 k_0 dx \\ -i\frac{k_0}{\eta_0} [n^2(x) - n_i^2 \sin^2 \theta_i] dx & 1 \end{bmatrix} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix}, \quad (10.67)$$

and canceling the common $E(x)$ and $H(x)$ terms, we have the linear differential system

$$\frac{d}{dx} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix} = ik_0 \begin{bmatrix} 0 & \eta_0 \\ \frac{1}{\eta_0} [n^2(x) - n_i^2 \sin^2 \theta_i] & 0 \end{bmatrix} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix}. \quad (10.68)$$

If we condense our notation by defining

$$n_\theta(x) := \sqrt{n^2(x) - n_i^2 \sin^2 \theta_i}, \quad (10.69)$$

(angle-corrected refractive index)

the linear system becomes

$$\frac{d}{dx} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix} = ik_0 \begin{bmatrix} 0 & \eta_0 \\ \frac{n_\theta^2(x)}{\eta_0} & 0 \end{bmatrix} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix}, \quad (10.70)$$

or

$$\begin{aligned} \frac{dE}{dx} &= ik_0 \eta_0 H(x) \\ \frac{dH}{dx} &= \frac{ik_0 n_\theta^2(x)}{\eta_0} E(x). \end{aligned} \quad (10.71)$$

if we write them out as separate equations. These equations allow us to propagate the fields from $x = d$ to $x = 0$, the procedure that takes the place of multiplying together the N matrices.

10.4.3 Reflection Coefficient

To relate the fields to the reflection coefficient, we can then use the same boundary conditions at the interfaces as in the matrix formalism. Rewriting the boundary conditions (10.34) for the transverse components of the fields at the incident (a), at $x = 0$, and substrate (b), at $x = d$, boundaries of the continuous film,

$$\begin{aligned} E_a^{(+)} &= E_{0i}^{(+)} + E_{0r}^{(+)} \\ E_b^{(+)} &= E_{0t}^{(+)} \\ H_a^{(+)} &= \alpha_i (E_{0i}^{(+)} - E_{0r}^{(+)}) \\ H_b^{(+)} &= \alpha_s E_{0t}^{(+)}. \end{aligned} \quad (10.72)$$

The coefficients α_i and α_s are defined in Eqs. (10.24). But instead of working with the individual fields, we will work with the ratios of the (parallel components of the) fields, which we will call *effective impedances* (effective because they only refer to parallel components)

$$\tilde{\eta}(x) := \frac{E^{(+)}(x)}{H^{(+)}(x)}. \quad (10.73)$$

At the incident boundary, Eqs. (10.72) give

$$\tilde{\eta}(0) = \frac{E_a^{(+)}}{H_a^{(+)}} = \frac{E_{0i}^{(+)} + E_{0r}^{(+)}}{\alpha_i (E_{0i}^{(+)} - E_{0r}^{(+)})} = \frac{1+r}{\alpha_i (1-r)} \quad (10.74)$$

and at the transmitting boundary,

$$\tilde{\eta}(d) = \frac{E_b^{(+)}}{H_b^{(+)}} = \frac{1}{\alpha_s} \quad (10.75)$$

In the region $0 < x < d$, we can deduce the dependence of $\tilde{\eta}$ by differentiating the definition (10.73):

$$\frac{d\tilde{\eta}}{dx} = \frac{1}{H^{(+)}} \frac{dE^{(+)}}{dx} - \frac{E^{(+)}(x)}{[H^{(+)}(x)]^2} \frac{dH^{(+)}}{dx}. \quad (10.76)$$

Then using Eqs. (10.71) to eliminate the field derivatives, we find that the effective impedance obeys the differential equation²

$$\frac{d\tilde{\eta}}{dx} = ik_0 \eta_0 - \frac{ik_0 n_\theta^2(x)}{\eta_0} \tilde{\eta}^2(x). \quad (10.77)$$

(propagation equation for fields)

²We have basically concocted an alternate derivation of the tapered-transmission-line formalism in Daniel H. Raguin and G. Michael Morris, “Analysis of antireflection-structured surfaces with continuous one-dimensional surface profiles,” Applied Optics, **32**, 2582 (1993) (doi: 10.1364/AO.32.002582); see Eq. (26) in particular.

The procedure for calculating the reflection coefficient is as follows: first, use Eq. (10.75) as the initial condition for $\tilde{\eta}(d)$, then use this differential equation to integrate *backwards* to $\tilde{\eta}(0)$, and finally, use Eq. (10.74) in the form

$$r = \frac{\alpha_i \tilde{\eta}(0) - 1}{\alpha_i \tilde{\eta}(0) + 1} \quad (10.78) \quad (\text{GRIN-layer reflection coefficient})$$

for the reflection coefficient.

10.4.4 Solution of Riccati Equations

The differential equation is nonlinear in $\tilde{\eta}$, which makes it difficult to solve in general. However, it is of a special form, being first-order and only quadratic. An ODE of this form is called a **Riccati equation**

$$\frac{dy}{dx} = a(x)y^2 + b(x)y + c(x), \quad (10.79) \quad (\text{Riccati equation})$$

which can be converted to a *linear* (but second-order) ODE by a change of variables. To do this, define $z(x)$ such that

$$y = -\frac{z'(x)}{a(x)z(x)}. \quad (10.80)$$

Differentiating this expression, we have

$$y' = \frac{z'}{a^2 z^2} (a' z + a z') - \frac{z''}{az} = \frac{a' z'}{a^2 z} + \frac{z'^2}{az^2} - \frac{z''}{az}. \quad (10.81)$$

But combining Eqs. (10.79) and (10.80), we have

$$y' = ay^2 + by + c = \frac{z'^2}{az^2} - \frac{bz'}{az} + c \quad (10.82)$$

Equating these last two expressions for y' , we have

$$z'' - \left(\frac{a'}{a} + b \right) z' + acz = 0 \quad (10.83)$$

after multiplying through by az . Assuming that we can solve *this* equation (which still may be tricky due to non-constant coefficients), the solution to the original equation (10.79) then follows from the transformation (10.80).

10.4.5 Example: Single, Homogeneous Film

As a simple check and illustration of the technique, we can work out a solution that we have already solved: the case where $n(x) = n_1$ is a constant over the full range $0 < x < d$. From Eq. (10.69), we have

$$n_\theta^2(x) = n_1^2 - n_1^2 \sin^2 \theta_i = n_\theta^2, \quad (10.84)$$

which is now constant, greatly simplifying the solution. Then we can take the coefficients

$$a = -\frac{ik_0 n_\theta^2}{\eta_0}, \quad b = 0, \quad c = ik_0 \eta_0, \quad (10.85)$$

and thus we should solve the equation

$$z'' + acz = 0, \quad (10.86)$$

with constant coefficients. We can certainly do this, because a and c are constants; we can parameterize the general solution to this equation as

$$z = \tilde{\beta} (\cos \sqrt{ac} x + \beta \sin \sqrt{ac} x), \quad (10.87)$$

where β and $\tilde{\beta}$ are undetermined coefficients. With the change of variables (10.80), the solution to the propagation equation (10.77) is

$$\tilde{\eta}(x) = -\sqrt{\frac{c}{a}} \left(\frac{-\sin \sqrt{ac}x + \beta \cos \sqrt{ac}x}{\cos \sqrt{ac}x + \beta \sin \sqrt{ac}x} \right). \quad (10.88)$$

With the boundary condition (10.75)

$$\tilde{\eta}(d) = \frac{1}{\alpha_s}, \quad (10.89)$$

we can equate this with the solution $\tilde{\eta}$ at $x = d$ to fix the coefficient

$$\beta = \frac{\alpha_s \sin \sqrt{ac}d - \sqrt{(a/c)} \cos \sqrt{ac}d}{\alpha_s \cos \sqrt{ac}d + \sqrt{(a/c)} \sin \sqrt{ac}d}. \quad (10.90)$$

Then we require the solution at the incident boundary,

$$\tilde{\eta}(0) = -\sqrt{\frac{c}{a}} \beta, \quad (10.91)$$

and the S-polarized reflection coefficient (10.78) is

$$r = \frac{\alpha_i \tilde{\eta}(0) - 1}{\alpha_i \tilde{\eta}(0) + 1} = \frac{\alpha_i \sqrt{(c/a)} \beta + 1}{\alpha_i \sqrt{(c/a)} \beta - 1}, \quad (10.92)$$

or with Eq. (10.90),

$$r = \frac{\alpha_i \alpha_s \sqrt{(c/a)} \sin \sqrt{ac}d - \alpha_i \cos \sqrt{ac}d + \alpha_s \cos \sqrt{ac}d + \sqrt{(a/c)} \sin \sqrt{ac}d}{\alpha_i \alpha_s \sqrt{(c/a)} \sin \sqrt{ac}d + \alpha_i \cos \sqrt{ac}d - \alpha_s \cos \sqrt{ac}d + \sqrt{(a/c)} \sin \sqrt{ac}d}. \quad (10.93)$$

To compare this with our previous results in the matrix formalism in Section 10.3.1, we can write argument of the trigonometric functions as

$$\sqrt{ac}d = k_0 n_\theta d = k_0 d \sqrt{n_1^2 - n_1^2 \sin^2 \theta_i} = k_0 d \sqrt{n_1^2 - n_1^2 \sin^2 \theta_{t1}} = n_1 k_0 d \cos \theta_{t1} = \delta, \quad (10.94)$$

and the other factor we need is $\sqrt{c/a}$ or $\sqrt{a/c}$. Note that we have to be careful about this, because c/a is negative, and there is a phase ambiguity: do we choose $\sqrt{c/a}$ or $\sqrt{a/c}$ to have a phase of $+\pi/2$? To go back to the source of this factor, recall from Eq. (10.88) that $\sqrt{c/a}$ was a shorthand for \sqrt{ac}/a ; since ac is positive, there is no phase ambiguity here, and we may proceed:

$$\sqrt{\frac{c}{a}} = \frac{\sqrt{ca}}{a} = \frac{k_0 n_\theta}{-ik_0 n_\theta^2 / \eta_0} = \frac{i\eta_0}{n_\theta} = \frac{i\eta_0}{n_1 \cos \theta_{t1}} = \frac{i}{\alpha_1}, \quad (10.95)$$

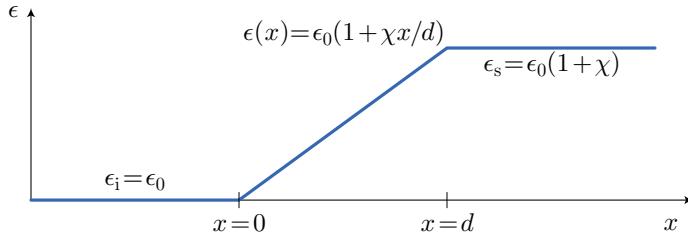
so that

$$r = \frac{-i \frac{\alpha_i \alpha_s}{\alpha_1} \sin \delta + \alpha_i \cos \delta - \alpha_s \cos \delta + i\alpha_1 \sin \delta}{-i \frac{\alpha_i \alpha_s}{\alpha_1} \sin \delta + \alpha_i \cos \delta + \alpha_s \cos \delta - i\alpha_1 \sin \delta}. \quad (10.96)$$

This is the same expression we wrote out before for the reflection coefficient in the matrix formalism, Eq. (10.41).

10.4.6 Example: Linear Permittivity Gradient

As a less-simple example, we will now work out the reflection coefficient due to a linear ramp in the permittivity, which interpolates linearly from the incident vacuum ($\epsilon_i = \epsilon_0$), to the substrate, of susceptibility χ [$\epsilon_s = \epsilon_0(1 + \chi)$]. This transition happens over a distance d .



This is a coarse-grained model of a subwavelength diffraction grating with triangular ridges, as illustrated here, where the substrate medium (e.g., glass) has susceptibility χ . Note that this is only a quantitative model of this particular grating for S-polarization and the polarization vector is parallel to the grooves; otherwise the effective index profile has a different functional form,³ because gratings are highly anisotropic, and the coarse-grained surface acts as if highly birefringent.



In terms of refractive indices, then, we have

$$\begin{aligned} n_i &= 1 \\ n(x) &= \sqrt{1 + \frac{\chi}{d}x} \\ n_s &= \sqrt{1 + \chi}. \end{aligned} \tag{10.97}$$

From Eq. (10.69), we have

$$n_\theta^2(x) = 1 + \frac{\chi}{d}x - \sin^2 \theta_i = \cos^2 \theta_i + \frac{\chi}{d}x, \tag{10.98}$$

and the Riccati coefficients in Eq. (10.79) become

$$a = -\frac{ik_0 n_\theta^2(x)}{\eta_0}, \quad b = 0, \quad c = ik_0 \eta_0, \tag{10.99}$$

and thus we should solve the equation

$$z'' - \left(\frac{a'}{a}\right) z' + acz = 0, \tag{10.100}$$

which becomes

$$z'' - \left(\frac{\chi/d}{\cos^2 \theta_i + (\chi/d)x}\right) z' + k_0^2 \left(\cos^2 \theta_i + \frac{\chi}{d}x\right) z = 0 \tag{10.101}$$

after putting in the explicit forms of the a and c coefficients.

To solve this equation, we will compare it to a known differential equation. The **Airy equation** is

$$y'' - xy = 0, \tag{10.102}$$

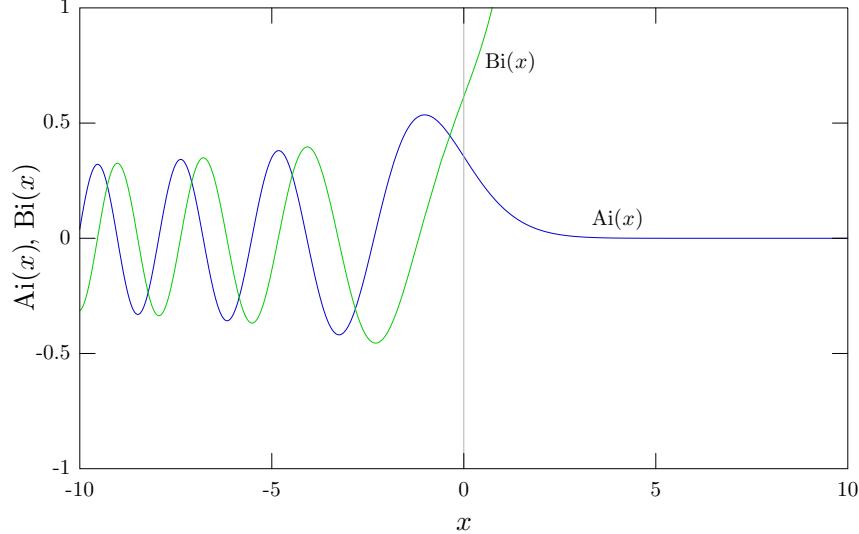
and two independent solutions of the Airy equation are the **Airy functions**⁴

$$\begin{aligned} \text{Ai}(x) &:= \frac{1}{\pi} \int_0^\infty dt \cos \left(\frac{t^3}{3} + xt \right) \\ \text{Bi}(x) &:= \frac{1}{\pi} \int_0^\infty dt \left[\exp \left(-\frac{t^3}{3} + xt \right) + \sin \left(\frac{t^3}{3} + xt \right) \right] \end{aligned} \tag{10.103}$$

³Daniel H. Raguin and G. Michael Morris, *op. cit.*, Eqs. (5) and (6).

⁴Milton Abramowitz and Irene A. Stegun, *Handbook of Mathematical Functions* (Dover, 1965), Section 10.4. Available online at http://people.math.sfu.ca/~cbm/aands/page_446.htm. The properties of the Airy function that we will use, including asymptotic forms, can be found there.

of the first (Ai) and second (Bi) kind. These functions are plotted below, where we will be interested mainly in the oscillatory part at negative x .



Taking the Airy equation and dividing through by x ,

$$\frac{1}{x}y'' - y = 0, \quad (10.104)$$

and then differentiating the equation, we find

$$\frac{1}{x}y''' - \frac{1}{x^2}y'' - y' = 0. \quad (10.105)$$

Changing variables to $z = y'$, we obtain the equation

$$z''(x) - \frac{1}{x}z'(x) - xz(x) = 0, \quad (10.106)$$

to which the solutions are the derivatives $\text{Ai}'(x)$ and $\text{Bi}'(x)$. Now changing the independent variable $x \rightarrow -(k_0^2\chi/d)^{1/3}x$, we obtain the equation

$$z'' - \frac{1}{x}z' + \frac{k_0^2\chi}{d}xz = 0, \quad (10.107)$$

to which the solutions are $\text{Ai}'[-(k_0^2\chi/d)^{1/3}x]$ and $\text{Bi}'[-(k_0^2\chi/d)^{1/3}x]$. Finally, we can change this into Eq. (10.101) by letting

$$x \rightarrow x + \frac{d}{\chi} \cos^2 \theta_i, \quad (10.108)$$

and thus we may write the general solution to Eq. (10.101) as

$$z(x) = \tilde{\beta} [\text{Ai}'(-\tilde{k}\tilde{x}) + \beta \text{Bi}'(-\tilde{k}\tilde{x})], \quad (10.109)$$

where β and $\tilde{\beta}$ are undetermined constants, and we are using the shorthand, shifted coordinate

$$\tilde{x} := \frac{dn_\theta^2(x)}{\chi} = x + \frac{d}{\chi} \cos^2 \theta_i \quad (10.110)$$

and scaled wave number

$$\tilde{k} := \left(\frac{k_0^2 \chi}{d} \right)^{1/3}. \quad (10.111)$$

Then using Eq. (10.80), the solution to the Riccati equation (10.77) in this case is

$$\begin{aligned}\tilde{\eta}(x) &= -\frac{\eta_0 d}{ik_0 \chi \tilde{x}} \tilde{k} \left(\frac{\text{Ai}''(-\tilde{k}\tilde{x}) + \beta \text{Bi}''(-\tilde{k}\tilde{x})}{\text{Ai}'(-\tilde{k}\tilde{x}) + \beta \text{Bi}'(-\tilde{k}\tilde{x})} \right) \\ &= \frac{k_0 \eta_0}{i \tilde{k}^3 \tilde{x}} \tilde{k}^2 \tilde{x} \left(\frac{\text{Ai}(-\tilde{k}\tilde{x}) + \beta \text{Bi}(-\tilde{k}\tilde{x})}{\text{Ai}'(-\tilde{k}\tilde{x}) + \beta \text{Bi}'(-\tilde{k}\tilde{x})} \right) \\ &= -\frac{ik_0 \eta_0}{\tilde{k}} \left(\frac{\text{Ai}(-\tilde{k}\tilde{x}) + \beta \text{Bi}(-\tilde{k}\tilde{x})}{\text{Ai}'(-\tilde{k}\tilde{x}) + \beta \text{Bi}'(-\tilde{k}\tilde{x})} \right),\end{aligned}\quad (10.112)$$

where the Airy equation (10.102) implies the second derivatives

$$\text{Ai}''(x) = x \text{Ai}(x), \quad \text{Bi}''(x) = x \text{Bi}(x). \quad (10.113)$$

The next step is to fix the undetermined constant β using the boundary condition (10.75)

$$\tilde{\eta}(d) = \frac{1}{\alpha_s}. \quad (10.114)$$

Letting

$$\tilde{\alpha} := \frac{\tilde{k}}{k_0 \eta_0} \quad (10.115)$$

and

$$\tilde{d} := \tilde{x}|_{x=d} = d \left(1 + \frac{\cos^2 \theta_i}{\chi} \right), \quad (10.116)$$

we find

$$\beta = -\frac{\alpha_s \text{Ai}(-\tilde{k}\tilde{d}) - i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d})}{\alpha_s \text{Bi}(-\tilde{k}\tilde{d}) - i\tilde{\alpha} \text{Bi}'(-\tilde{k}\tilde{d})}. \quad (10.117)$$

Then defining

$$\tilde{d}_0 := \tilde{x}|_{x=0} = \frac{d \cos^2 \theta_i}{\chi}, \quad (10.118)$$

the effective impedance at the incident boundary is

$$\tilde{\eta}(0) = \frac{1}{i\tilde{\alpha}} \left(\frac{\text{Ai}(-\tilde{k}\tilde{d}_0) + \beta \text{Bi}(-\tilde{k}\tilde{d}_0)}{\text{Ai}'(-\tilde{k}\tilde{d}_0) + \beta \text{Bi}'(-\tilde{k}\tilde{d}_0)} \right). \quad (10.119)$$

Finally, the reflection coefficient (10.78) becomes

$$\begin{aligned}r &= \frac{\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) + \alpha_i \beta \text{Bi}(-\tilde{k}\tilde{d}_0) - i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d}_0) - i\tilde{\alpha} \beta \text{Bi}'(-\tilde{k}\tilde{d}_0)}{\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) + \alpha_i \beta \text{Bi}(-\tilde{k}\tilde{d}_0) + i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d}_0) + i\tilde{\alpha} \beta \text{Bi}'(-\tilde{k}\tilde{d}_0)} \\ &= \frac{\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) - i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d}_0) + \beta [\alpha_i \text{Bi}(-\tilde{k}\tilde{d}_0) - i\tilde{\alpha} \text{Bi}'(-\tilde{k}\tilde{d}_0)]}{\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) + i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d}_0) + \beta [\alpha_i \text{Bi}(-\tilde{k}\tilde{d}_0) + i\tilde{\alpha} \text{Bi}'(-\tilde{k}\tilde{d}_0)]}.\end{aligned}\quad (10.120)$$

Then inserting Eq. (10.117), we finally find, condensing our notation yet further, the reflection coefficient

for S-polarization

$$\begin{aligned}
r_s &= \frac{A_0^- B^- - B_0^- A^-}{A_0^+ B^- - B_0^+ A^-} \\
A_0^\pm &:= \alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) \pm i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d}_0), \quad A^- := \alpha_s \text{Ai}(-\tilde{k}\tilde{d}) - i\tilde{\alpha} \text{Ai}'(-\tilde{k}\tilde{d}) \\
B_0^\pm &:= \alpha_i \text{Bi}(-\tilde{k}\tilde{d}_0) \pm i\tilde{\alpha} \text{Bi}'(-\tilde{k}\tilde{d}_0), \quad B^- := \alpha_s \text{Bi}(-\tilde{k}\tilde{d}) - i\tilde{\alpha} \text{Bi}'(-\tilde{k}\tilde{d}) \\
\tilde{d} &:= d \left(1 + \frac{\cos^2 \theta_i}{\chi} \right), \quad \tilde{d}_0 := \frac{d \cos^2 \theta_i}{\chi} \\
\tilde{k} &:= \left(\frac{k_0^2 \chi}{d} \right)^{1/3}, \quad \tilde{\alpha} := \frac{\tilde{k}}{k_0 \eta_0} \\
\alpha_i &:= \frac{\cos \theta_i}{\eta_0} = \frac{1}{\eta_0} \sqrt{\frac{\tilde{d}_0 \chi}{d}}, \quad \alpha_s := \frac{\sqrt{1+\chi}}{\eta_0} \cos \theta_t = \frac{1}{\eta_0} \sqrt{\chi + \cos^2 \theta_i} = \frac{1}{\eta_0} \sqrt{\frac{\tilde{d} \chi}{d}},
\end{aligned}$$

(reflection coefficient, S-polarization, linear GRIN film) (10.121)

where we used $\beta = -A^-/B^-$ in our condensed notation.

10.4.7 Asymptotic Forms

10.4.7.1 Sharp-Boundary Limit

As a simple check on our calculation, we can compute the reflection coefficient in the limit $d = 0$, corresponding to a hard boundary between vacuum and the dielectric, without any smooth index gradient. First, analyzing the numerator of the reflection coefficient (10.121), note that terms proportional to $\alpha_i \alpha_s$ and $\tilde{\alpha}^2$ will cancel, leaving

$$\begin{aligned}
A_0^- B^- - B_0^- A^- &= -i\alpha_i \tilde{\alpha} \text{Ai}(0) \text{Bi}'(0) + i\alpha_s \tilde{\alpha} \text{Ai}(0) \text{Bi}'(0) - i\alpha_s \tilde{\alpha} \text{Ai}'(0) \text{Bi}(0) + i\alpha_i \tilde{\alpha} \text{Ai}'(0) \text{Bi}(0) \\
&= i[\text{Ai}'(0) \text{Bi}(0) - \text{Ai}(0) \text{Bi}'(0)] \alpha_i \tilde{\alpha} + i[\text{Ai}(0) \text{Bi}'(0) - \text{Ai}'(0) \text{Bi}(0)] \alpha_s \tilde{\alpha} \\
&= a_0(\alpha_i - \alpha_s),
\end{aligned}$$

where we have defined

$$a_0 := i\tilde{\alpha} [\text{Ai}'(0) \text{Bi}(0) - \text{Ai}(0) \text{Bi}'(0)] = -i\tilde{\alpha} [\text{Ai}(0) \text{Bi}'(0) - \xi_0 \text{Ai}'(0) \text{Bi}(0)] = -\frac{2i\tilde{\alpha}}{\sqrt{3}\Gamma(1/3)\Gamma(2/3)}, \quad (10.123)$$

and the last two equalities here follow from the expressions

$$\text{Ai}(0) = \frac{1}{3^{2/3}\Gamma(2/3)}, \quad \text{Ai}'(0) = -\frac{1}{3^{1/3}\Gamma(1/3)}, \quad \text{Bi}(0) = \frac{1}{3^{1/6}\Gamma(2/3)}, \quad \text{Bi}'(0) = \frac{3^{1/6}}{\Gamma(1/3)}. \quad (10.124)$$

The exact form of a_0 is unimportant, as it will cancel here anyway. The denominator of the reflection coefficient (10.121) is the same, except A_0^- and B_0^- are replaced by A_0^+ and B_0^+ , which changes the signs of all terms involving α_s . Thus, the reflection coefficient is

$$r_s = \frac{\alpha_i - \alpha_s}{\alpha_i + \alpha_s} = \frac{\cos \theta_i - \sqrt{1+\chi} \cos \theta_t}{\cos \theta_i + \sqrt{1+\chi} \cos \theta_t}, \quad (10.125)$$

which is the correct Fresnel coefficient for S polarization incident from vacuum onto a dielectric.

10.4.7.2 Small-Reflection Limit

In the limit of large d , which smooths the transition between the vacuum and dielectric, we expect a small reflection coefficient. In this limit, we can use the asymptotic forms for real x for the Airy functions

$$\text{Ai}(-x) \sim \frac{1}{\sqrt{\pi}x^{1/4}} \sin \left(\frac{2x^{3/2}}{3} + \frac{\pi}{4} \right), \quad \text{Bi}(-x) \sim \frac{1}{\sqrt{\pi}x^{1/4}} \cos \left(\frac{2x^{3/2}}{3} + \frac{\pi}{4} \right), \quad (10.126)$$

and the derivatives

$$\text{Ai}'(-x) \sim -\frac{x^{1/4}}{\sqrt{\pi}} \cos\left(\frac{2x^{3/2}}{3} + \frac{\pi}{4}\right), \quad \text{Bi}'(-x) \sim \frac{x^{1/4}}{\sqrt{\pi}} \sin\left(\frac{2x^{3/2}}{3} + \frac{\pi}{4}\right). \quad (10.127)$$

Then examining the components of the refection coefficient (10.121), we note that both terms in each coefficient A_0^\pm , B_0^\pm , A^- , and B^- scale as $d^{-1/6}$, if we note that $\tilde{k}\tilde{d}$ and $\tilde{k}\tilde{d}_0$ scale as $d^{2/3}$, and $\tilde{\alpha}$ scales as $d^{-1/3}$.

In the reflection coefficient (10.121),

$$r_s = \frac{A_0^- B^- - B_0^- A^-}{A_0^+ B^- - B_0^+ A^-}, \quad (10.128)$$

we will begin by evaluating the first term in the numerator:

$$A_0^- B^- = \alpha_i \alpha_s \text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d}) - i\tilde{\alpha} \left[\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}'(-\tilde{k}\tilde{d}) + \alpha_s \text{Ai}'(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d}) \right] - \tilde{\alpha}^2 \text{Ai}'(-\tilde{k}\tilde{d}_0) \text{Bi}'(-\tilde{k}\tilde{d}). \quad (10.129)$$

Comparing the first and last terms here in the limit of large d , we can see from the asymptotic forms (10.126) and (10.127) that the dependence on trigonometric functions will be the same. To compare them further, we can see that

$$\frac{\alpha_i \alpha_s \text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d})}{\tilde{\alpha}^2 \text{Ai}'(-\tilde{k}\tilde{d}_0) \text{Bi}'(-\tilde{k}\tilde{d})} = \frac{\alpha_i \alpha_s}{\tilde{\alpha}^2 \tilde{k} \sqrt{\tilde{d}_0 \tilde{d}}} = \frac{\chi}{\eta_0^2 \tilde{\alpha}^2 \tilde{k} d} = \frac{k_0^2 \chi}{\tilde{k}^3 d} = 1, \quad (10.130)$$

where we used Eqs. (10.121) to reduce the ratio. Thus, the first and last terms in Eq. (10.129) cancel in the large- d limit, and we have

$$A_0^- B^- = -i\tilde{\alpha} \left[\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}'(-\tilde{k}\tilde{d}) + \alpha_s \text{Ai}'(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d}) \right]. \quad (10.131)$$

Proceeding to the second term in the numerator of Eq. (10.128), we see that this is the same as the first term (10.131) if we interchange d and d_0 , and then α_i with α_s . Then from the last two equations in (10.121), we have

$$\frac{\alpha_s}{\alpha_i} = \sqrt{\frac{\tilde{d}}{\tilde{d}_0}}, \quad (10.132)$$

and writing out the numerator of (10.121),

$$\begin{aligned} A_0^- B^- - B_0^- A^- &= -i\alpha_i \tilde{\alpha} \left[\text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}'(-\tilde{k}\tilde{d}) + \sqrt{\frac{\tilde{d}}{\tilde{d}_0}} \text{Ai}'(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d}) \right] \\ &\quad - i\alpha_i \tilde{\alpha} \left[\sqrt{\frac{\tilde{d}}{\tilde{d}_0}} \text{Ai}(-\tilde{k}\tilde{d}) \text{Bi}'(-\tilde{k}\tilde{d}_0) + \text{Ai}'(-\tilde{k}\tilde{d}) \text{Bi}(-\tilde{k}\tilde{d}_0) \right] \end{aligned} \quad (10.133)$$

we can put in the asymptotic forms (10.126) and (10.127) to see that this vanishes.

In the denominator, we have different signs, and so the first cancellation of the first and last terms in Eq. (10.129) does not occur, because the sign of the $\tilde{\alpha}^2$ term changes (as does the sign of the $\tilde{\alpha}\alpha_s$ term), and so we have

$$A_0^+ B^- = 2\alpha_i \alpha_s \text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d}) - i\tilde{\alpha} \left[\alpha_i \text{Ai}(-\tilde{k}\tilde{d}_0) \text{Bi}'(-\tilde{k}\tilde{d}) - \alpha_s \text{Ai}'(-\tilde{k}\tilde{d}_0) \text{Bi}(-\tilde{k}\tilde{d}) \right]. \quad (10.134)$$

Again, the other term in the denominator has the same form, but with \tilde{d}_0 and \tilde{d} interchanged, and with $\alpha_i \rightarrow -\alpha_s$ and $\alpha_s \rightarrow -\alpha_i$:

$$A_0^- B^+ = 2\alpha_i \alpha_s \text{Ai}(-\tilde{k}\tilde{d}) \text{Bi}(-\tilde{k}\tilde{d}_0) + i\tilde{\alpha} \left[\alpha_s \text{Ai}(-\tilde{k}\tilde{d}) \text{Bi}'(-\tilde{k}\tilde{d}_0) - \alpha_i \text{Ai}'(-\tilde{k}\tilde{d}) \text{Bi}(-\tilde{k}\tilde{d}_0) \right]. \quad (10.135)$$

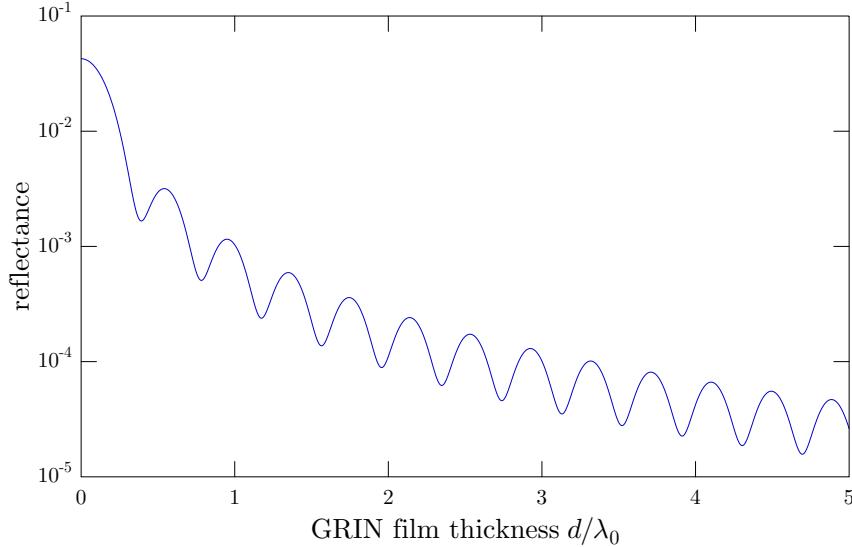
From the asymptotic forms (10.126) and (10.127), along similar lines to Eq. (10.133), we see that the imaginary terms are equal in magnitude but opposite in sign, so that the denominator becomes

$$\begin{aligned} A_0^+ B^- - A_0^- B^+ &= 2\alpha_i \alpha_s [\text{Ai}(-\tilde{k}\tilde{d}_0)\text{Bi}(-\tilde{k}\tilde{d}) + \text{Ai}(-\tilde{k}\tilde{d})\text{Bi}(-\tilde{k}\tilde{d}_0)] \\ &\quad + 2i\tilde{\alpha} [\alpha_s \text{Ai}(-\tilde{k}\tilde{d})\text{Bi}'(-\tilde{k}\tilde{d}_0) - \alpha_i \text{Ai}'(-\tilde{k}\tilde{d})\text{Bi}(-\tilde{k}\tilde{d}_0)] \\ &= \frac{2\alpha_i \alpha_s}{\pi (\tilde{k}^2 \tilde{d}_0 \tilde{d})^{1/4}} \sin \left[\frac{2}{3} \tilde{k}^{3/2} (\tilde{d}_0^{3/2} - \tilde{d}^{3/2}) \right] - \frac{2i\tilde{\alpha} \alpha_i}{\pi} \left(\frac{\tilde{d}}{\tilde{d}_0} \right)^{1/4} \cos \left[\frac{2}{3} \tilde{k}^{3/2} (\tilde{d}_0^{3/2} - \tilde{d}^{3/2}) \right], \end{aligned} \quad (10.136)$$

which is nonvanishing. Since we have a vanishing numerator and a nonvanishing denominator, we have established that $r_s \rightarrow 0$ as $d \rightarrow \infty$.

Note that the denominator scales as $d^{-1/3}$ for large d . The terms of the same order ended up canceling in the numerator. The next-order terms in the the asymptotic forms (10.126) and (10.127) are smaller by factors that scale as $x^{-3/2}$ for large x . Since x here refers to $\tilde{k}\tilde{d}$ and $\tilde{k}\tilde{d}_0$, both of which scale as $d^{2/3}$ the next-order part of the numerator should scale as $1/d$ compared to the denominator. Thus, for large d , the reflection coefficient scales as $1/d$ for large d , or the reflectance scales as $R \sim 1/d^2$.

To illustrate all this, below is plotted the reflectance at normal incidence for a vacuum–glass interface (assuming $n = 1.52$ for glass), connected by a linear permittivity gradient.



Here, we can see that the reflectance starts at 4.3% at $d = 0$, which is what we expect for a vacuum–glass interface. We also see oscillations in the reflectance, and in the large- d regime, the envelope of the oscillations shows the expected $1/d^2$ scaling. To gain a bit more insight into the oscillations, note that asymptotically, the Airy functions conspire to form trigonometric functions as in Eq. (10.136), with argument

$$\frac{2}{3} \tilde{k}^{3/2} (\tilde{d}^{3/2} - \tilde{d}_0^{3/2}) = \frac{2}{3} \left(\frac{k_0^2 \chi}{d} \right)^{1/2} \left[d^{3/2} \left(1 + \frac{1}{\chi} \right)^{3/2} - \left(\frac{d}{\chi} \right)^{3/2} \right] = \frac{2k_0 d}{3\chi} [(1 + \chi)^{3/2} - 1], \quad (10.137)$$

where to keep things simple we are considering only normal incidence. This has the form of the optical path length of the GRIN film,

$$\int_0^d dx' n(x') = \int_0^d dx' \sqrt{1 + \frac{\chi}{d} x'} = \frac{2d}{3\chi} [(1 + \chi)^{3/2} - 1], \quad (10.138)$$

multiplied by k_0 , and so the oscillations here are similar to what we expect in for a thin, *homogeneous* film, provided we interpret the result in terms of the more general path length.

10.4.8 P-Polarization

So far we have calculated the reflection coefficient in the case of S-polarization. But recall that we can adapt the calculation to P-polarization by moving the angle cosines from the numerators to the denominators of the α coefficients. Beginning with Eq. (10.61),

$$\mathbf{F}_j \approx \begin{bmatrix} 1 & -i\delta_j \\ -i\alpha_j\delta_j & 1 \end{bmatrix}. \quad (10.139)$$

we use

$$\alpha_j = \frac{n_j}{\eta_0 \cos \theta_{tj}} \quad (10.140)$$

along with

$$\delta_j = k_0 n_j (\delta x) \cos \theta_{tj} \quad (10.141)$$

to write the alternate transfer matrix

$$\mathbf{F}_j \approx \begin{bmatrix} 1 & -i\eta_0 k_0 \cos^2 \theta_{tj} \delta x \\ -i\frac{k_0}{\eta_0} n_j^2 \delta x & 1 \end{bmatrix} = \begin{bmatrix} 1 & -i\eta_0 k_0 \frac{(n_j^2 - n_i^2 \sin^2 \theta_i)}{n_j^2} \delta x \\ -i\frac{k_0}{\eta_0} n_j^2 \delta x & 1 \end{bmatrix}. \quad (10.142)$$

Passing over to the continuous limit, the differential system for the fields (10.70) becomes

$$\frac{d}{dx} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix} = ik_0 \begin{bmatrix} 0 & \frac{\eta_0 n_\theta^2(x)}{n^2(x)} \\ \frac{n^2(x)}{\eta_0} & 0 \end{bmatrix} \begin{bmatrix} E(x) \\ H(x) \end{bmatrix}, \quad (10.143)$$

where the definition (10.69) of $n_\theta(x)$ still holds. Again E and H refer *only* to the components of the fields parallel to the film surface (now E_y and H_z in our coordinate system, for P-polarization). Written as two coupled equations, we have

$$\begin{aligned} \frac{dE}{dx} &= ik_0 \eta_0 \frac{n_\theta^2(x)}{n^2(x)} H(x) \\ \frac{dH}{dx} &= \frac{ik_0 n^2(x)}{\eta_0} E(x). \end{aligned} \quad (10.144)$$

Then proceeding with the effective impedance $\tilde{\eta}(x)$ defined in Eq. (10.73), we can eliminate the derivatives in Eq. (10.76) to find in place of Eq. (10.77)

$$\frac{d\tilde{\eta}}{dx} = ik_0 \eta_0 \frac{n_\theta^2(x)}{n^2(x)} - \frac{ik_0 n^2(x)}{\eta_0} \tilde{\eta}^2. \quad (\text{propagation equation, P-polarization}) \quad (10.145)$$

That is, part of the space dependence $n_\theta^2(x)/n^2(x)$ of the variable index moves from the nonlinear term to the constant term. The procedure is then the same as before for calculating the reflection coefficient: first, use Eq. (10.75)

$$\tilde{\eta}(d) = \frac{E_b^{(+)}}{H_b^{(+)}} = \frac{1}{\alpha_s} \quad (10.146)$$

as the initial condition for $\tilde{\eta}(d)$ (with suitably redefined α_s), then use this differential equation to integrate backwards to $\tilde{\eta}(0)$, and finally, use Eq. (10.74),

$$r = \frac{\alpha_i \tilde{\eta}(0) - 1}{\alpha_i \tilde{\eta}(0) + 1}, \quad (10.147)$$

for the reflection coefficient, with the modified α_i .

10.4.8.1 Example: Linear Permittivity Gradient

Now we return to the example of the linear permittivity gradient from Section 10.4.6, and work out the reflection coefficient for P-polarization.

The Riccati coefficients in Eq. (10.79) in this case become

$$a = -\frac{ik_0 n^2(x)}{\eta_0}, \quad b = 0, \quad c = ik_0 \eta_0 \frac{n_\theta^2(x)}{n^2(x)}, \quad (10.148)$$

where Eq. (10.98) still holds for $n_\theta(x)$ in this case. We should then solve the equation

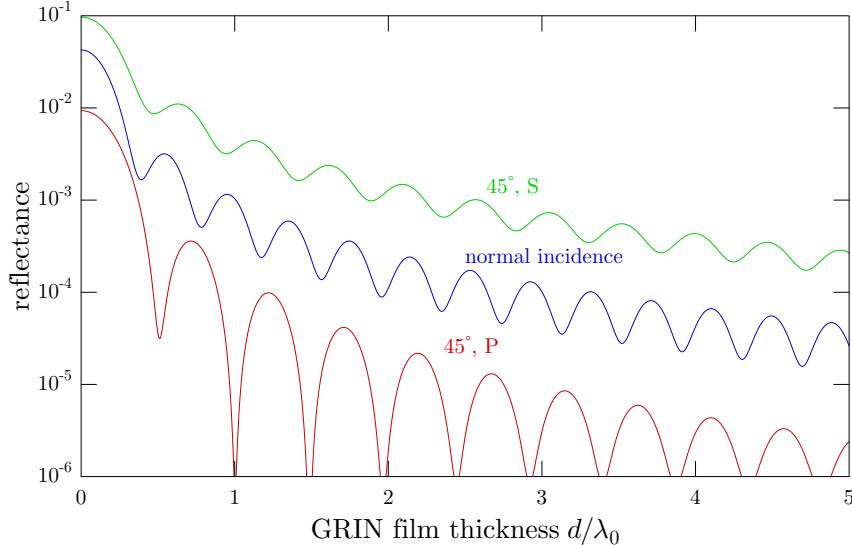
$$z'' - \left(\frac{a'}{a} \right) z' + acz = 0, \quad (10.149)$$

which becomes

$$z'' - \left(\frac{\chi/d}{1 + (\chi/d)x} \right) z' + k_0^2 \left(\cos^2 \theta_i + \frac{\chi}{d} x \right) z = 0, \quad (10.150)$$

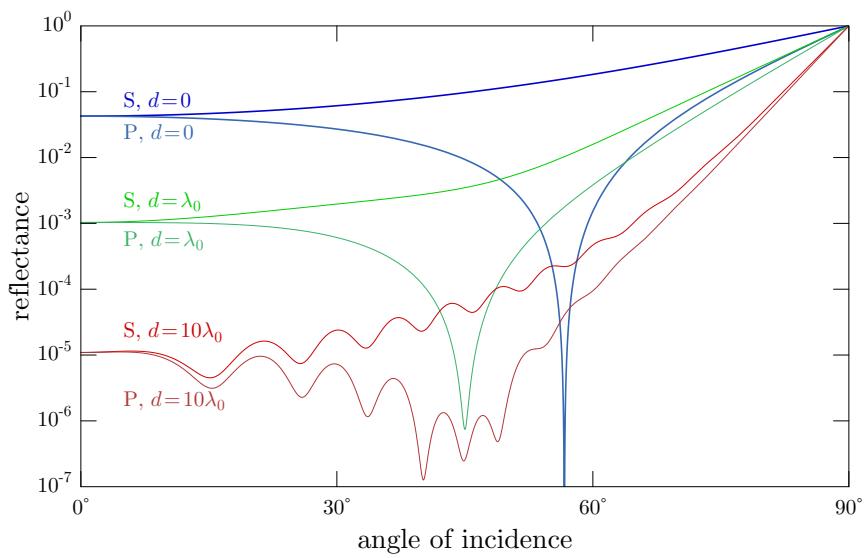
which is the same as Eq. (10.101), except that the cosine in the coefficient of z' is no longer there. Unfortunately, the variable-transformation tricks we used in the S-polarization case do not carry over here, and this equation does not have an obvious solution—except, that is, at normal incidence, when this case is equivalent to the S-polarization case.

Nevertheless, it is possible to solve Eq. (10.145) numerically to obtain the reflection coefficient. Doing this for the vacuum–glass ($n = 1.52$) interface for normal incidence, and 45° incidence for both polarizations leads to the plot below.



In the $d = 0$ limit, S-polarization reflects more than does P, because 45° is not far from Brewster's angle. This behavior persists for all d , with similar oscillations in each case (but with longer period in the angled-incidence cases).

Note, however, that while something analogous to the polarizing effect at Brewster's angle, the suppression is not as complete as for a simple interface, as shown in the plot below, which shows both polarizations for three film thicknesses (with the $d = 0$ case corresponding to the usual vacuum–glass interface).



However, notice that the reflectance is suppressed over a wide range of angles by the GRIN film.

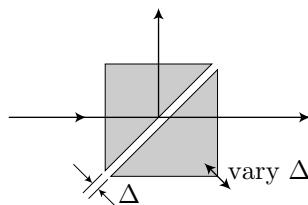
10.5 Exercises

Problem 10.1

Consider a planar glass window of refractive index n and thickness d , used as an etalon for light of wavelength λ incident at angle θ with respect to the window surface. Derive an expression for the changes in resonance wavelength $\delta\lambda$ and frequency $\delta\nu$, given a *small* change $\delta\theta_i$ in the incident angle. (That is, linearize the solution in $\delta\theta_i$.)

Problem 10.2

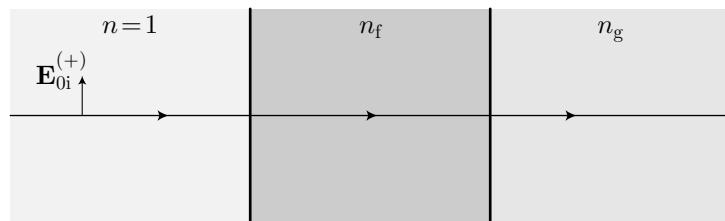
Consider a variable beam splitter consisting of two 45° - 45° - 90° prisms⁵ of refractive index n , separated by an air ($n = 1$) gap Δ and operated as shown with light of (vacuum) wavelength λ_0 . Assume that n is large enough that frustrated internal reflection occurs at the gap at the incident angle of 45° .



- (a) Write down an expression for the intensity reflection coefficient due to the air gap (ignore reflections from all other prism surfaces). Everything you write down should be defined in terms of parameters given in this problem, and make sure to treat both possible polarizations.
- (b) Obtain the limits of your expression for small and large Δ , and comment on whether or not they make sense.
- (c) Plot the reflectance of the beamsplitter as a function of Δ/λ_0 , assuming $n = 1.52$ for glass. Also include in your plot for comparison the simple estimate $1 - e^{-2\Delta/\delta}$, where δ is the skin depth for internal reflection, which is based on the normalized intensity of the evanescent wave where it enters the second prism (hence, this fraction of the intensity transmits through the second prism).

Problem 10.3

Light of wavelength λ_0 is incident from air ($n = 1$) onto a single dielectric thin film (of index n_f , and thickness $\lambda/4$, where λ is the wavelength inside the film), which covers a glass substrate (index n_g).



- (a) Write down an expression for the film reflectance, assuming the light is at normal incidence, using the results of the reflection-summation formalism.
- (b) Derive the value of n_f that makes a perfect anti-reflection coating.

Problem 10.4

Repeat the derivation of the thin-film matrix formalism for P-polarization, showing that the transfer-matrix formalism is the same as for the S-polarization case that we already derived, provided the α_j

⁵for a realization of this beam splitter, see D. Bertani, M. Cetica, and R. Polloni, "A simple variable-ratio beam splitter for holography," *Journal of Physics E: Scientific Instruments* **16**, 602 (1983) (doi: 10.1088/0022-3735/16/7/007). For application to Q-switched laser cavities, see Ian N. Court and Frederick K. von Willisen, "Frustrated Total Internal Reflection and Application of Its Principle to Laser Cavity Design," *Applied Optics* **3**, 719 (1964) (doi: dx.doi.org/10.1364/AO.3.000719).

coefficients are suitably redefined. Compute the film reflection and transmission coefficients for this case.

Problem 10.5

Discuss the advantage of a two-layer antireflection coating over a single-layer coating. Assume that you are only concerned with suppressing reflection at one incident wavelength and normal incidence, and distinguish *practical* vs. *in principle* performance of the coating in your answer. (Two or three sentences should suffice here.)

Problem 10.6

Plot the intensity reflection coefficients as a function of angle for light incident from air onto crown glass with a single-layer antireflection coating of MgF_2 . (Make plots for both polarizations.) Assume $n = 1.38$ for MgF_2 , and a coating thickness of $\lambda/4$ (where λ is the wavelength *inside* the coating!), where the design wavelength is $\lambda = 550 \text{ nm}$ (in vacuum). How thick is the coating in nm?

Problem 10.7

Plot the intensity reflection coefficients as a function of λ_0 for light incident from air onto crown glass with a double-layer antireflection coating. The thin-film stack consists of a $\lambda/4$ layer of ZrO_2 ($n = 2.1$) directly on top of the crown glass, followed by a $\lambda/4$ layer of CeF_3 ($n = 1.65$). Assume a design wavelength of 550 nm (in vacuum) and extend the plot over the visible spectrum (400-700 nm). Consider only the case of normal incidence. How thick are the layers in nm?

Problem 10.8

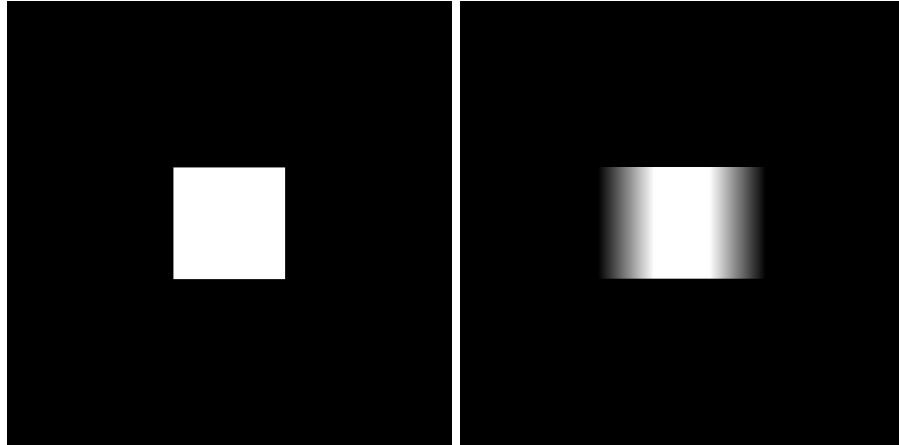
Consider a “slab” (film) of glass of index n surrounded on both sides by vacuum. Use the transfer-matrix method to derive the reflection coefficient for the film, assuming it is very thin ($d \ll \lambda_0, \lambda$), Taylor expanding to linear order in the film thickness d (at every stage, to keep it simple), and at normal incidence.

Chapter 11

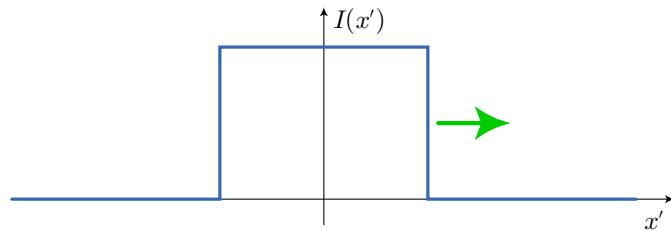
Fourier Analysis II: Convolution

11.1 Motion Blurring in Photography

To motivate the idea behind a convolution, let's consider the common situation of photographing a moving object. As an example, we'll choose a white box on a black background as the object, as shown below, to the left. As far as the camera is concerned, this is only the image that you will record in the limit of a very short shutter time. For longer shutter times, the image will be blurred, as shown below to the right.



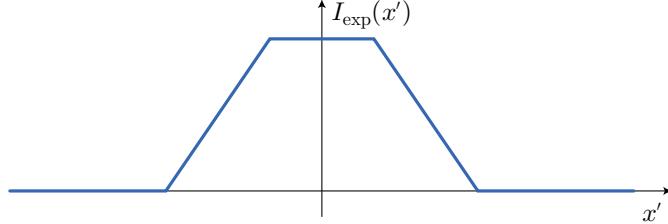
Let's represent this more mathematically, and consider the one-dimensional intensity function $I(x)$, representing the horizontal cross section through the middle of the figure. Again, this object is moving, say to the right.



During the shutter exposure, we must sum over the image in each location it occupies while it is moving along. We can represent this as the exposure-summed intensity

$$I_{\text{exp}}(x) = \int_{-d/2}^{d/2} I(x - x') dx'. \quad (11.1)$$

Here, we assume that the object moves a distance $d = vt_{\text{exp}}$ at speed v during the exposure time t_{exp} . In the integral, we add the intensity function centered at each x' within this range, and for concreteness, we take the range of the image's motion to be from $-d/2$ to $d/2$. The result is a blurred object, which would look something like the plot below, if the blurring effect is not too heavy.



Suppose that we introduce the box function of width d :

$$h(x) = \begin{cases} 1, & |x| \leq d/2, \\ 0, & |x| > d/2. \end{cases} \quad (11.2)$$

Then we can rewrite the exposure integral as

$$I_{\text{exp}}(x) = \int_{-\infty}^{\infty} I(x - x') h(x') dx'. \quad (11.3)$$

Here, $h(x)$ is unity to represent the locations of the object where the shutter is open, and it is zero when the shutter is closed. It need not be the case that $h(x)$ is only unity or zero; for example, $h(x)$ could ramp smoothly from zero to one as the shutter smoothly opens, or $h(x)$ could be more complicated if the object's velocity is nonuniform, since it spends different amounts of time in different locations. Considering this integral further, let's change variables by letting $x' \rightarrow x' + x$ and then $x' \rightarrow -x'$:

$$\begin{aligned} I_{\text{exp}}(x) &= \int_{-\infty}^{\infty} I(-x') h(x' + x) dx' \\ &= \int_{-\infty}^{\infty} I(x') h(x - x') dx'. \end{aligned} \quad (11.4)$$

What this is saying is that the blurring effect is equivalent when we have a *stationary* object, but a *moving* camera, represented by a moving function $h(x - x')$ in the integral (that is, we have just exchanged the dependences of I and h). This integral is precisely what we mean by a convolution.

11.2 Convolution

More abstractly, the **convolution** is a mathematical operation that maps a pair of functions to a function. Roughly speaking, the convolution is an operation that “smears” one function with another.

Formally, we can define the convolution $f * g$ of two functions f and g as the function given by the integral

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x') g(x - x') dx'. \quad (11.5)$$

(convolution of functions)

The interpretation of this integral is as follows: as a function of x , we displace (scan the location) of $g(x')$. It is most intuitive to consider the (common) case of a *localized* function $g(x')$, which we will refer to as the **convolution kernel**. Then the value of the convolution at location x is the “amount” of $f(x')$ that “passes through” $g(x')$ when g is displaced by an amount x . Notice that this is proportional to the *projection* of $f(x')$ onto $g(x - x')$.

As an example, consider the perfectly localized convolution kernel $g(x) = \delta(x)$. The convolution with a function $f(x)$ is

$$(f * \delta)(x) = \int_{-\infty}^{\infty} f(x') \delta(x - x') dx' = f(x). \quad (11.6)$$

The effect of convolution with a delta function is simply to do nothing; there is no “smearing” due to a perfectly localized function. Typically, we will use centered kernels; the effect of a displaced kernel is simply to displace the convolution by the same amount. For example, if

$$g(x) = \delta(x - x_0), \quad (11.7)$$

then

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x') \delta(x - x_0 - x') dx' = f(x - x_0), \quad (11.8)$$

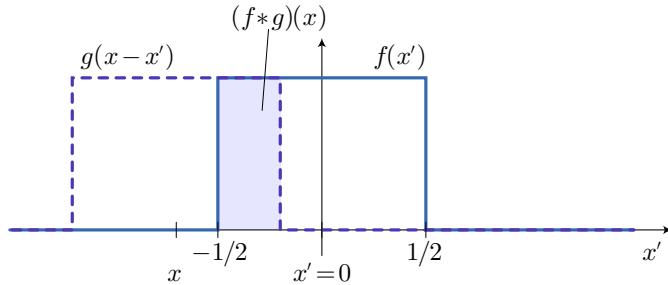
and thus, convolution with a displaced delta function is just a displacement of the function.

11.2.1 Example: Convolution of Box Functions

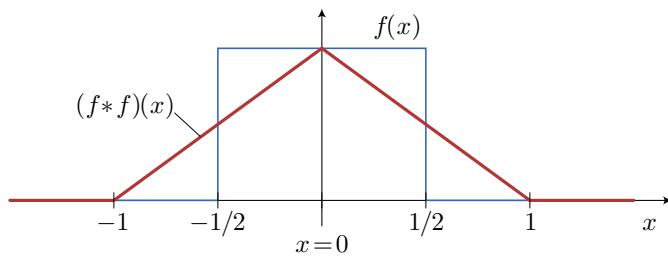
As a slightly more complicated example, consider the convolution of box functions, both given by

$$f(x) = g(x) = \begin{cases} 1, & |x| \leq 1/2, \\ 0, & |x| > 1/2. \end{cases} \quad (11.9)$$

The convolution consists of displacing $g(x')$ by x , multiplying the functions together, and integrating. For this simple case (box functions of unit height), the convolution (product) just turns out to be the area where the two functions overlap.



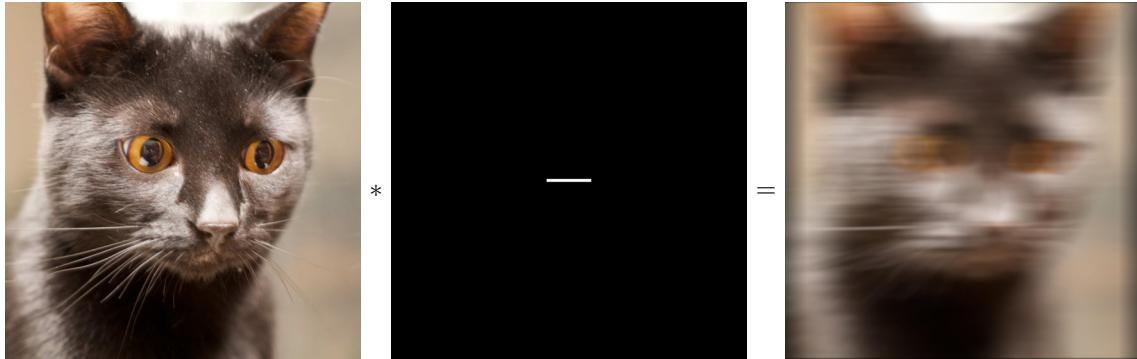
When the displacement is large, $|x| > 1$, the boxes don't overlap at all, so the convolution is zero. Otherwise, the overlap area varies linearly with the displacement, so the convolution is a triangle function, peaking at unity.



(Note that $f * f$ in the figure is the same as $f * g$ for this special case of $f = g$.) We see now the “smoothing” or “blurring” effect of the convolution. The original functions were discontinuous, but the convolution is continuous. The convolution is also wider than the original functions. As we will see, continued convolutions will make the distribution look Gaussian.

11.2.2 Example: Photographic Blurring

Let's return to the original photographic example. A photograph is a two-dimensional image, but the above convolution (and convolution theorem) generalize readily to two dimensions (just do a convolution in both directions, or equivalently, do a two-dimensional Fourier transform, multiply, and invert the transform). Here, we will start off with the original photo on the left.



In the middle, we have the convolution kernel. To be precise, this is a 2048×2048 image, which is all zero except for a unit-height region in the center, 16 pixels high (this is intended to be as small as possible, but is this large so that the bright region shows up in this figure) and 256 pixels wide. Think of this as essentially a delta function in the vertical direction, and a square pulse in the horizontal direction. There is essentially no blurring in the vertical direction in the convolved image, but significant blurring (on the scale of the width of the white line) in the image. This models motion of the camera during an exposure along the white line; more complicated trajectories of the camera would correspond to a more complicated white squiggle in the kernel image. Note the dark borders in the convolution image occur because the image is assumed to be zero (black) outside its borders.

11.3 Convolution Theorem

The **convolution theorem** gives an easy way to evaluate the convolution integral in Eq. (11.5), both in an intuitive and a computational sense. The convolution theorem states that the Fourier transform of the convolution is the product of the Fourier transforms of the individual functions:

$$\mathcal{F}[f * g] = \mathcal{F}[f]\mathcal{F}[g]. \quad (11.10)$$

(convolution theorem)

Before working through the proof, we need to review the Fourier transform, and in particular the alternate convention for Fourier transforms between position and spatial frequency ($x-k$) vs. time and frequency ($t-\omega$).

11.3.1 Spatial Fourier Transforms

Previously, in Chapter 3, we've been dealing with time-frequency Fourier transforms. However, we will soon be dealing with Fourier transforms between space and spatial frequency. So before proceeding with the convolution theorem, we'll have a short digression to introduce the different convention. Recall that the formulae for the Fourier transform and the inverse transform are

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(\omega) e^{-i\omega t} d\omega, \quad \tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt. \quad (11.11)$$

(Fourier transform– ω - t convention)

The basis function was $\exp(-i\omega t)$, and this determines the form of the transform integrals. Recall that the right-going plane wave has the form $\exp[i(kx - \omega t)]$, so the basis harmonic function for the *spatial* part is $\exp(ikx)$. So to get the spatial convention from the temporal convention, we must make the replacements

$$t \rightarrow x, \quad \omega \rightarrow k, \quad i \rightarrow -i. \quad (11.12)$$

Thus, the transform equations for spatial functions are

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(k) e^{ikx} dk, \quad \tilde{f}(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} dx, \quad (11.13)$$

(Fourier transform– k - x convention)

where again $\tilde{f}(k) = \mathcal{F}[f(x)] = \mathcal{F}[f](k)$.

11.3.2 Proof

To prove the convolution theorem, we'll just compute the explicit form of $\mathcal{F}[f * g]$. This will be very much a physicist's proof, not a mathematician's proof, in that we'll just assume the functions are nice enough that all the integrals simply exist.

$$\begin{aligned}
 \mathcal{F}[f * g] &= \mathcal{F} \left[\int_{-\infty}^{\infty} dx' f(x') g(x - x') \right] \\
 &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dx' f(x') g(x - x') e^{-ikx} \\
 &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dx' f(x') e^{-ikx'} g(x - x') e^{-ik(x-x')} \\
 &= \int_{-\infty}^{\infty} dx' f(x') e^{-ikx'} \int_{-\infty}^{\infty} dx g(x - x') e^{-ik(x-x')}.
 \end{aligned} \tag{11.14}$$

Letting $x \rightarrow x + x'$ in the second integral,

$$\begin{aligned}
 \mathcal{F}[f * g] &= \int_{-\infty}^{\infty} f(x') e^{-ikx'} dx' \int_{-\infty}^{\infty} g(x) e^{-ikx} dx \\
 &= \mathcal{F}[f] \mathcal{F}[g].
 \end{aligned} \tag{11.15}$$

Thus, to convolve two functions, just follow this recipe: Fourier transform both functions, multiply them together, then compute the inverse Fourier transform. That is,

$$f * g = \mathcal{F}^{-1} \{ \mathcal{F}[f] \mathcal{F}[g] \}, \tag{11.16}$$

(convolution-theorem recipe)

if we write this procedure out mathematically.

11.3.3 Example: Convolution of Two Gaussians

Since it's easy to compute the Fourier transform of Gaussian distributions, let's use the convolution theorem to convolve two Gaussians. Let's write the two functions as

$$f(x) = A e^{-x^2/\alpha^2}, \quad g(x) = B e^{-x^2/\beta^2}. \tag{11.17}$$

We computed the Fourier transform of a Gaussian in Chapter 3, so we can write

$$\mathcal{F}[f](k) = \tilde{f}(k) = A\alpha\sqrt{\pi}e^{-\alpha^2 k^2/4}, \quad \mathcal{F}[g](k) = \tilde{g}(k) = B\beta\sqrt{\pi}e^{-\beta^2 k^2/4}. \tag{11.18}$$

Then the product of the Fourier transforms is

$$(\mathcal{F}[f] \mathcal{F}[g])(k) = AB\alpha\beta\pi e^{-(\alpha^2 + \beta^2)k^2/4}. \tag{11.19}$$

Finally, we invert the Fourier transform to obtain the convolution. To do this, we can adapt the formulae for $f(x)$ and $\tilde{f}(k)$ by first letting $\alpha \rightarrow \sqrt{\alpha^2 + \beta^2}$ and then $A \rightarrow AB\alpha\beta\sqrt{\pi}/\sqrt{\alpha^2 + \beta^2}$. Then we obtain the same Fourier transform above, so the inversion yields

$$(f * g)(x) = \mathcal{F}^{-1} \left[AB\alpha\beta\pi e^{-(\alpha^2 + \beta^2)k^2/4} \right] = \frac{AB\alpha\beta\sqrt{\pi}}{\sqrt{\alpha^2 + \beta^2}} \exp \left(-\frac{x^2}{\alpha^2 + \beta^2} \right). \tag{11.20}$$

Recall that the standard (normalized) form of the Gaussian is

$$\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x - \mu)^2}{2\sigma^2} \right), \tag{11.21}$$

where the μ is the mean and σ is the standard deviation (σ^2 is the variance), which we'll come back to shortly. The standard deviation is one measure of the width of a Gaussian function. Note that $f(x)$ has standard deviation $\alpha/\sqrt{2}$, $g(x)$ has standard deviation $\beta/\sqrt{2}$, and $(f * g)(x)$ has standard deviation $\sqrt{(\alpha^2 + \beta^2)/2}$. Thus, the convolution of Gaussians is still Gaussian, but the blurring effect of the convolution makes the convolved Gaussian wider than the original functions. Indeed, the convolved width is the width of the originals summed in quadrature.

11.4 Application: Error Analysis

One important application of the convolution is in probability theory, which carries over to analysis of statistical error of experimental data in physics. The mathematical problem is as follows: let X_1 and X_2 be *independent* random variables with **probability density functions** $f_1(x)$ and $f_2(x)$, respectively. That is, the probability that $X_{1,2}$ is between x and $x + dx$ is $f_{1,2}(x) dx$. These could represent the errors in measurements of quantities x_1 and x_2 , respectively, so that the measurement results are $x_1 + X_1$ and $x_2 + X_2$. Then we can ask, what is the probability density of $X_1 + X_2$, which corresponds to the error of the combined (derived) measurement of $x_1 + x_2 + 2$?

To answer this, we can note that $X_1 + X_2 = x$ for any pair of values of X_1 and X_2 that happen to add up to x . But then we must sum over all such pairs. The probability that both X_1 and X_2 will *both* have particular probabilities is the product of the individual probabilities since the variables are independent. Thus, expressing what we said in equation form,

$$\begin{aligned} \text{Prob}(X_1 + X_2 \text{ between } x \text{ and } x + dx) &= \\ \sum_{x',x''} \text{Prob}(X_1 \text{ between } x' \text{ and } x' + dx') \times \text{Prob}(X_2 \text{ between } x'' \text{ and } x'' + dx'' | x = x' + x''). \end{aligned} \quad (11.22)$$

We can translate this statement in terms of the probability densities and implement the constraint as a delta function. Letting $f_+(x)$ denote the probability density of $X_1 + X_2$,

$$f_+(x) dx = \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dx'' f_1(x') f_2(x'') \delta(x' + x'' - x) dx. \quad (11.23)$$

The factor of dx here tempers the infinite height of the delta function, so that the delta function registers unity when the constraint is met (it is useful to think of the delta function as being zero everywhere, but $1/dx$ at the origin, so that multiplying by dx and summing over all x gives unity). Evaluating the x'' integral, we see that the probability density of the sum is the convolution of the individual densities:

$$f_+(x) dx = \int_{-\infty}^{\infty} dx' f_1(x') f_2(x - x') dx = (f_1 * f_2)(x) dx. \quad (11.24)$$

(probability density of sum)

Note that this result is general in that it doesn't assume any particular form for $f_1(x)$ or $f_2(x)$.

In the special case where both distributions are Gaussian, from Eqs. (11.20) and (11.21) we can see that the variances are related by

$$\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2]. \quad (11.25)$$

(variance of sum)

The **variance** here is defined by

$$\text{Var}[X] := \langle (X - \langle X \rangle) \rangle \quad (11.26)$$

(variance)

for a random variable X . The angle brackets denote the **expectation value**,

$$\langle h(X) \rangle := \int_{-\infty}^{\infty} h(x) f(x) dx, \quad (11.27)$$

(expectation value)

where $h(x)$ is an arbitrary function and $f(x)$ is the probability density function of X . For the Gaussian probability density in the form of Eq. (11.21), the variance turns out to be σ^2 . Recalling that the standard deviation is the square root of the variance,

$$\sigma_X := \sqrt{\text{Var}[X]}, \quad (11.28) \quad (\text{standard deviation})$$

we can see that for Gaussians, when you add the random variables, the standard deviation adds in quadrature:

$$\sigma_{X_1+X_2} = \sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2}. \quad (11.29) \quad (\text{error propagation for sums})$$

This is the **law of error propagation** for addition in error analysis: if two independent measurements have uncertainties σ_1 and σ_2 , then the uncertainty of the sum is $\sqrt{\sigma_1^2 + \sigma_2^2}$. Actually, in terms of variances, these results turn out to hold also for non-Gaussian variables, so long as the variances are well-defined.

11.5 Application: Central Limit Theorem

Often, we make many independent measurements of some quantity; however, we might not know the underlying distribution, in which case it's difficult to make use of a result like Eq. (11.24). Fortunately, there is a very important result that often makes it unnecessary to know the underlying distribution. This is also the reason that the Gaussian probability distribution is so important.

Let X_1, \dots, X_N be independent, identically distributed random variables. Let $f(x)$ be the probability density function of each of the X_j . Defining the sum by

$$S_N := \sum_{j=1}^N X_j, \quad (11.30) \quad (\text{sum of random variables})$$

we will now ask, what is the probability density $f_{S_N}(x)$ of S_N ? Evidently, we can iterate Eq. (11.24) to obtain

$$f_{S_N}(x) = (f * f * \dots * f)(x), \quad (11.31) \quad (\text{probability density of } S_N)$$

where the result is the successive convolution of N copies of f (for $N - 1$ total convolution operations). However, it turns out that this distribution becomes simple for large enough N .

The **central limit theorem** states that, provided that the mean and variance of the X_j exist, with the mean $\mu = \langle X_j \rangle$ and variance $\sigma^2 = \text{Var}[X_j]$, the distribution $f_{S_N}(x)$ becomes asymptotically Gaussian for large N with

$$\langle S_N \rangle = N\mu, \quad \text{Var}[S_N] \longrightarrow N\sigma^2. \quad (11.32) \quad (\text{central limit theorem})$$

(The mean is an exact result, whereas the distribution and variance are valid for large N .) This is a rough statement, since “becomes asymptotically Gaussian” is an imprecise statement. So let's clean this up a bit.

The central limit theorem alternately that the probability density function $f_{Z_N}(x)$ of the centered, scaled statistic

$$Z_N := \frac{S_N - N\mu}{\sigma\sqrt{N}} \quad (11.33)$$

converges to the “standard normal” (Gaussian) distribution

$$f_{Z_N}(x) \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad (11.34) \quad (\text{asymptotic scaled distribution})$$

which is the special Gaussian with mean 0 and unit variance.

Let's prove this now¹. To evaluate the convolutions in Eq. (11.31), we need to employ the convolution theorem. Taking the Fourier transform of $f(x)$,

$$\begin{aligned}\tilde{f}(k) &= \int_{-\infty}^{\infty} f(x) e^{-ikx} dx \\ &= \sum_{j=0}^{\infty} \int_{-\infty}^{\infty} f(x) \frac{(-ikx)^j}{j!} dx \\ &= \sum_{j=0}^{\infty} \frac{(-ik)^j}{j!} \langle X^j \rangle \\ &= 1 - ik\mu - \frac{k^2(\sigma^2 + \mu^2)}{2} + O(k^3).\end{aligned}\tag{11.35}$$

Here, we Taylor-expanded e^{-ikx} and then used the fact that the terms of the expansion were proportional to expectation values $\langle X^j \rangle$, especially the first two, which follow from

$$\langle X_j \rangle = \mu, \quad \text{Var}[X_j] = \sigma^2 = \langle X_j^2 \rangle - \mu^2.\tag{11.36}$$

In particular, note that in probability theory the **characteristic function** of a probability density, given by

$$\tilde{f}(-k) = \langle e^{ikX} \rangle,\tag{11.37}$$

is an important tool for the manipulation of probabilities.

This is more cumbersome than necessary, so let's recompute the expansion in Eq. (11.35) for the centered, scaled variable

$$Z_j = \frac{X_j - \mu}{\sigma\sqrt{N}},\tag{11.38}$$

with corresponding probability density $f_Z(x)$. The centering effectively zeroes the mean, and the rescaling changes the factor in front of the variance, with the result

$$\tilde{f}_Z(k) = 1 - \frac{k^2}{2N} + O\left[\left(\frac{k}{\sqrt{N}}\right)^3\right].\tag{11.39}$$

Since

$$Z_N = \sum_{j=1}^N Z_j,\tag{11.40}$$

the convolution theorem says that to calculate the transform of the N -fold convolution, we just compute $\tilde{f}_Z(k)$ to the N th power:

$$\tilde{f}_{Z_N}(k) = [\tilde{f}_Z(k)]^N = \left(1 - \frac{k^2}{2N} + O\left[\left(\frac{k}{\sqrt{N}}\right)^3\right]\right)^N.\tag{11.41}$$

As N becomes large, we can neglect the higher order terms beyond the first, and then use the formula

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x\tag{11.42}$$

to see that for large N , the transform becomes

$$\tilde{f}_{Z_N}(k) = \exp\left(-\frac{k^2}{2}\right).\tag{11.43}$$

But now the inverse Fourier transform of $\exp(-k^2/2)$ is $\exp(-x^2/2)/\sqrt{2\pi}$, so f_{Z_N} converges to a standard normal distribution as $N \rightarrow \infty$.

¹This is the physicist's proof; the rigorous version is in T. W. Körner, *Fourier Analysis* (Cambridge, 1988), starting on p. 349.

11.5.1 Central Limit Theorem Application: Random Walk

The central limit theorem has one (of many) of its important applications in statistical mechanics. Suppose a random walker takes a random step of size X with probability density $f(x)$ between periodic intervals of duration Δt . Let's assume that

$$\langle X \rangle = 0, \quad \text{Var}[X] = \sigma^2. \quad (11.44)$$

After N steps (N large), what has the walker done? The central limit theorem says that the probability density of the accumulated displacement

$$S_N = \sum_{j=1}^N X_j \quad (11.45)$$

is Gaussian with zero mean and variance $N\sigma^2$. That is, the width (standard deviation) is $\sigma\sqrt{N}$. The probability distribution thus becomes asymptotically Gaussian with a time-dependent width of

$$\sigma(t) = \sigma\sqrt{\frac{t}{\Delta t}}. \quad (11.46) \quad (\text{asymptotic distribution width})$$

This random-walk behavior is characteristic of a **diffusion process**, which is a transport process by which the distribution grows as $t^{1/2}$,

$$\Delta x \sim D t^{1/2}, \quad (11.47) \quad (\text{diffusion law})$$

where for the random walker the **diffusion coefficient** is $D = \sigma/\sqrt{\Delta t}$. Note that within certain restrictions (we'll explore this more in the homework), the final distribution is Gaussian, independent of the one-step distribution. Some one-step distributions, such as the Cauchy distribution, lead to **anomalous diffusion** where the diffusion coefficient diverges.

11.5.2 Central Limit Theorem Application: Standard Deviation of the Mean

Returning again to error analysis, suppose we make independent measurements X_1, \dots, X_N of some quantity in the laboratory. The **sample mean** is

$$\mu_N := \frac{1}{N} \sum_{j=1}^N X_j. \quad (11.48)$$

We can rewrite this as

$$\mu_N = \frac{S_N}{N} = \mu + \frac{\sigma Z_N}{\sqrt{N}}, \quad (11.49)$$

where the first term represents the *true mean*, and the second is the experimental error (statistical fluctuation in the sample mean). Applying the central limit theorem, Z_N is approximately standard normal for large N , so μ_N is Gaussian with mean μ and standard deviation

$$\sigma_{\text{mean}} = \frac{\sigma}{\sqrt{N}}, \quad (11.50) \quad (\text{standard deviation of the sample mean})$$

where σ is the standard deviation of a single measurement. Thus, the **standard deviation of the mean** (also called the **standard error**) is σ/\sqrt{N} . This is why, by making many measurements, it is possible to increase the accuracy of a measured quantity. (However, the convergence is fairly slow with N , so for a high-accuracy measurement, N must be very large, or σ must be small.)

11.6 Application: Impulse Response and Green Functions

Let's consider a simple mechanical system, such as a damped harmonic oscillator subject to a forcing function $f(t)$:

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = f(t). \quad (11.51) \quad (\text{forced, damped harmonic oscillator})$$

The three terms on the left-hand side represent, respectively, the acceleration, damping, and harmonic restoring force. Rather than solving this directly for an arbitrary forcing function, let's assume a particular form for the forcing function: an impulse (delta function) at time t' :

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = \delta(t - t'). \quad (11.52)$$

The solution to this is fairly straightforward. There is no force before the impulse ($t < t'$), so the oscillator is at rest,

$$x(t < t') = 0, \quad (11.53)$$

assuming the boundary conditions $x(-\infty) = 0$ and $\dot{x}(-\infty) = 0$. At $t = t'$, the impulse "kicks" the oscillator. We can see this by solving Eq. (11.52) and integrating over a short interval around $t = t'$.

$$\int_{t'-\varepsilon}^{t'+\varepsilon} \ddot{x} dt = \int_{t'-\varepsilon}^{t'+\varepsilon} [\delta(t - t') - \gamma\dot{x} - \omega_0^2 x] dt. \quad (11.54)$$

In the limit as $\varepsilon \rightarrow 0$, the last two terms in the right-hand integral become negligible (being proportional to ε). Thus, as $\varepsilon \rightarrow 0$, we find

$$\Delta\dot{x} = 1 \quad (11.55)$$

across the impulse. This makes intuitive sense: an impulsive force changes the velocity, but there isn't yet time for any position change to take place. Since the velocity just *before* t' is zero, the velocity just *after* t' is 1. Thus, the solution for $t > t'$ is the usual damped harmonic oscillator solution with the initial conditions $x(t') = 0$ and $\dot{x}(t') = 1$, and so we can write (without deriving the solution)

$$x(t > t') = \frac{1}{\omega_\gamma} e^{-(\gamma/2)(t-t')} \sin[\omega_\gamma(t-t')], \quad (11.56)$$

where

$$\omega_\gamma := \sqrt{\omega_0^2 - (\gamma/2)^2} \quad (11.57)$$

is the oscillation frequency in the presence of damping.

Let's rewrite this as the **impulse-response solution**, the solution $x(t)$ to the equation in response to a unit impulsive force at $t = t'$:

$$g(t, t') = \begin{cases} 0, & t < t' \\ \frac{1}{\omega_\gamma} e^{-(\gamma/2)(t-t')} \sin[\omega_\gamma(t-t')], & t \geq t'. \end{cases} \quad (\text{Green function for damped harmonic oscillator}) \quad (11.58)$$

Then $g(t, t')$ satisfies Eq. (11.52), where we write out the derivative operators separately:

$$\left(\frac{\partial^2}{\partial t^2} + \gamma \frac{\partial}{\partial t} + \omega_0^2 \right) g(t, t') = \delta(t - t'). \quad (11.59) \quad (\text{Green-function condition})$$

We note also the property of the delta function

$$f(t) = \int_{-\infty}^{\infty} dt' f(t') \delta(t - t'), \quad (11.60)$$

so we can multiply through Eq. (11.59) by $f(t')$ and integrate over t' , with the result

$$\left(\frac{\partial^2}{\partial t^2} + \gamma \frac{\partial}{\partial t} + \omega_0^2 \right) \int_{-\infty}^{\infty} dt' f(t') g(t, t') = f(t). \quad (11.61)$$

This is equivalent to the original equation of motion (11.51),

$$\left(\frac{\partial^2}{\partial t^2} + \gamma \frac{\partial}{\partial t} + \omega_0^2 \right) x(t) = f(t). \quad (11.62)$$

provided we identify

$$x(t) = \int_{-\infty}^{\infty} f(t') g(t, t') dt'. \quad (11.63) \quad (\text{Green-function solution})$$

Thus, given the impulse-response function $g(t, t')$, we can write down the solution for *any* forcing function $f(t)$ in terms of an integral. In fact, noting that in this particular case that the time dependence is such that $g(t, t') = g(t - t')$, we can write

$$x(t) = \int_{-\infty}^{\infty} f(t') g(t - t') dt' = (f * g)(t). \quad (\text{Green-function solution as convolution}) \quad (11.64)$$

so that the general solution is the *convolution* of $f(t)$ with $g(t) := g(t, 0)$. This form holds whenever the only explicit time dependence in the system is in the forcing function. This is an example of a convolution method for solving linear system. This works because of the superposition principle and because any function can be represented as a superposition of delta functions, which is essentially the content of Eq. (11.60). This convolution or impulse-response method is an example of a more general method called the **Green-function solution** ($g(t, t')$ is the Green function for the damped harmonic oscillator that is initially undisturbed).

All this works more generally, as the Green-function equation (11.59) has the general form

$$L(t)g(t, t') = \delta(t - t'), \quad (11.65)$$

where $L(t)$ is a linear operator. If $L(t)$ is time-independent, as it is for the harmonic oscillator, then $g(t, t') = g(t - t')$:

$$Lg(t - t') = \delta(t - t'). \quad (11.66)$$

Of course then we can just shift t to write

$$Lg(t) = \delta(t). \quad (11.67)$$

Then we can write the formal solution for the Green function as

$$g(t) = L^{-1}\delta(t). \quad (11.68) \quad (\text{formal solution for general Green function})$$

All of the integration above is essentially just working out the specific expression for L^{-1} , which is of course different for every differential equation.

11.6.1 Frequency Domain

Just for fun, let's look at the Fourier transform of the Green function. A quick trick for doing so is to take the Fourier transform of Eq. (11.59). This is easy, since it is essentially just assuming that the time dependence is of the form $e^{-i\omega t}$. Thus, to take the Fourier transform, we just make the following replacements:

$$\frac{\partial}{\partial t} \rightarrow -i\omega, \quad g(t, t') \rightarrow \tilde{g}(\omega, t'), \quad \delta(t - t') \rightarrow \int_{-\infty}^{\infty} \delta(t - t') e^{i\omega t} dt = e^{i\omega t'}. \quad (11.69)$$

The resulting equation is

$$(-\omega^2 - i\gamma\omega + \omega_0^2)\tilde{g} = e^{i\omega t'}. \quad (11.70)$$

Solving for the Fourier transform \tilde{g} ,

$$\tilde{g}(\omega, t') = \frac{e^{i\omega t'}}{\omega_0^2 - \omega^2 - i\gamma\omega}, \quad (11.71)$$

which we see is a Lorentzian function in frequency. To employ the convolution theorem, we really just wanted the Fourier transform for the impulse at $t' = 0$ (the t' is only there to facilitate the convolution), so

$$\tilde{g}(\omega) = \tilde{g}(\omega, t' = 0) = \frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}. \quad (\text{transfer function of harmonic oscillator}) \quad (11.72)$$

We can also assume we know the Fourier transform of the driving function,

$$\tilde{f}(\omega) = \mathcal{F}[f(t)], \quad (11.73)$$

so the convolution theorem gives an alternate way to write the general solution:

$$x(t) = \mathcal{F}^{-1} [\tilde{f}(\omega) \tilde{g}(\omega)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \frac{\tilde{f}(\omega) e^{-i\omega t}}{\omega_0^2 - \omega^2 - i\gamma\omega}. \quad (\text{transfer-function solution}) \quad (11.74)$$

In this case, we see that in the frequency domain, the kernel $g(t)$ acts as a bandpass filter (centered on ω_0 , with a width controlled by γ) that passes some of the frequencies of the driving function $f(t)$ to produce the output $x(t)$. Thus $\tilde{g}(\omega)$ is the **transfer function** for the damped harmonic oscillator. The same idea holds in electronics, and what we have done here is formally equivalent to a signal $f(t)$ going through a band-pass filter ($R-L-C$ resonant circuit).

In the language of the linear operator L as in Eq. (11.68), we have computed the Fourier transform of this equation in the sense

$$\tilde{g}(\omega) = \frac{1}{\mathcal{F}[L]},$$

(formal Green-function solution, frequency domain) (11.75)

where the Fourier transform of the operator L is given by the same replacements in Eqs. (11.69). The Fourier transform of L thus defines transfer function, or the filtering action of the system or oscillator.

11.7 Application: Spectral Transmission

Finally, we can get to an application in optics. Recall from Chapter 7 that a Fabry–Perot cavity can act as a spectrum analyzer, so that we can view the frequency content of an arbitrary input wave. We only treated the Fabry–Perot cavity in the case of monochromatic light, however, so let's explore this in more detail.

Suppose we have a Fabry–Perot cavity with transmission function

$$T_{\text{cav}} = \frac{T_{\text{cav,max}}}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2\left(\frac{\pi\nu}{\text{FSR}}\right)}. \quad (11.76)$$

This depends only on a single frequency, so it isn't so apparent how scanning the cavity length and an input spectrum would fit in. So let's massage things a little. As a spectrum analyzer, we want to analyze the transmission with respect to small length changes, so we can write

$$\begin{aligned} \frac{\pi\nu}{\text{FSR}} &= \frac{\pi(\nu_q + \delta\nu)}{\text{FSR}} \\ &= \pi q + \frac{\pi\delta\nu}{\text{FSR}} \\ &= \frac{2\pi d}{\lambda} + \frac{\pi\delta\nu}{\text{FSR}} \\ &= \frac{2\pi d_0}{\lambda} + \frac{2\pi\delta d}{\lambda} + \frac{\pi\delta\nu}{\text{FSR}}. \end{aligned} \quad (11.77)$$

Here, d_0 is the nominal cavity length, δd is the (small) length change while scanning the cavity, and $\delta\nu$ is a small frequency deviation from the nominal resonance. Defining $\phi_0 := 2\pi d_0/\lambda$ as the overall offset phase due to the nominal cavity length (essentially a constant), we can write

$$\frac{\pi\nu}{\text{FSR}} = \frac{\pi(\delta\nu - \delta\nu_d)}{\text{FSR}} + \phi_0, \quad (11.78)$$

where

$$\delta\nu_d := -\frac{\text{FSR}}{\lambda/2} \delta d \quad (11.79)$$

is the change in the resonance frequency due to the cavity length change. Thus, we can write the functional dependence of $T_{\text{cav}} = T_{\text{cav}}(\delta\nu - \delta\nu_d)$ on both the input frequency and cavity length changes.

If the input spectrum is the “line-shape function” $g(\nu)$, we will assume that the line shape is sharply peaked about some center frequency ν_0 . Thus, we can define a *centered* line-shape function

$$g_c(\delta\nu) := g(\nu_0 + \delta\nu), \quad (11.80)$$

where $\delta\nu = \nu - \nu_0$. Then we can write the total transmission through the cavity as the amplitude $g_c(\delta\nu)$ at frequency $\delta\nu$, weighted by the cavity transmission coefficient at the same frequency, integrated over all frequencies:

$$T_{\text{cav,total}}(\delta d) \approx \int_{-\infty}^{\infty} T_{\text{cav}}(\delta\nu - \delta\nu_d) g_c(\delta\nu) d(\delta\nu). \quad (11.81)$$

(spectrum-analyzer output)

The approximation here is the extension of the integration to all frequencies, which is valid if $g(\nu)$ is sharply peaked. Thus, the cavity transmission is the convolution of the input spectrum with the cavity transmission spectrum. Since the latter is periodic in frequency (due to the free spectral range), the total transmission is still periodic, but the individual transmission peaks are broadened by the input line shape due to the convolution. As a spectrum analyzer, the cavity line shape (for a given finesse) represents a *resolution limit*, since any input spectrum on transmission through the cavity will have an apparent line width of at least $\delta\nu_{\text{FWHM}}$ due to the convolution.

11.8 Exercises

Problem 11.1

Show that the order of the convolution does not matter:

$$(f * g)(x) = (g * f)(x). \quad (11.82)$$

Problem 11.2

Show that the area of a convolution is the product of the areas.

$$\int_{-\infty}^{\infty} dx (f * g)(x) = \left[\int_{-\infty}^{\infty} dx f(x) \right] \left[\int_{-\infty}^{\infty} dx' g(x') \right]. \quad (11.83)$$

(In particular, note that convolution with a normalized kernel does not change the area of a function.)

Problem 11.3

Show that convolution within the integral of a product can be associated by flipping the convolution kernel. That is, show that

$$\int_{-\infty}^{\infty} dx (f * g)(x) h(x) = \int_{-\infty}^{\infty} dx f(x) (g^- * h)(x), \quad (11.84)$$

where $g^-(x) := g(-x)$. (Note that in the important case of an even convolution kernel, $g^-(x) = g(x)$.)

Problem 11.4

Prove the convolution derivative formula

$$(f * g)'(x) = (f' * g)(x) = (f * g')(x). \quad (11.85)$$

Problem 11.5

Let $f(x)$ denote the unit box function,

$$f(x) = \begin{cases} 1, & |x| \leq 1/2 \\ 0 & \text{elsewhere} \end{cases} \quad (11.86)$$

- (a) Let $f^{*N}(x)$ denote the convolution of $f(x)$ with itself $N - 1$ times. Write down the asymptotic form for $f^{*N}(x)$ in the limit of large N .
- (b) Make plots of $f(x)$, $(f * f)(x)$, $(f * f * f)(x)$, and $(f * f * f * f)(x)$. Include the corresponding asymptotic distribution on each plot. Comment on your results.

Problem 11.6

The “Gaussian blur” in image-processing programs is just a convolution with a Gaussian kernel. Other types of image blurring can be thought of as convolutions with other kernels (e.g., the trajectory of motion for camera shake, or the response function of a lens system for an unfocused camera). As a model of this, compute the convolution of a single Fourier component

$$f(x) = \cos(kx) \quad (11.87)$$

with a (normalized) Gaussian blurring kernel

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}. \quad (11.88)$$

Interpret your result in the limits of large and small spatial frequency k , and large and small blurring width σ .

Problem 11.7

Consider independent, identically distributed random variables X_1, \dots, X_N with Cauchy (Lorentzian) probability density functions

$$f(x) = \frac{1}{\pi(1+x^2)}. \quad (11.89)$$

(a) Show that the Fourier transform is given by

$$\tilde{f}(k) = e^{-|k|}. \quad (11.90)$$

Hint: it's much easier to compute the *inverse* Fourier transform of $\tilde{f}(k)$ and use the fact that the Fourier transform is invertible for reasonable functions.

(b) Show that the probability density of the mean

$$\mu_N := \frac{1}{N} \sum_{j=1}^N X_j \quad (11.91)$$

is also $f(x)$.

(c) A naive application of the central limit theorem leads to the conclusion that μ_N has a Gaussian probability distribution for large N , but in (b) you have shown that this is not the case. Explain why there is no contradiction.

Problem 11.8

The harmonically forced oscillator has the following equation of motion:

$$\ddot{x} + \gamma \dot{x} + \omega_0^2 x = f_0 \sin(\omega t). \quad (11.92)$$

Use the Green function method in the time domain to show that the motion of the driven, damped oscillator (in steady state) is

$$x(t) = \frac{f_0 \sin(\omega t + \delta)}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2 \gamma^2}}, \quad (11.93)$$

where

$$\tan \delta = \frac{\omega \gamma}{\omega_0^2 - \omega^2}. \quad (11.94)$$

You may find the following integral formula useful:

$$\int_0^\infty e^{-ax} \sin(bx) \sin(cx + d) dx = \frac{b[2ac \cos d + (a^2 + b^2 - c^2) \sin d]}{(a^2 + b^2)^2 + 2(a^2 - b^2)c^2 + c^4} \quad (a, b, c, d \text{ real}, a > 0). \quad (11.95)$$

Problem 11.9

Work out the Green function $g(t, t')$ for the differential equation

$$\frac{\partial x}{\partial t} = f(t), \quad (11.96)$$

for arbitrary “forcing” function $f(t)$. Write the solution $x(t)$ as a convolution involving $f(t)$ and $g(t, t') = g(t - t')$. Assume the boundary condition $g(t \rightarrow -\infty, t') = 0$.

Problem 11.10

Work out the Green function $g(t, t')$ for the differential equation

$$\frac{\partial^n x}{\partial t^n} = f(t), \quad (11.97)$$

for positive integer n and arbitrary “forcing” function $f(t)$, given the boundary condition that $g(t < t') = 0$ and any derivatives up to the $(n-2)$ th of $g(t, t')$ vanish at $t = t'$. Write the solution $x(t)$ as a convolution involving $f(t)$ and $g(t, t') = g(t - t')$.

Problem 11.11

Derive the Green function for the ODE

$$\dot{x} + \gamma x = f(t), \quad (11.98)$$

and write down the general solution $x(t)$ in terms of the forcing function $f(t)$.

Problem 11.12

In this problem you will calculate the Green function for the Poisson equation

$$\nabla^2 \phi(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\epsilon_0} \quad (11.99)$$

for the electric potential ϕ in terms of the source charge density $\rho(\mathbf{r})$. We can do this by using the three-dimensional Fourier-transform,

$$f(\mathbf{r}) = \frac{1}{(2\pi)^3} \int d^3k \tilde{f}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}}, \quad \tilde{f}(\mathbf{k}) = \int d^3r f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}. \quad (11.100)$$

The Green function satisfies

$$\nabla^2 G(\mathbf{r}, \mathbf{r}') = -\delta^3(\mathbf{r} - \mathbf{r}') = -\delta(x - x') \delta(y - y') \delta(z - z'). \quad (11.101)$$

- (a) Start by finding the Fourier transform $\tilde{G}(\mathbf{k}, \mathbf{r}')$ of $G(\mathbf{r}, \mathbf{r}')$ by computing the Fourier transform of Eq. (11.101), and solving the resulting equation.
- (b) Then find $G(\mathbf{r}, \mathbf{r}')$ by inverting the Fourier transform. Evaluate the \mathbf{k} integral in spherical coordinates. You may find the following integral formula useful:

$$\int_0^\infty dx \frac{\sin(ax)}{ax} = \frac{\pi}{2|a|} \quad (a \in \mathbb{R}). \quad (11.102)$$

- (c) Finally, use your expression for $G(\mathbf{r}, \mathbf{r}')$ to write down the Green-function solution (i.e., an integral solution) to Poisson's equation for an arbitrary charge density $\rho(\mathbf{r})$. Comment on your result, and the physical significance of the Green function.

Problem 11.13

In this problem we will work with the Hermite–Gaussian function

$$H_n(x) e^{-x^2} = (-1)^n \frac{d^n}{dx^n} e^{-x^2}, \quad (11.103)$$

consisting of the product of a Hermite polynomial and a Gaussian factor. The expression on the right-hand side follows from the explicit formula for the Hermite polynomials [see Eq. (6.102) in Problem 6.18]:

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}. \quad (11.104)$$

- (a) Show that the Fourier transform is given by

$$\mathcal{F}[H_n(x) e^{-x^2}](k) = \sqrt{\pi}(-ik)^n e^{-k^2/4}. \quad (11.105)$$

- (b) Derive the convolution formula

$$[H_m(x) e^{-x^2}] * [H_n(x) e^{-x^2}] = \frac{\sqrt{\pi}}{2^{(m+n+1)/2}} H_{m+n}\left(\frac{x}{\sqrt{2}}\right) e^{-x^2/2} \quad (11.106)$$

for these Hermite–Gauss functions.

(c) Derive the convolution formula

$$e^{-x^2/\alpha^2} * \left[H_n\left(\frac{x}{\beta}\right) e^{-x^2/\beta^2} \right] = \frac{\sqrt{\pi} \alpha \beta^{n+1}}{(\alpha^2 + \beta^2)^{(n+1)/2}} H_n\left(\frac{x}{\sqrt{\alpha^2 + \beta^2}}\right) e^{-x^2/(\alpha^2 + \beta^2)}. \quad (11.107)$$

That is, a convolution of the Hermite–Gauss function with a Gaussian leads to a stretched version of the original Hermite–Gauss function.

Problem 11.14

Another Hermite–Gauss function can be written

$$H_n(x) e^{-x^2/2} = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2}, \quad (11.108)$$

differing from Eq. (11.108) in the factor of 2 in the exponential. This is a better representation of the profile of a Hermite–Gaussian beam [see Eq. (6.74), where the factor of $\sqrt{2}$ appears in the Hermite-polynomial argument but not in the exponential], compared to the Hermite–Gauss function from Problem 11.13. However, this function is somewhat more difficult to work with.

(a) Recalling the generating function for the Hermite polynomials, Eq. (6.103),

$$g(x, t) := e^{2xt - t^2} = \sum_{n=0}^{\infty} H_n(x) \frac{t^n}{n!}, \quad (11.109)$$

show that

$$\mathcal{F}[g(x, t) e^{-x^2/2}](k) = \sqrt{2\pi} g(k, -it) e^{-k^2/2}, \quad (11.110)$$

and then show

$$\mathcal{F}[H_n(x) e^{-x^2/2}](k) = \sqrt{2\pi} (-i)^n H_n(k) e^{-k^2/2}. \quad (11.111)$$

That is, up to a constant factor, the Fourier transform of a Hermite–Gauss function has the same form as the original.

(b) What does the result from (a) say about Fraunhofer diffraction of a Hermite–Gauss beam, and how is this (qualitatively) consistent with the Hermite–Gaussian solution to the paraxial wave equation?

(c) Write down an integral expression for the convolution of two Hermite–Gauss functions of different orders.

Chapter 12

Fourier Optics

12.1 Fourier Transforms in Multiple Dimensions

Before introducing Fourier optics, we must extend the spatial convention for Fourier transforms that we introduced before. As opposed to time, space takes up three dimensions, so we need a Fourier transform in multiple dimensions. In *Cartesian* coordinates, this turns out to be simple: just take a one-dimensional Fourier transform along each direction. For example, in two dimensions, the inverse transform is

$$f(x, y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \tilde{f}(k_x, k_y) e^{i(k_x x + k_y y)}, \quad (\text{inverse Fourier transform, 2D}) \quad (12.1)$$

and the Fourier transform is

$$\tilde{f}(k_x, k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y) e^{-i(k_x x + k_y y)}, \quad (\text{Fourier transform, 2D}) \quad (12.2)$$

where again $\tilde{f}(k_x, k_y) = \mathcal{F}[f(x, y)]$. Fourier transforms can be extended to arbitrarily many dimensions, but two is sufficient for our purposes here.

12.2 Wave Propagation in Homogeneous Media

12.2.1 Fingerprints of Propagation

The electromagnetic wave equation gives us a formal method for propagating an optical wave forward in space. Essentially, we can do this because we know how plane waves propagate, and any field can be regarded as a superposition of plane waves. That's where the "Fourier" comes into Fourier optics: we use a Fourier transform to break up an arbitrary field into plane-wave components. Let's see how this works.

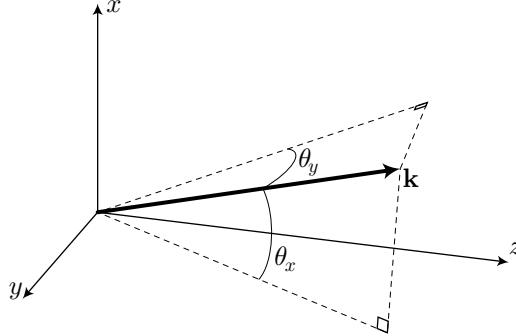
Recall the *scalar* plane wave (we won't do any vector theory, which is much more complicated) solution to the wave equation:

$$E^{(+)}(\mathbf{r}) = E_0^{(+)} e^{i\mathbf{k}\cdot\mathbf{r}} = E_0^{(+)} e^{i(k_x x + k_y y + k_z z)}, \quad (\text{scalar plane wave}) \quad (12.3)$$

where the wave vector $\mathbf{k} = (k_x, k_y, k_z)$ indicates the propagation direction of the wave. Let's let the optical axis lie along the z -axis. Then the wave vector \mathbf{k} has a direction determined by its angles with respect to the z -axis,

$$\theta_x = \sin^{-1} \left(\frac{k_x}{k} \right), \quad \theta_y = \sin^{-1} \left(\frac{k_y}{k} \right), \quad (\text{propagation angles of plane wave}) \quad (12.4)$$

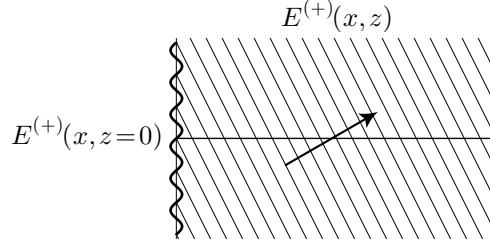
where k_x and k_y are the **transverse spatial frequencies**.



That is, in the $z = 0$ plane, the electric field has harmonic spatial dependence in the x - and y -directions:

$$E^{(+)}(x, y, z = 0) = E_0^{(+)} e^{ik_x x} e^{ik_y y}. \quad (12.5)$$

This is the central point of Fourier optics: harmonic phase variation in the (x, y) -plane (say, at $z = 0$) corresponds to a plane wave in a particular direction determined by Eqs. (12.4). That is, the wavelength of the transverse harmonic variation determines the direction of propagation—the wavelength of the wave is fixed (since we are assuming monochromatic light), and so only the proper angle can match the wave to the phase variation at $z = 0$.



In this sense, the transverse harmonic variation is a “fingerprint” of propagation in a certain direction. Often, we will only want to look at these in the paraxial approximation, where these relations are simpler. In the paraxial regime, we can write

$$\theta_x \approx \frac{k_x}{k}, \quad \theta_y \approx \frac{k_y}{k}, \quad (12.6) \quad (\text{paraxial propagation angles})$$

since $k_{x,y} \ll k$.

12.2.2 Decomposition

A general field profile $E^{(+)}(x, y)$ at $z = 0$ can thus be written as a superposition of plane waves via the Fourier transform:

$$E^{(+)}(x, y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \tilde{E}^{(+)}(k_x, k_y) e^{i(k_x x + k_y y)}. \quad (12.7)$$

We know how individual plane waves travel with z ,

$$e^{i(k_x x + k_y y)} \longrightarrow e^{i(k_x x + k_y y)} e^{ik_z z}, \quad (12.8)$$

where, since $k^2 = k_x^2 + k_y^2 + k_z^2$,

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2}. \quad (12.9)$$

So to propagate the full field forward in z , we just multiply each plane wave in the superposition by $e^{ik_z z}$:

$$E^{(+)}(x, y, z) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \tilde{E}^{(+)}(k_x, k_y) e^{i(k_x x + k_y y)} e^{ik_z z}. \quad (12.10)$$

It is important to keep in mind the functional dependence $k_z = k_z(k_x, k_y)$ here.

12.2.3 Reverse Waves

There is also a subtle point here in writing down Eq. (12.9), since

$$k_z = -\sqrt{k^2 - k_x^2 - k_y^2}. \quad (12.11)$$

is also a solution consistent with $k^2 = k_x^2 + k_y^2 + k_z^2$. These solutions correspond to k vectors with a component along the $-z$ -direction, and thus correspond to *backward*-propagating solutions. Thus, in writing down Eq. (12.9), we are building in the *assumption* that we have an entirely forward-propagating wave (i.e., all plane-wave components are propagating in the $+z$ -direction). Without this assumption, the time-independent wave equation

$$(\partial_z^2 + \partial_x^2 + \partial_y^2 + k^2) E^{(+)}(\mathbf{r}) = 0 \quad (12.12)$$

is not well-posed as an initial-value problem by specifying $E^{(+)}(x, y, z = 0)$ and propagating to other z . This is because the wave equation is second-order in z , and thus requires more information [e.g., specifying both $E^{(+)}(x, y, z = 0)$ and $\partial_z E^{(+)}(x, y, z = 0)$, or our forward-propagating assumption]. Of course, the *paraxial* wave equation (6.7)

$$(i2k\partial_z + \nabla_{\tau}^2) \psi = 0 \quad (12.13)$$

is only first-order in z , and thus only requires the initial condition $\psi(x, y, z = 0)$ to be well-posed. Recall that in deriving the wave equation, we explicitly assumed forward propagation, by assuming that all plane-wave components were close to a central “carrier” wave. Recall that the paraxial wave equation is formally equivalent to the Schrödinger equation in quantum mechanics, where the initial-value problem is well-posed by specifying only $\psi(\mathbf{r}, t = 0)$ and evolving to arbitrary time. The quantum-mechanical analog of our problem here is the Klein–Gordon equation, which for plane waves is basically the wave equation. In the Klein–Gordon equation, the analogues to backward-propagating waves are negative-energy solutions, which can be interpreted as antiparticles (particles traveling backwards in time). In optics, backward waves are formed from forward waves by gradients in the refractive index (such as refractive interfaces, which cause reflections); plane waves are eigenstates of homogeneous media, so only an inhomogeneity can scatter waves between different plane-wave states. (In quantum mechanics, these inhomogeneities correspond to time-dependent potentials).

12.2.4 Fourier-Transform Recipe

The theoretical development thus leads us to a fairly simple (in principle) recipe for propagating the field forward in space, given the form of the field in some plane. We can see that the information in the entire three-dimensional (monochromatic) field is thus encoded in any two-dimensional slice, a fact which motivates the field of holography. But for now, we wish to use this recipe as a calculational tool for understanding diffraction, so to summarize:

1. Start with an initial field $E^{(+)}(x, y)$ at $z = 0$.
2. Compute the Fourier transform of the initial field:

$$\tilde{E}^{(+)}(k_x, k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy E^{(+)}(x, y) e^{-i(k_x x + k_y y)}. \quad (\text{decomposition of initial field into plane waves}) \quad (12.14)$$

3. Multiply the transform $E^{(+)}(k_x, k_y)$ by the **free-space transfer function** $e^{ik_z z}$:

$$\tilde{E}^{(+)}(k_x, k_y) \longrightarrow \tilde{E}^{(+)}(k_x, k_y) e^{ik_z z}. \quad (12.15) \quad (\text{\textit{k}-space propagation})$$

Again, we need to apply this to the Fourier transform rather than the initial function because $k_z = k_z(k_x, k_y)$. In the paraxial approximation, we can alternatively use the paraxial form (12.18) below for the transfer function.

4. Finally, invert the Fourier transform to find the propagated field:

$$E^{(+)}(x, y, z) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \left[\tilde{E}^{(+)}(k_x, k_y) e^{ik_z z} \right] e^{i(k_x x + k_y y)}. \quad (\text{field propagated to arbitrary } z) \quad (12.16)$$

This recipe works because the superposition principle holds in linear optics and because it is easy to propagate plane waves.

12.2.5 Paraxial Propagation

The integral in Eq. (12.16) is in general difficult to evaluate because of the functional form of k_z in the exponent. However, in the paraxial approximation, the integral greatly simplifies. Since, $k_{x,y} \ll k$, we can write

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2} \approx k \left[1 - \frac{1}{2} \left(\frac{k_x^2 + k_y^2}{k^2} \right) \right] = k - \frac{k_x^2 + k_y^2}{2k}. \quad (12.17)$$

Thus, the free-space transfer function becomes

$$e^{ik_z z} \approx e^{ikz} e^{-i(k_x^2 + k_y^2)z/2k}, \quad (\text{free-space transfer function, paraxial approximation}) \quad (12.18)$$

where the first factor is an overall phase factor corresponding to plane-wave propagation along the optical axis, and the second factor generates the evolution of the spatial profile.

Note that in the paraxial limit, because k_x and k_y are assumed to be small compared to k , we can still keep the limits of integration in Eq. (12.16) to $\pm\infty$ because we assume that $\tilde{E}^{(+)}(k_x, k_y)$ has negligible “stuff” at large $k_{x,y}$.

12.2.5.1 Solution of the Paraxial Wave Equation

We have shown that the transfer function (12.18) produces the proper solution to the wave equation, once we have made the paraxial approximation. Of course, then, this should represent the solution to the paraxial wave equation (6.7)

$$\left(\nabla_{\tau}^2 + i2k \frac{\partial}{\partial z} \right) \psi = 0, \quad (12.19)$$

where the transverse Laplacian is

$$\nabla_{\tau}^2 := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \quad (12.20)$$

and the field envelope ψ is defined by

$$E^{(+)}(\mathbf{r}) = \psi(\mathbf{r}) e^{ikz}. \quad (12.21)$$

If we write the wave equation as

$$\frac{\partial}{\partial z} \psi = \left(\frac{i}{2k} \nabla_{\tau}^2 \right) \psi, \quad (12.22)$$

we have a differential equation of the form

$$\frac{\partial}{\partial z} \psi = A \psi, \quad (12.23)$$

for which the formal solution is

$$\psi(z) = e^{Az} \psi(0). \quad (12.24)$$

Here, “formal” means that we have to be careful that the exponential of the operator A exists (it turns out that this is an okay assumption); if A is a scalar, then there is no problem. Since A is an operator, the

operator exponential is defined by the Taylor expansion (powers of operators are obviously well-defined). Thus, the solution for the paraxial wave equation is

$$\psi(z) = e^{i\nabla_r^2 z/2k} \psi(0) = e^{i(\partial_x^2 + \partial_y^2)z/2k} \psi(0). \quad (12.25)$$

Recall that when operating on a plane wave with spatial dependence $e^{i\mathbf{k}\cdot\mathbf{r}}$, the derivatives are equivalent to k -vector components:

$$\frac{\partial}{\partial x} \longrightarrow ik_x, \quad \frac{\partial}{\partial y} \longrightarrow ik_y. \quad (12.26)$$

Thus, for plane waves (where the vector \mathbf{k} is well-defined), we can replace the derivatives according to these replacements, so that the solution is

$$\psi(z) = e^{-i(k_x^2 + k_y^2)z/2k} \psi(0). \quad (12.27)$$

When k_x and k_y are *not* well defined, they can be thought of as shorthands for the derivatives. Then changing back to the regular field,

$$E^{(+)}(z) = e^{ikz} e^{-i(k_x^2 + k_y^2)z/2k} E^{(+)}(0). \quad (12.28)$$

Thus, we recover the paraxial transfer function (12.18). Again for k_x and k_y to make sense as numbers, the electric field should be decomposed into its plane-wave components by a Fourier transform.

12.2.6 Nonparaxial Propagation and the Diffraction Limit

If the paraxial approximation isn't valid, we have to live with the square-root form of $k_z(k_x, k_y)$, and the general propagation problem isn't easy (although it is still straightforward on a computer). However, we can still get some insight by examining the propagation equation (12.16) a bit more.

Recall that through the Fourier transform, $E^{(+)}(x, y)$ is composed of different spatial frequencies. In particular, broad features in $E^{(+)}(x, y)$ are represented by small spatial frequencies, whereas fine details are represented by large spatial frequencies. Very fine details on spatial scales smaller than λ are represented by spatial frequencies $k_{x,y} > k$, which may seem strange in view of the form of Eq. (12.16). In fact, these high frequencies correspond to *evanescent waves* and don't propagate, as we can see from the form of the propagation factor:

$$e^{ik_z z} = e^{i\sqrt{k^2 - k_x^2 - k_y^2} z} = e^{-\sqrt{k_x^2 + k_y^2 - k^2} z} \quad (\text{if } k_x^2 + k_y^2 > k^2). \quad (12.29)$$

Here we have observed that $k^2 - k_x^2 + k_y^2 < 0$, and thus factored a minus sign out of the square root, producing a factor of i outside the (now real) square root. This produces either an exponentially damping solution, which is physically sensible (note that the backward version of this wave would be exponentially growing in z , which is only physically sensible if it propagates along $-z$). Thus, in this form, the wave damps exponentially with a length scale

$$\delta(k_x, k_y) = (k_x^2 + k_y^2 - k^2)^{-1/2}, \quad (12.30)$$

so that

$$e^{ik_z z} = e^{-z/\delta(k_x, k_y)} \quad (12.31)$$

for the evanescent-wave solutions.

This effect gives rise to a **diffraction limit** in imaging: propagation over long distances wipes out fine ($\text{sub-}\lambda$) details in the initial image $E^{(+)}(x, y)$, because the components that carry these details rapidly decay away. Note that this is still true in the ideal limit where all $k_{x,y}$ up to k can propagate, corresponding to propagation up to 180° from the optical axis; typical imaging systems aperture the light to a much smaller solid angle, wiping out yet more of the fine detail. In any case, this is why it is often said that it is impossible to optically resolve objects that are smaller than the optical wavelength λ .

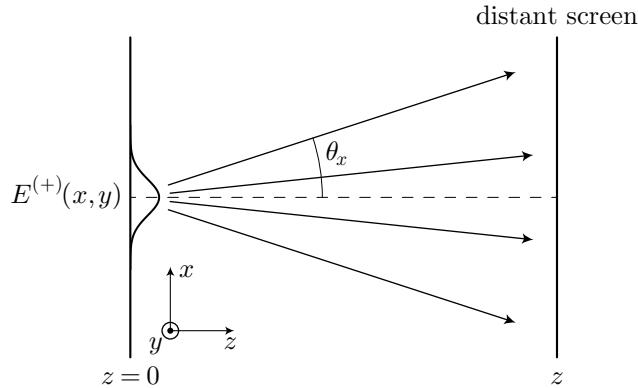
12.3 Fraunhofer Diffraction

12.3.1 Far-field Propagation

Fraunhofer diffraction refers to what happens to an optical pattern after propagation over a very large distance z in the paraxial approximation. Often, the “diffraction pattern” refers to the far-field pattern of an aperture (say, a slit or a pinhole) illuminated by a uniform, monochromatic field, such as an expanded laser beam (plane wave) at normal incidence. We will speak more generally here, however, and refer to the diffraction pattern of an arbitrary initial field $E^{(+)}(x, y)$ in the $z = 0$ plane.

It turns out that the Fraunhofer diffraction pattern is simple to calculate: with the proper scaling factors, *it's just the Fourier transform of the initial pattern*. We'll go through a heuristic derivation of this result here, and then do the mathematical version in the next section.

Consider the diffraction setup shown here.



The initial field $E^{(+)}(x, y)$ breaks up into its Fourier components. As we argued before in Section 12.2.1, each Fourier component corresponds to a wave propagating at a particular angle. In the paraxial approximation, we can write this out as

$$\theta_{x,y} \approx \frac{k_{x,y}}{k}. \quad (12.32)$$

Thus, we can write the Fourier transform of $E^{(+)}(x, y)$ as

$$\tilde{E}^{(+)}(k_x, k_y) = \tilde{E}^{(+)}(k\theta_x, k\theta_y). \quad (12.33)$$

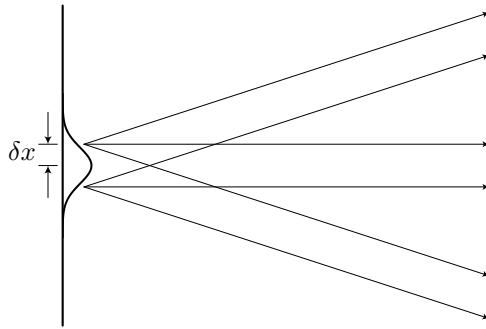
Noting that in the paraxial approximation, $\theta_x \approx x/z$ and $\theta_y \approx y/z$, we can write

$$E^{(+)}(x, y, z) = \frac{k}{2\pi z} \tilde{E}^{(+)}\left(\frac{kx}{z}, \frac{ky}{z}\right), \quad \text{where } \tilde{E}^{(+)}(k_x, k_y) = \mathcal{F}[E^{(+)}(x, y)],$$

(Fraunhofer diffraction pattern) (12.34)

where we have neglected an overall phase. The factor $k/2\pi z$ is less obvious here, but is what is required to make the total power of the diffracted wave equal to the power of the initial wave (in particular, the $1/z$ dependence enforces the inverse-square law). We will establish it more firmly in the next section. For now it is sufficient to realize that Eq. (12.34) gives the Fraunhofer diffraction pattern in terms of the Fourier transform, and in terms of angle the *relative* field components in terms of the angles $\theta_{x,y}$ are given by $\tilde{E}^{(+)}(k\theta_x, k\theta_y)$. While we are being fairly heuristic here, we will clean this up and be more formal when we redo this via Fresnel diffraction.

The Fraunhofer diffraction pattern is clearly an approximation to the “exact” pattern obtained by the propagation recipe. So when is this approximation valid? There are two conditions to be satisfied for the Fraunhofer approximation to hold. The first is that the size of the diffraction pattern must be much larger than the size of the initial aperture pattern $E^{(+)}(x, y)$.



We can view each component at angle θ as a “beam” with a size determined by the size of $E^{(+)}(x, y)$, rather than as a plane wave. Then each beam at each angle must separate spatially from the others by the time the field hits the screen; then we can treat each Fourier component as a separate “spot” on the screen. In other words, it must be that $E^{(+)}(x, y)$ looks like a point source to an observer on the screen. Letting δx_{source} denote the size (“radius”) of $E^{(+)}(x, y)$ in the x -direction (the same argument will hold for the y -direction), we can recall from the uncertainty relation (presented in temporal form in Eq. (7.36), but equally applicable to space) that the spatially confined field implies a spread in spatial frequency of

$$\delta k_x \sim \frac{1}{\delta x_{\text{source}}}. \quad (12.35)$$

Using Eq. (12.32), the uncertainty relation implies a maximum angle of

$$\max(\theta_x) \sim \frac{\lambda}{\delta x_{\text{source}}}. \quad (12.36)$$

An angle θ_x corresponds to a separation over distance z of $z\theta_x$, so for the size of the diffraction pattern to be much larger than the initial pattern, we require $\lambda z / \delta x_{\text{source}} \gg \delta x_{\text{source}}$, or

$$\delta x_{\text{source}} \ll \sqrt{\lambda z}. \quad (12.37)$$

(Fraunhofer small-source condition)

We can write this condition equivalently as

$$z \gg \frac{(\delta x_{\text{source}})^2}{\lambda}, \quad (12.38)$$

(Fraunhofer far-field condition)

and so this is the far-field requirement for the approximation to be valid.

One thing that we have neglected comes from the propagation phase factor

$$e^{ik_z z} \longrightarrow e^{ikz} \exp \left[-i \left(\frac{k_x^2 + k_y^2}{2k} \right) z \right] \quad (12.39)$$

that accompanies free-space propagation. We can drop the overall phase $\exp(ikz)$, but the other phase factor represents some phase variation across the screen. At the screen, this phase factor becomes

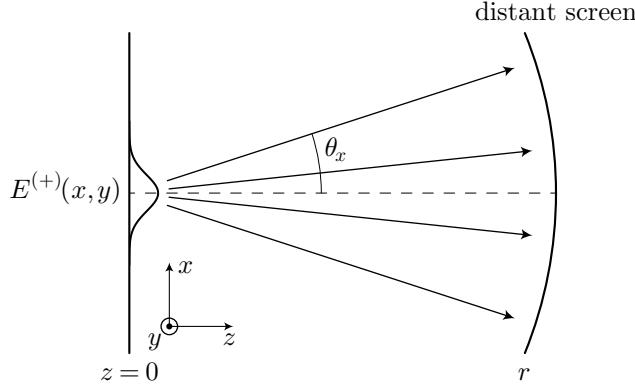
$$\exp \left(-ik \frac{x^2 + y^2}{2z} \right), \quad (12.40)$$

since we make the identification $k_x \rightarrow kx/z$ and $k_y \rightarrow ky/z$. This phase factor is in general not negligible, even in the paraxial regime. For example, to make a paraxial argument, we could consider only a region of “radius” δx_{screen} at the screen, where

$$\delta x_{\text{screen}} \ll \sqrt{\lambda z}. \quad (12.41)$$

This is essentially a small-angle restriction, more restrictive than the usual paraxial restriction, and because we require the diffraction pattern to be much larger than the source (which obeys the same condition),

this restriction would amount to looking only at the $(k_x, k_y) = (0, 0)$ region of the diffraction pattern. To understand this phase factor, note that we considered each Fourier component propagating away from the source at different angles. For all of them to propagate over the same distance r , we would have to consider a spherical screen of radius r , a distance r from the source.



Thus, we can correctly write

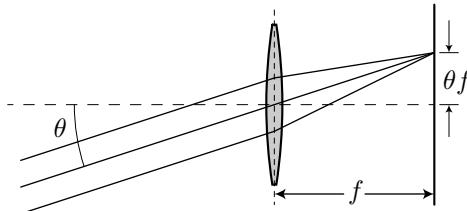
$$E^{(+)}(x, y, r) = \frac{k}{2\pi r} \tilde{E}^{(+)}\left(\frac{kx}{r}, \frac{ky}{r}\right), \quad \text{where } \tilde{E}^{(+)}(k_x, k_y) = \mathcal{F}[E^{(+)}(x, y)], \quad r = \sqrt{x^2 + y^2 + z^2}$$

(Fraunhofer diffraction pattern) (12.42)

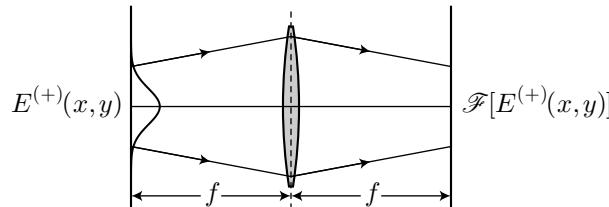
for the diffraction pattern at constant r , whereas the expression (12.34) should include the phase factor (12.40). The phase factor here is the spherical (parabolic in the paraxial approximation) phase correction for continuing this field onto a flat screen. When computing the screen intensity, this phase factor vanishes anyway, but strictly speaking the far-field pattern is only the Fourier transform without this factor on a spherical screen.

12.3.2 Thin Lens as a Fourier Transform Computer

The whole Fourier-transform argument above also works for a thin lens. Recall that for a thin lens, parallel rays propagating at an angle θ with respect to the optical axis get mapped to a transverse position θf in the focal plane of the lens.



With this fact in mind, let's consider the setup shown here, where a field profile is followed by a thin lens placed one focal length f away, and a screen is another focal length past the lens.



We can now make a few observations:

1. The lens images $E^{(+)}(x, y)$ to $z = \infty$.

2. Plane waves at different angles $\theta_{x,y}$ get mapped to different points $\theta_{x,y}f$ in the focal plane (on the screen). This is the wave-optics version of the parallel-ray argument above.
3. The lens maps the image at $z = -\infty$ to the focal plane.
4. Because an image at infinity emits parallel rays, we can identify the images at $z = \pm\infty$ as being equivalent. The field at $z = -\infty$ is the Fraunhofer diffraction pattern of $E^{(+)}(x, y)$ because the image at $z = \infty$ *in the absence of the lens* is the Fraunhofer pattern of the original image.
5. The lens then maps the field at $-\infty$ to the focal plane.

Thus, we can conclude that the image at the focal plane of the lens (i.e., the screen) is the Fourier transform of $E^{(+)}(x, y)$, but where f is now the propagation length scale:

$$E_{\text{focalplane}}^{(+)}(x, y) = \frac{k}{2\pi f} \tilde{E}^{(+)}\left(\frac{kx}{f}, \frac{ky}{f}\right). \quad (12.43)$$

(Fourier-transform via lens)

Thus, at least within the paraxial approximation, **thin lenses act as analog Fourier transform computers**. Note that the phase factor (12.40) is negligible here, because in the paraxial limit, both the source and image fields are narrowly confined near the optical axis. In the Fraunhofer-diffraction case, the image was large due to the far-field requirement, but the lens takes care of compressing the image down to a small size here.

12.3.3 Example: Diffraction from a Double Slit

As an easy example, let's compute the Fraunhofer diffraction pattern due to two narrow slits separated by distance a . The slits are uniformly illuminated by a monochromatic plane wave at normal incidence to the mask. Assuming the slit widths are much smaller than a , we can model the slits as δ -functions:

$$E^{(+)}(x) = A [\delta(x - a/2) + \delta(x + a/2)]. \quad (12.44)$$

The Fourier transform is given by

$$\begin{aligned} \tilde{E}^{(+)}(k_x) &= A \int_{-\infty}^{\infty} e^{-ik_x x} [\delta(x - a/2) + \delta(x + a/2)] dx \\ &= A [e^{ik_x a/2} + e^{-ik_x a/2}] \\ &= 2A \cos\left(\frac{k_x a}{2}\right). \end{aligned} \quad (12.45)$$

Thus, the far-field diffraction pattern at distance z is

$$E^{(+)}(x, z) = \sqrt{\frac{k}{2\pi z}} \tilde{E}^{(+)}\left(\frac{kx}{z}\right) = \sqrt{\frac{2k}{\pi z}} A \cos\left(\frac{kax}{2z}\right) = \frac{2A}{\sqrt{\lambda z}} \cos\left(\frac{\pi ax}{\lambda z}\right). \quad (\text{double-slit diffraction pattern}) \quad (12.46)$$

Note that we are using $\sqrt{k/2\pi z}$ as the normalization factor, rather than $k/2\pi z$, because this is a one-dimensional transform (we associate half of the full factor with each dimension—note that the units come out correctly this way). In terms of angle, the diffraction pattern is (up to an overall factor)

$$\tilde{E}^{(+)}(k\theta_x) = 2A \cos\left(\frac{ka\theta_x}{2}\right) = 2A \cos\left(\frac{\pi a\theta_x}{\lambda}\right). \quad (\text{double-slit angular diffraction pattern}) \quad (12.47)$$

This function vanishes when the argument is $\pm\pi/2, \pm3\pi/2$, and so on, which corresponds to angles θ_x of $\pm\lambda/2a, \pm3\lambda/2a$, and so on. Thus, the angular separation between fringes is $\Delta\theta = \lambda/a$. This result is consistent with the result of a simple relative-path-length calculation (Problem 5.1).

12.3.4 Example: Diffraction from a Sinusoidal Intensity-Mask Grating

Another simple example is the diffraction of a plane wave by a sinusoidal intensity mask (or *intensity grating*). If the intensity mask has a transmission function $t_{\text{mask}} = \cos^2 gx$, and is illuminated by a normally incident plane wave $E_0^{(+)}$, then the field is

$$E^{(+)}(x) = t_{\text{mask}} E_0^{(+)} = E_0^{(+)} \cos^2 gx = \frac{E_0^{(+)}}{2} + \frac{E_0^{(+)}}{4} \cos(2gx) = \frac{E_0^{(+)}}{2} + \frac{E_0^{(+)}}{4} [e^{i2gx} + e^{-i2gx}]. \quad (12.48)$$

Using the Fourier-transform relation

$$\mathcal{F}[e^{i\alpha x}] = \int_{-\infty}^{\infty} e^{-ix(k_x - \alpha)} dx = 2\pi\delta(k_x - \alpha), \quad (12.49)$$

we can easily compute the Fourier transform of Eq. (12.48):

$$\mathcal{F}[E^{(+)}(x)] = \tilde{E}^{(+)}(k_x) = \pi E_0^{(+)} \delta(k_x) + \frac{\pi E_0^{(+)}}{2} [\delta(k_x - 2g) + \delta(k_x + 2g)]. \quad (12.50)$$

Thus, the angular diffraction pattern is

$$\tilde{E}^{(+)}(k\theta_x) = \pi E_0^{(+)} \delta(k\theta_x) + \frac{\pi E_0^{(+)}}{2} [\delta(k\theta_x - 2g) + \delta(k\theta_x + 2g)]. \quad (12.51)$$

Using $\delta(ax) = \delta(x)/|a|$, this becomes

$$\tilde{E}^{(+)}(k\theta_x) = \frac{\pi E_0^{(+)}}{k} \delta(\theta_x) + \frac{\pi E_0^{(+)}}{2k} [\delta(\theta_x - 2g/k) + \delta(\theta_x + 2g/k)]. \quad (\text{angular diffraction pattern of sinusoidal grating}) \quad (12.52)$$

Thus, there are diffraction peaks at angles $\theta_x = 0$ and $\pm 2g/k = \pm g\lambda/\pi$. There are only 3 peaks, so this grating is special because it's sinusoidal: there are no higher-order diffraction peaks from this grating (we'll say more about this in a bit). It's important to see the Fourier-transform behavior at work here again—as the period of the grating gets longer, g decreases, and so the diffraction pattern narrows, as you'd expect from the uncertainty principle.

Briefly, the corresponding pattern on a distant screen is

$$E^{(+)}(x, z) = \frac{k}{2\pi z} \tilde{E}^{(+)}(kx/z) = \frac{E_0^{(+)}}{2} \delta(x) + \frac{E_0^{(+)}}{4} [\delta(x - 2gz/k) + \delta(x + 2gz/k)] \quad (12.53)$$

when the screen is at distance z .

12.3.5 Example: Diffraction from an Arbitrary Grating

For a *general* periodic transmission mask of period π/g , with a normally incident plane wave, we can repeat the above analysis by decomposing the function into its Fourier coefficients:

$$E^{(+)}(x) = \sum_{n=-\infty}^{\infty} E_0^{(+)} c_n e^{i2gnx}, \quad (12.54)$$

where the c_n are the Fourier coefficients of the transmission function. Then the Fourier transform is

$$\tilde{E}^{(+)}(k_x) = 2\pi E_0^{(+)} \sum_{n=-\infty}^{\infty} c_n \delta(k_x - 2gn), \quad (12.55)$$

so the angular diffraction pattern is

$$\tilde{E}^{(+)}(k\theta_x) = \frac{2\pi E_0^{(+)}}{k} \sum_{n=-\infty}^{\infty} c_n \delta(\theta_x - 2gn/k). \quad (\text{angular diffraction pattern of arbitrary grating}) \quad (12.56)$$

Thus, for an arbitrary pattern, there are many diffraction peaks, with the **diffraction peak of order n** located at

$$\theta_n = \frac{2gn}{k}, \quad (12.57)$$

(angle of n th diffraction order)

and this peak has amplitude c_n (or relative intensity $|c_n|^2$).

12.4 Fresnel Diffraction

12.4.1 Convolution Revisited

If we stay in the paraxial approximation, but we don't require that the observation screen is in the far field, then we have the problem of **Fresnel diffraction**. Actually, Fresnel diffraction is also used to refer to the case where the wavefronts at $z = 0$ are not planar, but this is already accounted for in the Fraunhofer formalism that we presented in the last section, as it only modifies the particular form of the initial function $E^{(+)}(x, y)$. Thus, we will be concerned with the propagation of the optical wave to *any* distance, not just the far field.

Actually we were already doing this when we introduced the Fourier recipe for wave propagation from Section 12.2.4. Recalling this procedure, we start with a field $E^{(+)}(x, y)$ at $z = 0$. Again, this is completely general, so this could be an arbitrary wave $E_{\text{in}}^{(+)}(x, y)$ incident on a phase or amplitude mask (aperture) with transmission t_{aperture} , so that

$$E^{(+)}(x, y) = t_{\text{aperture}} E_{\text{in}}^{(+)}(x, y). \quad (12.58)$$

Then the *frequency-domain solution* (Eq. (12.16)) is

$$E^{(+)}(x, y, z) = \mathcal{F}^{-1} \left[\mathcal{F} [E^{(+)}(x, y)] e^{ik_z z} \right]. \quad (12.59)$$

In the paraxial approximation, but not necessarily in the far field, this becomes

$$E^{(+)}(x, y, z) = \mathcal{F}^{-1} \left[\mathcal{F} [E^{(+)}(x, y)] e^{ikz} e^{-i(k_x^2 + k_y^2)z/2k} \right]. \quad (12.60)$$

Now we'll use the identity

$$\mathcal{F}^{-1} \left[e^{-i(k_x^2 + k_y^2)z/2k} \right] = \frac{k}{2\pi iz} e^{ik(x^2 + y^2)/2z}, \quad (12.61)$$

which follows from the Gaussian Fourier-transform formula (see Section 3.2.1), but with a little care. Since the exponent is purely imaginary, the transform integral doesn't actually converge. But letting $z \rightarrow z - i\beta$, where $\beta > 0$, we can do the resulting convergent integral and then let $\beta \rightarrow 0$ afterwards (this is saying that arbitrarily high spatial frequencies, corresponding to arbitrary fine detail, don't physically contribute). So, using Eq. (12.61) in Eq. (12.60), we can write

$$E^{(+)}(x, y, z) = \mathcal{F}^{-1} \left[\mathcal{F} [E^{(+)}(x, y)] \mathcal{F} \left[\frac{ke^{ikz}}{2\pi iz} e^{ik(x^2 + y^2)/2z} \right] \right]. \quad (12.62)$$

By the convolution theorem,

$$\begin{aligned} E^{(+)}(x, y, z) &= E^{(+)}(x, y) * g(x, y; z) \\ &= \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E^{(+)}(x', y') g(x - x', y - y'; z), \end{aligned} \quad (\text{paraxial propagation as convolution}) \quad (12.63)$$

where

$$g(x, y; z) := \frac{ke^{ikz}}{2\pi iz} e^{ik(x^2 + y^2)/2z} = \frac{e^{ikz}}{i\lambda z} e^{ik(x^2 + y^2)/2z}. \quad (\text{Green function for paraxial propagation}) \quad (12.64)$$

This is the *convolution* or *space-domain* solution to the paraxial-propagation or Fresnel-diffraction problem.

12.4.2 Paraxial Impulse Response

So what exactly is this convolution kernel $g(x, y)$? Suppose we have an optical source that is a unit point, $E^{(+)}(x, y, z = 0) \rightarrow \delta(x, y)$, and let's propagate it forward. The Fourier transform is easy:

$$\mathcal{F}[\delta(x, y)] = 1. \quad (12.65)$$

Then multiplying by the propagation factor $\exp(ikz) = \exp(ikz) \exp[-i(k_x^2 + k_y^2)z/2k]$ and inverting the transform, the propagation induces the transformation

$$\delta(x, y) \longrightarrow \frac{ke^{ikz}}{2\pi iz} e^{ik(x^2+y^2)/2z} = g(x, y; z).$$

(paraxial propagation of point source) (12.66)

Thus, $g(x, y)$ is the *impulse-response function* or Green function for the diffraction problem, and Eq. (12.63) is the Green function solution to the diffraction problem (in the paraxial approximation).

12.4.3 Far-Field (Fraunhofer) Limit

In the far field, we should be able to recover the Fraunhofer diffraction pattern and thus rigorously derive the results of Section 12.3. Let's start by writing out the Green function solution (12.63) explicitly:

$$E^{(+)}(x, y, z) = \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E^{(+)}(x', y') \frac{ke^{ikz}}{2\pi iz} e^{ik[(x-x')^2+(y-y')^2]/2z}. \quad (12.67)$$

We can assume the Fraunhofer limit $\delta x_{\text{source}} \ll \sqrt{\lambda z}$ (far field/small source), so that the phase factor $\exp[ik(x'^2 + y'^2)/2z]$ is approximately unity (i.e., neglecting x'^2 compared to x^2), so Eq. (12.67) becomes

$$E^{(+)}(x, y, z) = \frac{ke^{ikz}}{2\pi iz} e^{ik(x^2+y^2)/2z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E^{(+)}(x', y') \exp\left[-i\left(\frac{kx}{z}\right)x'\right] \exp\left[-i\left(\frac{ky}{z}\right)y'\right]. \quad (12.68)$$

Now defining $\kappa_x := kx/z$ and $\kappa_y := ky/z$, we can write

$$E^{(+)}(x, y, z) = \frac{ke^{ikz}}{2\pi iz} e^{ik(x^2+y^2)/2z} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E^{(+)}(x', y') e^{i(\kappa_x x + \kappa_y y)}. \quad (\text{Fraunhofer diffraction pattern}) \quad (12.69)$$

This is just the Fourier transform, so

$$E^{(+)}(x, y, z) = \frac{ke^{ikz}}{2\pi iz} e^{ik(x^2+y^2)/2z} \tilde{E}^{(+)}(\kappa_x, \kappa_y) = \frac{ke^{ikz}}{2\pi iz} e^{ik(x^2+y^2)/2z} \tilde{E}^{(+)}\left(\frac{kx}{z}, \frac{ky}{z}\right). \quad (\text{Fraunhofer diffraction pattern}) \quad (12.70)$$

Thus, we recover the Fourier transform form for the Fraunhofer diffraction pattern, including the correct scaling factor and the phase factors that we merely indicated in our heuristic treatment. This includes the phase factor $\exp[ik(x^2 + y^2)/2z]$ for projecting a spherical wave front onto the flat screen at constant z . We can see this directly by noting that in the paraxial approximation,

$$z = \sqrt{r^2 - x^2 - y^2} \approx r - \frac{x^2 + y^2}{2z}, \quad (12.71)$$

so that upon making this replacement in the $\exp(ikz)$ factor, the diffraction pattern becomes

$$E^{(+)}(x, y, r) = \frac{ke^{ikr}}{2\pi ir} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E^{(+)}(x', y') e^{i(\kappa_x x + \kappa_y y)}, \quad (\text{Fraunhofer diffraction pattern, constant } r) \quad (12.72)$$

where we have also replaced $z \approx r$ in the denominator. In this form, the extra phase factor is gone when the diffraction pattern is viewed on a spherical screen at constant r .

12.4.4 Example: Fresnel Diffraction from a Slit

As one of the simplest examples, let's consider the Fresnel diffraction pattern from a single slit of width a . In this case, Eq. (12.63) becomes

$$E^{(+)}(x, z) = E_0^{(+)} \sqrt{\frac{k}{2\pi iz}} e^{ikz} \int_{-a/2}^{a/2} e^{ik(x-x')^2/2z} dx' , \quad (12.73)$$

where we assume the y' -integration has already been completed, and we've also suppressed the y -dependence. Again, in suppressing the y -dependence, note that we are also suppressing half of the normalization factor $k/2\pi iz$ from the 2D Gaussian transform (12.61)—the y profile of the beam will simply remain a plane wave, not diffracting. Letting $x' \rightarrow x' + x$,

$$E^{(+)}(x, z) = E_0^{(+)} \sqrt{\frac{k}{2\pi iz}} e^{ikz} \int_{-x-a/2}^{-x+a/2} e^{ikx'^2/2z} dx' . \quad (12.74)$$

Now letting $x' \rightarrow \sqrt{\pi z/k} x'$,

$$E^{(+)}(x, z) = E_0^{(+)} \sqrt{\frac{-i}{2}} e^{ikz} \int_{\sqrt{k/\pi z}(-x-a/2)}^{\sqrt{k/\pi z}(-x+a/2)} e^{i\pi x'^2/2} dx' . \quad (12.75)$$

Now defining the **Fresnel integrals**

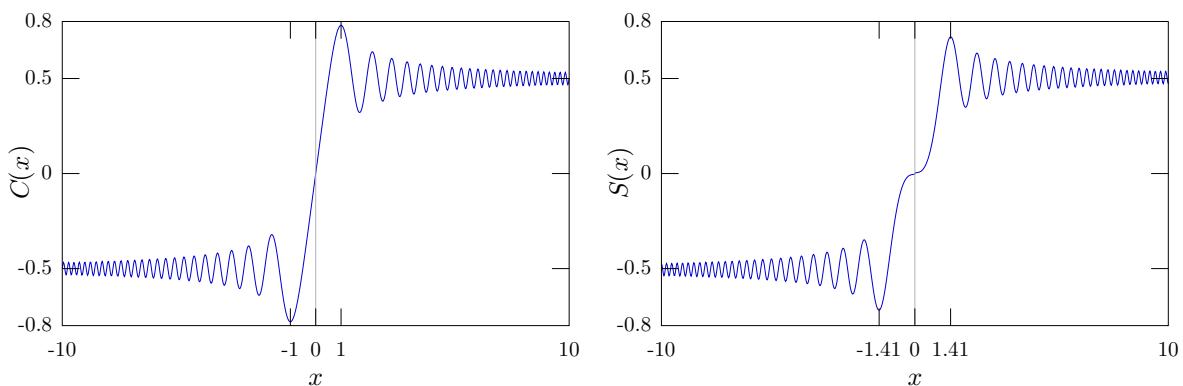
$$C(x) := \int_0^x \cos\left(\frac{\pi t^2}{2}\right) dt \quad S(x) := \int_0^x \sin\left(\frac{\pi t^2}{2}\right) dt , \quad (12.76) \quad (\text{Fresnel integrals})$$

we can write the diffraction pattern in standard form as

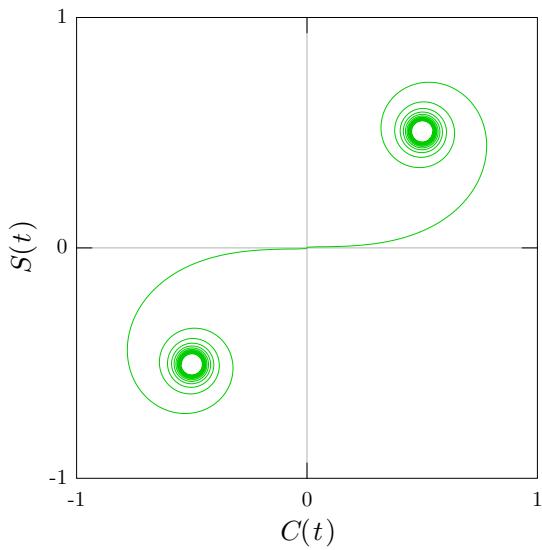
$$\begin{aligned} E^{(+)}(x, z) = E_0^{(+)} \sqrt{\frac{-i}{2}} e^{ikz} & \left\{ C\left[\sqrt{\frac{k}{\pi z}}\left(x + \frac{a}{2}\right)\right] - C\left[\sqrt{\frac{k}{\pi z}}\left(x - \frac{a}{2}\right)\right] \right. \\ & \left. + iS\left[\sqrt{\frac{k}{\pi z}}\left(x + \frac{a}{2}\right)\right] - iS\left[\sqrt{\frac{k}{\pi z}}\left(x - \frac{a}{2}\right)\right] \right\} , \end{aligned} \quad (\text{Fresnel diffraction by a slit}) \quad (12.77)$$

where we have used the fact that $C(x)$ and $S(x)$ are odd functions, because they are both integrals from 0 to x of an even integrand. The Fresnel integrals are easily evaluated on a computer these days.

The Fresnel integrals are plotted here. Note their oscillatory nature, which ultimately gives rise to the fringes in the diffraction patterns, and that they asymptotically settle to $\pm 1/2$.

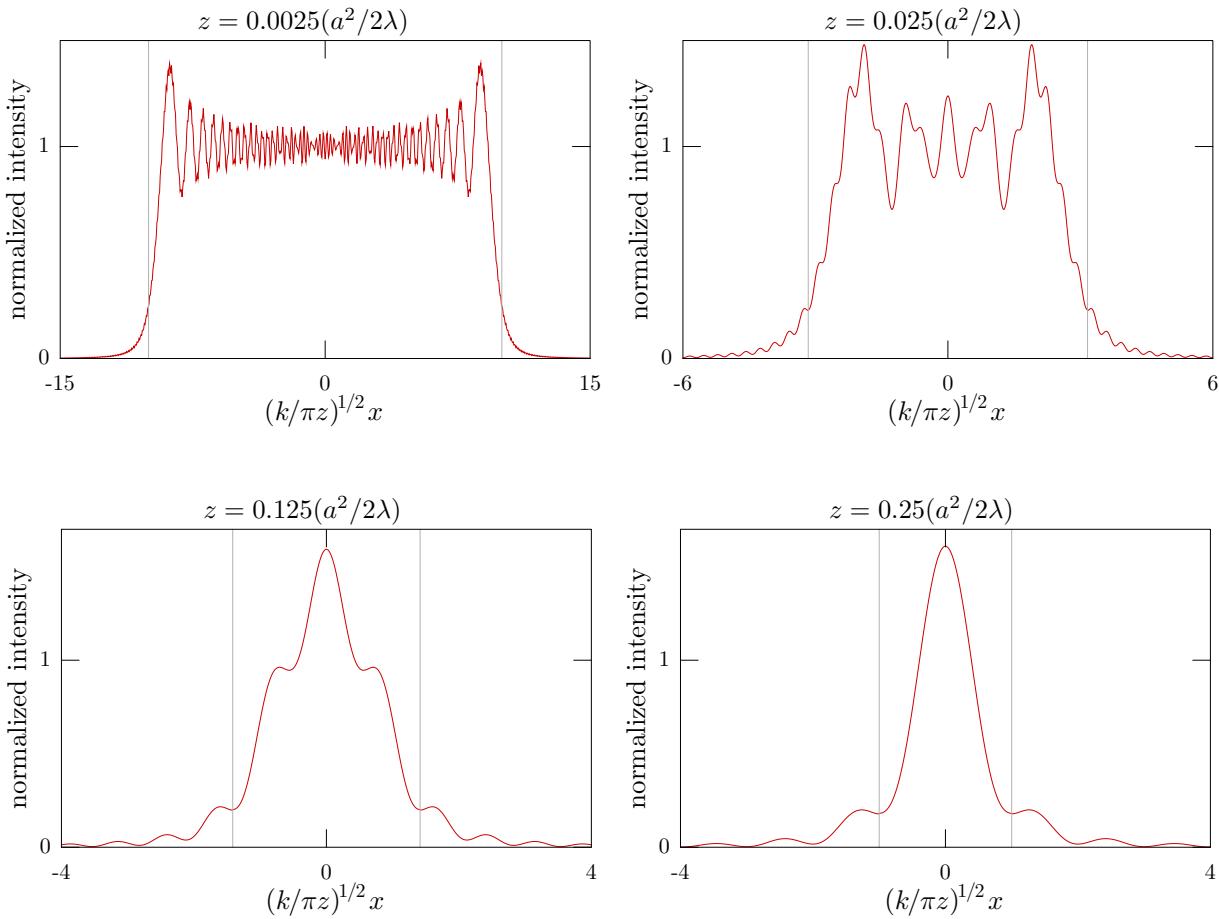


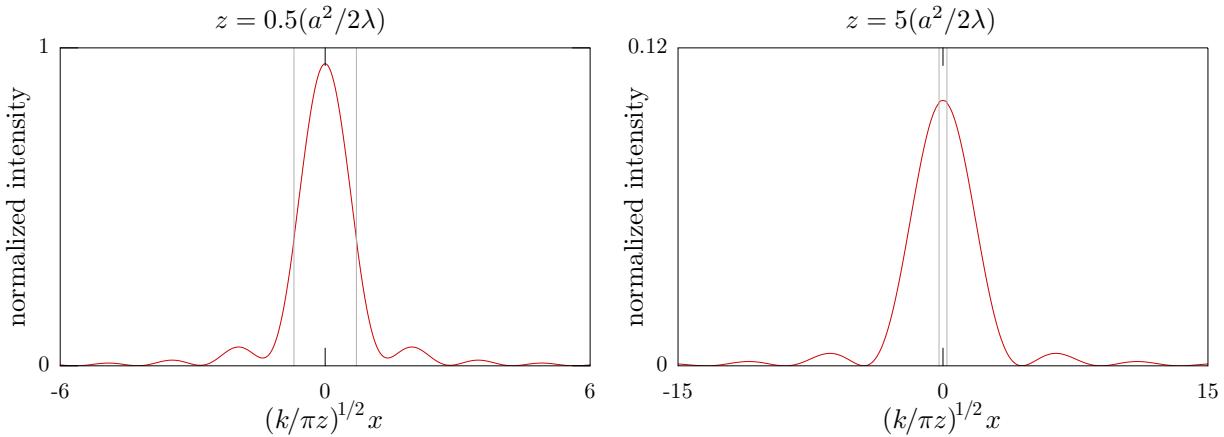
The Fresnel integrals are famously plotted as the “Cornu spiral,” which is a parametric plot of $S(t)$ vs. $C(t)$.



The Fresnel integrals can be evaluated graphically using the Cornu spiral, but their evaluation much easier on a computer.

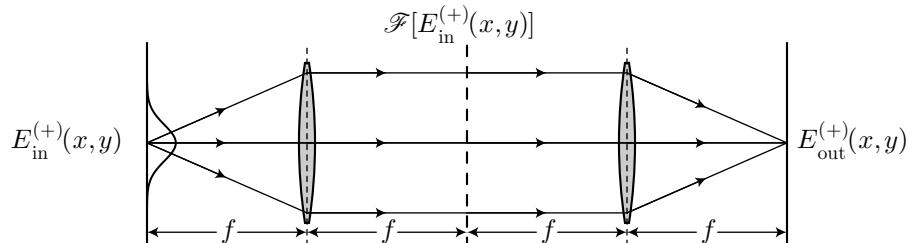
Some diffraction patterns for the single slit are shown here. The slit edges are marked by vertical lines. Notice the gradual transition from something rectangular to the familiar far-field sinc pattern as the distance increases.





12.5 Spatial Filters

The fact that a thin lens computes the Fourier transform of an optical field profile opens up a lot of intriguing possibilities—namely, that of *optical information processing*. The basic setup for doing this is a system of two lenses. For simplicity we'll assume that the two lenses have the same focal length f . Then the propagation is over a distance $4f$ as shown here.



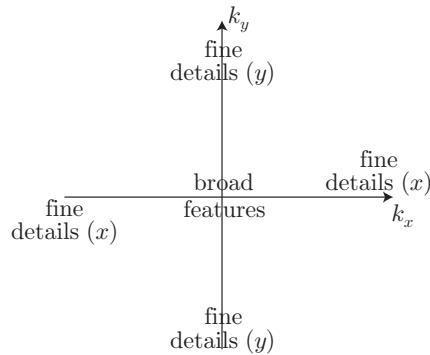
For this reason this is called the **4-f imaging system** for spatially filtering an image. The Fourier transform of the original field profile appears in the plane midway between the two lenses. The second lens undoes the Fourier transform and reconstructs the original image on the output plane. (Technically, the second lens performs a second Fourier transform, which is equivalent to the inverse transform with $x \rightarrow -x$ and $y \rightarrow -y$, so the final image is inverted. We'll ignore this in the discussions here.)

The point of this setup is that we can put in filters to modify the Fourier transform of the image, in either the amplitude or phase. In doing so, we can perform many useful enhancements or modifications to the image. The aperture in the Fourier plane defines the *transfer function* for the system.

Recall that in the original wave is $E_{in}^{(+)}(x, y)$, then in the Fourier-transform plane, the field is given in terms of the Fourier transform $\tilde{E}_{in}^{(+)}(k_x, k_y)$ by

$$E_{\text{focalplane}}^{(+)}(x, y) = \frac{k}{2\pi f} \tilde{E}_{in}^{(+)}\left(\frac{kx}{f}, \frac{ky}{f}\right). \quad (12.78)$$

The spatial locations in the Fourier plane correspond to different spatial frequencies and thus to details in the images of different length scales.



The low frequencies (and thus broad image features) are near the center of the transform plane, with the dc component being right in the center. The fine details are encoded in the high frequencies away from the center. This observation motivates some of the simplest spatial filters.

1. **Low-pass filter.** A circular aperture in the Fourier plane will block the high spatial frequencies and pass the low ones. The effect is to remove the fine details, blurring the image. If the filter is an aperture of diameter a , the cutoff frequency is given by the condition $k_r = kr/f$, giving a cutoff frequency of

$$k_{\text{cutoff}} = \frac{ka}{2f}. \quad (12.79)$$

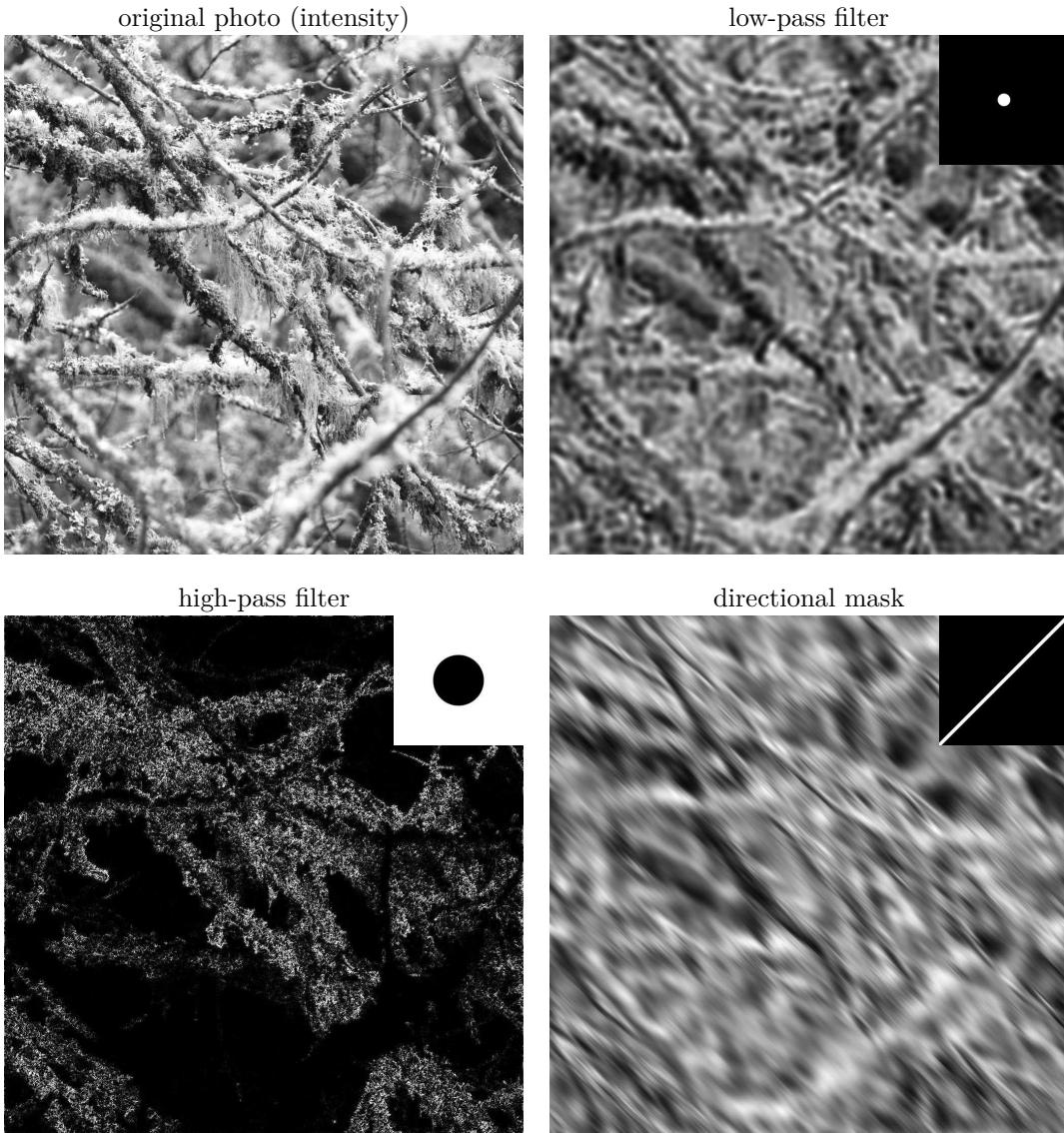
That is, structures smaller than the length scale of

$$\lambda_{\text{cutoff}} = \frac{2\pi}{k_{\text{cutoff}}} = \frac{4\pi f}{ka} = \frac{2\lambda f}{a} \quad (12.80)$$

are removed from the image. From the convolution theorem, as we'll come back to later, the output image is the convolution of the original image with the Fourier transform of the aperture. The aperture's Fourier transform is the Airy disk, and so the image is "smeared" by the Airy disk, causing the blurring.

2. **High-pass filter.** A circular, absorbing spot in the Fourier does the opposite, blocking the low frequencies and passing the high. For a spot of diameter a , the cutoff is the same as for the low-pass filter. Thus, the image filter keeps only those structures smaller than $2\lambda f/a$. This filter is also called an *edge detector* because sharp edges generate high spatial frequencies and thus make it through the filter.
3. **Directional filter.** The high- and low-pass filters can also work only in one direction, with the obvious modification of the aperture. For example, a vertical slit preferentially passes vertical structures while blocking horizontal structures.

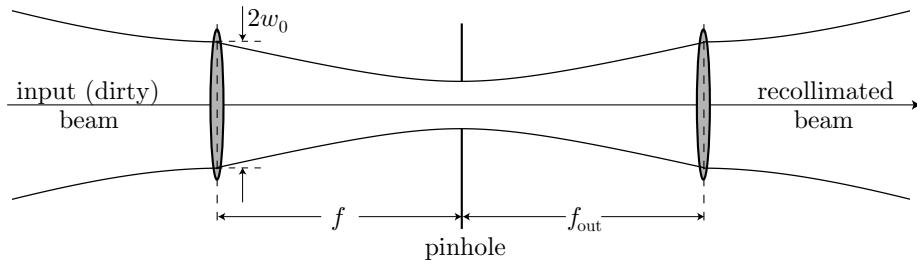
Examples of these filters are shown here in simulations. The original photo (of Oregon moss on trees) has structures on many different scales. The low-pass filter wipes out the fine details of the moss, leaving only the branches. The high-pass filter does just the opposite: the structure of the branches is mostly wiped out, while the moss detail is preserved. The directional mask in the final photo is oriented at 45° ; notice that only the branches *perpendicular* to the slit survive the filter, while the parallel branches are mostly erased. That's because branches parallel to the slit represent high spatial frequencies perpendicular to the slit, which get wiped out.



12.5.1 Spatial Filtering of a Gaussian Beam

One important application of spatial filtering in real optical setups is the cleaning of a Gaussian beam. When propagating through real optical elements, Gaussian beams get contaminated with spatial intensity noise due to dust, imperfections, diffraction at the edges of objects, etalon fringes (when, say, passing through an uncoated window), aberrations, and so on. For example, an acousto-optic modulator acting as a switch tends to distort the beam profile, and so it is often desirable to restore it. The idea is to use a low-pass filter, here in the form of a pinhole. The Gaussian beam is focused through the pinhole, and under the right conditions a nearly ideal Gaussian profile appears on the other side.

Let's simplify things from the basic $4-f$ setup. We'll use a lens of focal length f to focus the beam through the pinhole, and let's assume that the waist of the unfocused beam occurs at the lens. The pinhole is much smaller than the unfocused beam, so the pinhole is a distance f from the pinhole. After the pinhole, a second lens of focal length f_{out} recollimates the beam.



In the absence of the pinhole, this is simply a telescope beam expander, and so the diameter of the beam is magnified by a factor f_{out}/f . The most important part of the setup is the pinhole, so let's focus on that.

Let the incident beam have radius w_0 . Then we can use the results in Section 6.5 from the *ABCD* Law, which gives the beam radius at the pinhole:

$$w_{0,\text{pinhole}} = \frac{\lambda f}{\pi w_0}. \quad (12.81)$$

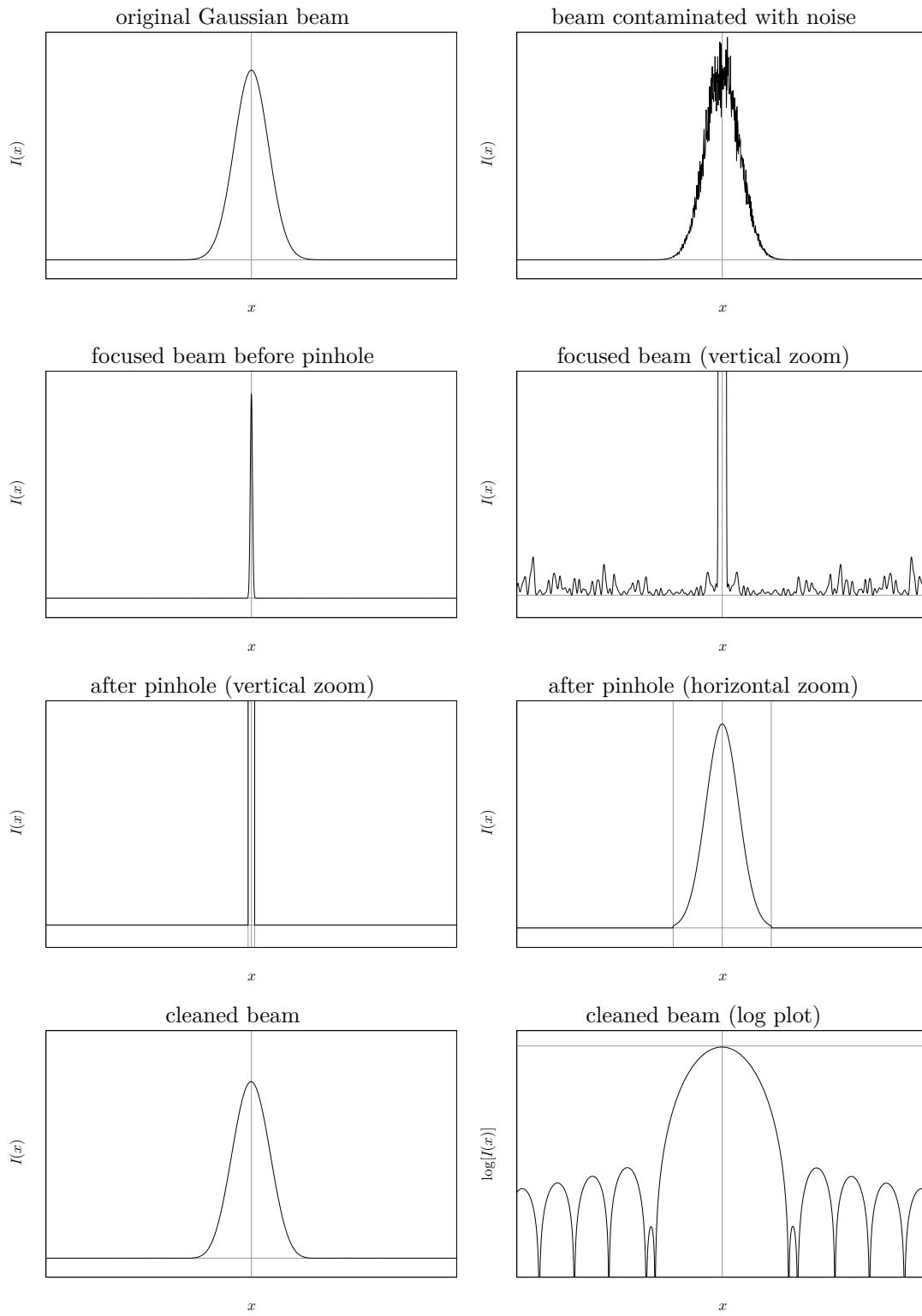
This assumes, of course, a *clean* Gaussian beam. The commonly recommended pinhole diameter is one that passes 99% of the incident power of a clean beam, which is

$$D_{\text{pinhole}} = 2 \cdot 1.5 \cdot w_{0,\text{pinhole}}. \quad (12.82)$$

In this case, nearly all of the Gaussian beam passes through the pinhole. But noise, especially high-frequency noise, has its power distributed away from the pinhole in the Fourier-transform plane. So the pinhole blocks the power associated with spatial intensity noise on the beam. Thus, the recollimated beam is a very clean Gaussian to good approximation.

There are two caveats here. One is that a small amount of intensity noise is left, corresponding to the same length scales as the beam diameter. This is usually not noticeable, and corresponds to shifts or slight distortion of the profile. In cases where there is significant low-frequency noise (such as the highly distorted output of a single-mode diode laser), it may be necessary to use a smaller pinhole and lose much more power. The other caveat is that the output isn't quite Gaussian. From the convolution theorem, the output is the convolution of a Gaussian and the Airy disk, so there are generally faint rings around a spatially filtered beam. These can be clipped off if necessary with an aperture after the recollimating lens.

All this information is best visualized in the simulated example shown below. The first plot shows the intensity profile of a clean Gaussian beam. The second is the same profile, but with intensity noise superposed. Now the task is to clean this up as much as possible. Focusing the beam produces a very narrow Gaussian profile, shown in the next plot. The Gaussian itself is rather clean, but what is only apparent with a large vertical magnification (fourth plot) is that the baseline is contaminated with broadly distributed noise. This is what we expect from the Fourier transform argument above. We clip this noise away with pinhole (fifth and sixth plots, edges of pinhole marked by vertical lines). The very tips of the Gaussian profile are clipped as well, but essentially all of the noise is gone. The beam on the output is a nice-looking Gaussian (seventh plot), but notice that the amplitude is reduced compared to the first plot: power is lost in the cleaning process, since we are essentially projecting the beam into the Gaussian mode. The faint Airy rings are not visible on this plot, but they are on a logarithmic plot (last plot). The side lobes in this plot do not occur on the original beam. These are visible to the human eye, which itself has a logarithmic response to intensity.



12.5.2 Visualization of Phase Objects

The “obvious” spatial filters in the beginning of this section (low-pass, high-pass, directional) are intuitive examples of what you can do with optical spatial filters. But these days, it seems far easier just to digitize the image and do the same processing on a computer. So why do optical information processing?

Sometimes it is convenient or useful to do so optically, as in cleaning a Gaussian beam. But sometimes it is impossible to digitize or even record optical information before optical processing. An important class of examples is in imaging **phase objects**, or objects that only influence the phase of probe light. We can model a phase object by a transmission coefficient of the form

$$t(x, y) \sim e^{i\phi(x, y)}. \quad (12.83)$$

If this object is illuminated by a plane wave $E_0^{(+)}$ (in the $z = 0$ plane), the resulting field becomes

$$E_0^{(+)} \rightarrow E_0^{(+)} t(x, y) \sim E_0^{(+)} e^{i\phi(x, y)}. \quad (12.84)$$

The intensity is

$$I(x, y) \sim \frac{2|E_0^{(+)}|^2}{\eta}, \quad (12.85)$$

which is a constant and thus has no information about the phase $\phi(x, y)$. The point is that the transmission function has unit amplitude everywhere, and so doesn’t modify the intensity of the transmitted light. So we can’t just record the light on film and post-process the information. We could get at this information via interferometry or holographic techniques. Instead, we will look at how spatial filters can reveal this information.

12.5.2.1 Zernike Phase-Contrast Imaging

One widely used phase-imaging method was developed by Frits Zernike, who won the 1953 Nobel Prize in Physics for this work. This method is commonly used, for example in biological microscopy to view transparent objects such as bacteria (that would otherwise require staining).

Let’s consider a phase object with a small phase shift of the form

$$t(x, y) = e^{i\phi(x, y)} \approx 1 + i\phi(x, y) + O(\phi^2). \quad (12.86)$$

As we saw before, when uniformly illuminated, the intensity transmitted is

$$I \sim |1 + i\phi(x, y)|^2 = 1 + O(\phi^2), \quad (12.87)$$

so that there is no intensity change (we are discarding the higher order terms in ϕ to be consistent with the initial expansion). Essentially, this is because the linear cross-terms cancel when squaring the field.

The idea behind the Zernike method starts with the observation that the unity part of Eq. (12.86) is the *dc component* of the signal, while $\phi(x, y)$ represents spatial structure, and thus has a wideband spectrum. The vanishing cross-terms are thus a product of the dc component and the rest of the spectrum. The idea is, what if we modify one of these to prevent the cancellation? Specifically, let’s try a $\pi/2$ phase shift of the dc component:

$$I \sim \left| e^{i\pi/2} + i\phi(x, y) \right|^2 = |i[1 + \phi(x, y)]|^2 = 1 + 2\phi(x, y) + O(\phi^2). \quad (12.88)$$

Now the linear term stays, so the intensity reflects the phase information. Actually, since the intensity change is *linear* in the phase (for small phase shifts), this method gives an image that is simple to interpret: other than an overall background, the intensity is a direct image of the phase. The other imaging methods we will discuss here produce more complicated functions of the phase.

Since the intensity increases with increasing phase shift, this method is called **positive phase contrast imaging**. We can also try a $3\pi/2$ phase shift of the dc component, with the result

$$I \sim \left| e^{i3\pi/2} + i\phi(x, y) \right|^2 = |(-i)[1 - \phi(x, y)]|^2 = 1 - 2\phi(x, y) + O(\phi^2). \quad (12.89)$$

Now the intensity *decreases* with increasing phase, so this version is called **negative phase contrast imaging**.

The requisite phase shift is accomplished here by a filter in the Fourier plane of the 4-*f* system that could be, for example, a glass plate with a transparent dielectric thin-film dot at the center that causes a relative phase shift of $\pm\pi/2$ of the small portion at the center of the image.

12.5.2.2 Central Dark-Ground Method

An alternative to the Zernike method is the **central dark-ground method**. The idea is very similar, but instead of a mask with a phase-shifting dot, the mask has an opaque dot at the center to *block* the dc component. The details are left as an exercise (see Problem 12.12), but we can summarize the results. The advantage is that the background (dc component) is suppressed, giving a dark background. The disadvantage is that the intensity varies nonlinearly with phase (as ϕ^2 for small ϕ).

12.5.2.3 Schlieren Method

The last method we will discuss here is important in the visualization of fluid flows, but requires considerably more mathematical sophistication to handle than the last two cases. The upshot of all this is that the **schlieren method** shows *gradients* in the phase of the object in a particular direction.

The name comes from the German word *schlieren*, which means “streaks.” One application of this method was to test glass lenses for impurities introduced during the fabrication process that would cause inhomogeneities in the refractive index. These would show up as streaks in the schlieren test, and hence the name.

The setup is the same 4-*f* filtering setup. But rather than some sort of spot in the Fourier plane to modify the dc component, the schlieren method uses a knife edge in the Fourier plane to block half of the frequency spectrum. So we can represent the transmission function of the Fourier-plane mask as the Heaviside step function:

$$t_{\text{mask}}(x, y) = U_H(x) := \begin{cases} 1, & x > 0 \\ 0, & x < 0. \end{cases} \quad (12.90)$$

Since the *y*-direction is trivial to handle, we’ll only do a one-dimensional analysis. Following the field through the 4-*f* setup,

1. **Input field:** $E_{\text{in}}^{(+)}(x)$
2. **Transform plane:** $\tilde{E}_{\text{in}}^{(+)}(k_x)$
3. **After knife edge:** $\tilde{E}_{\text{in}}^{(+)}(k_x) U_H(k_x)$
4. **Image plane:** $\mathcal{F}[\tilde{E}_{\text{in}}^{(+)}(k_x) U_H(k_x)] = [E_{\text{in}}^{(+)} * \tilde{U}_H](x)$

The last expression follows from the convolution theorem, and $\tilde{U}_H(x)$ is the Fourier transform of the Heaviside function, which we must now evaluate. Note that for simplicity, we aren’t being particularly careful with distinguishing Fourier transforms vs. inverse Fourier transforms, which means we are dropping factors of 2π (though this will turn out not to matter) and ignoring some image flips (which *do* matter—the final image should be flipped compared to what we will derive).

So let’s now compute $\mathcal{F}[U_H(x)]$. Unfortunately, this is a little tricky, so we’ll spend some time working this transform out. It turns out that the answer is

$$\mathcal{F}[U_H(x)] \equiv \tilde{U}_H(k_x) = \pi\delta(k) + \frac{1}{ik}.$$

(Fourier transform of step function) (12.91)

Rather than derive this, we’ll simply verify that the inverse transform recovers the Heaviside function. The inverse transform is

$$\mathcal{F}^{-1}[\tilde{U}_H(k_x)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \pi\delta(k)e^{ikx} dk + \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\frac{1}{ik}\right) e^{ikx} dk. \quad (12.92)$$

Note the special notation in the integral in the last term. This is because this integral is singular due to the $1/k$ dependence. To handle the singularity, we take the **Cauchy principal value** by splitting the integral about the singularity, and then approaching the singularity symmetrically:

$$\int_{-\infty}^{\infty} \cdots \equiv \lim_{\delta \rightarrow 0} \left[\int_{-\infty}^{-\delta} \cdots + \int_{\delta}^{\infty} \cdots \right]. \quad (12.93)$$

(Cauchy principle value)

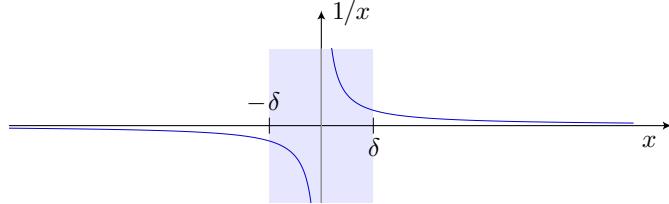
In this way, the singular part cancels; this works because the $1/k$ changes sign on either side of $k = 0$. Thus,

$$\int_{-\infty}^{\infty} \frac{dx}{x} = 0, \quad (12.94)$$

even though

$$\int_{-\infty}^{\infty} \frac{dx}{x} \quad (12.95)$$

diverges. This amounts to excluding the symmetric, shaded region in the graph below for the integrand $1/x$, and then taking the limit as the width of the “excised” region vanishes.



Of course, if the singularity occurs elsewhere, the splitting is understood to be there.

Now back to evaluating Eq. (12.92).

$$\mathcal{F}^{-1} [\tilde{U}_H(k_x)] = \frac{1}{2} + \frac{1}{2\pi i} \lim_{\delta \rightarrow 0} \left[\int_{\delta}^{\infty} \frac{e^{ikx}}{k} dk + \int_{-\infty}^{-\delta} \frac{e^{ikx}}{k} dk \right]. \quad (12.96)$$

Flipping the integration limits in the last term,

$$\mathcal{F}^{-1} [\tilde{U}_H(k_x)] = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin(kx)}{k} dk = \frac{1}{2} + \frac{1}{2} \operatorname{sgn}(x) = U_H(x). \quad (12.97)$$

Here we used the result

$$\int_0^{\infty} \frac{\sin(kx)}{k} dk = \pm \frac{\pi}{2}, \quad (12.98)$$

where the result is $\pi/2$ for $x > 0$ and $-\pi/2$ for $x < 0$ (proving this requires a contour integral, which we won’t get into here). Also, $\operatorname{sgn}(x)$ is the **signum function**, which returns the *sign* of the input (i.e., it is $+1$ for $x > 0$, -1 for $x < 0$, and 0 for $x = 0$).

So now we’ve verified that we have the correct Fourier transform of $U_H(x)$. But that’s not quite what we wanted. What we really need is the *inverse* Fourier transform of $U_H(k_x)$, to get $\tilde{U}_H(x)$. So to handle the overall scaling and sign conventions, we just need to divide our result by 2π and let $i \rightarrow -i$ to obtain

$$\tilde{U}_H(x) = \frac{1}{2} \delta(x) + \frac{i}{2\pi x}. \quad (12.99)$$

Now, as we concluded above in tracing the field through the filtering system, the output field is the convolution of the input field with $\tilde{U}_H(x)$:

$$E_{\text{out}}^{(+)} = (E_{\text{in}}^{(+)} * \tilde{U}_H)(x). \quad (12.100)$$

Recalling the form of the convolution integral from Eq. (11.5), we can thus write the output field for the schlieren setup as

$$E_{\text{out}}^{(+)} = \frac{E_{\text{in}}^{(+)}(x)}{2} + \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{E_{\text{in}}^{(+)}(x')}{x - x'} dx'. \quad (\text{output of schlieren imaging system}) \quad (12.101)$$

If $E_{\text{in}}^{(+)}(x)$ is produced by a uniformly illuminated phase object, then for small phase shifts

$$E_{\text{in}}^{(+)}(x) = E_0^{(+)} e^{i\phi(x)} \approx E_0^{(+)} (1 + i\phi(x)). \quad (12.102)$$

Then the output intensity is

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \left| \frac{1}{2} + \frac{i\phi(x)}{2} - \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\phi(x')}{x - x'} dx' + O(\phi^2) \right|^2, \quad (12.103)$$

where we have already used

$$\int_{-\infty}^{\infty} \frac{dx'}{x - x'} = 0. \quad (12.104)$$

Multiplying out the intensity expression and keeping only first-order terms in ϕ , the result is

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \frac{1}{4} - \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\phi(x')}{x - x'} dx'. \quad (\text{intensity output of schlieren imaging system}) \quad (12.105)$$

Thus, we now see how the phase information is reflected in the intensity.

The integral that appears in Eqs. (12.101) and (12.105) is of special significance, and it is called the **Hilbert transform**. For a function $f(x)$, the Hilbert transform \hat{f} of f is given by

$$\hat{f}(x) := \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(x')}{x' - x} dx'. \quad (12.106) \quad (\text{Hilbert transform})$$

As for the Fourier transform, we can also use an operator notation

$$\hat{f} \equiv \mathcal{H}[f] \quad (12.107) \quad (\text{Hilbert-transform notation})$$

to simplify some algebra.

So, we see from Eq. (12.101) that the output *field* of the schlieren system is (up to an offset and scaling factor) the Hilbert transform of the input *field*, and from Eq. (12.105) we see that the output *intensity* is the Hilbert transform of the input *phase*, in the limit of small phase shifts. So we can write Eq. (12.105) more compactly as

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \frac{1}{4} + \frac{\hat{\phi}(x)}{2}. \quad (12.108) \quad (\text{schlieren intensity as Hilbert transform})$$

But what does this mean intuitively? Well, the Hilbert transform is the convolution of a function $f(x)$ with $1/x$ (up to a minus sign). But $1/x$ diverges at $x = 0$, and right where it diverges it suddenly changes sign. When you multiply $f(x)$ by $1/x$ and integrate, you thus expect that it will take on a large value only when something upsets the cancellation at $x = 0$. That is, the value will be large when $f(x)$ has a large slope at $x = 0$, so that the two sides of $1/x$ are unequally weighted. So the Hilbert transform is “something like” a derivative. The complication is that $1/x$ falls off so slowly that the Hilbert transform picks up long-range information about the function, while the derivative is defined by information arbitrarily close to a single point. In this sense, there isn’t even really a limit (like “slowly varying functions”) for arbitrary functions in which the Hilbert transform closely approximates the derivative. But as far as visualizing the schlieren image is concerned, the derivative is a reasonable intuitive guide.

Let's go through a sloppy argument¹ that "shows" that the schlieren image is the derivative of the phase. We'll start off with the property that the Hilbert transform is its own inverse, up to a minus sign:

$$\mathcal{H}^{-1} = -\mathcal{H}, \quad (12.109)$$

(involution property of Hilbert transform)

or $\mathcal{H}^2 = -1$. This statement can be proved by using the fact that the Hilbert transform is linear, decomposing a test function into harmonic components (via the Fourier transform), and computing the Hilbert transform of the harmonic functions (see Problem 12.16). In any case, we can use Eq. (12.109) to write

$$\phi(x) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{\phi}(x')}{x' - x} dx'. \quad (12.110)$$

Differentiating, we find

$$\frac{d\phi}{dx} = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{\phi}(x')}{(x' - x)^2} dx' = -\frac{1}{\pi} \hat{\phi}(x) * \frac{1}{x^2}. \quad (12.111)$$

Here is the first problem: the integral with $1/x^2$ doesn't converge for generic functions $\hat{\phi}$, even using the Cauchy principle value. But let's go with it, since it gives a good intuitive result. Now something even *sloppier*: if $\hat{\phi}$ varies slowly, then $1/x^2$ is something like a δ -function. Again, the integral of $1/x^2$ is divergent, while the integral of the δ -function is not, and $1/x^2$ has no length scale, so it isn't clear what we mean by slowly varying. *But*, if we make the replacement, then the convolution is with the δ -function and so, dropping constant factors,

$$\frac{d\phi}{dx} \sim \hat{\phi}(x). \quad (12.112)$$

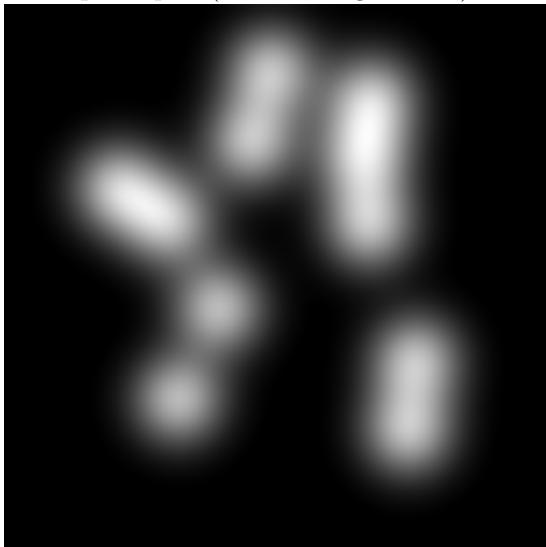
Thus, the Hilbert transform in Eq. (12.105) varies as the derivative of $\phi(x)$. So, the conclusion is that the schlieren image is approximately an intensity background plus the derivative of the phase. But if we're concerned about actually being *correct*, it's better to stick with the Hilbert transform.

12.5.2.4 Numerical Examples

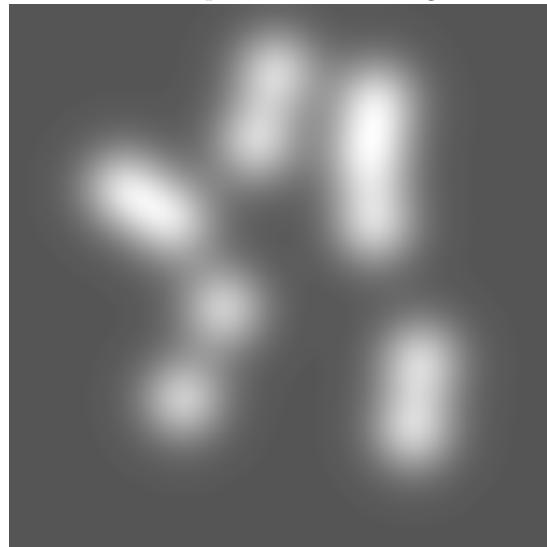
Now we'll look at simulated images of a phase object to get a better idea of how the different phase-visualization methods compare. The original image is a plot of the phase shift, where the phase shifts are concentrated in randomly placed Gaussian blobs. The first set of images is in the small-phase regime (the largest phase shift is 0.2π), so all our linearized conclusions above hold. The other three images show intensity patterns at the output of the 4-*f* imaging system (neglecting inversion and magnification). The Zernike image almost perfectly reproduces the original phase, but has an intensity background. The dark-ground image has a dark background, but the Gaussian spots are somewhat distorted. The derivative nature of the schlieren image is apparent in the last image, giving a "three-dimensional" sort of look.

The second set of images is in the large phase-shift regime, where the largest phase shift is 3π . The images are visually more interesting but harder to interpret. Similar comments to the first set of images hold here, except that the "wrapping" of the large phase shifts is apparent in the image.

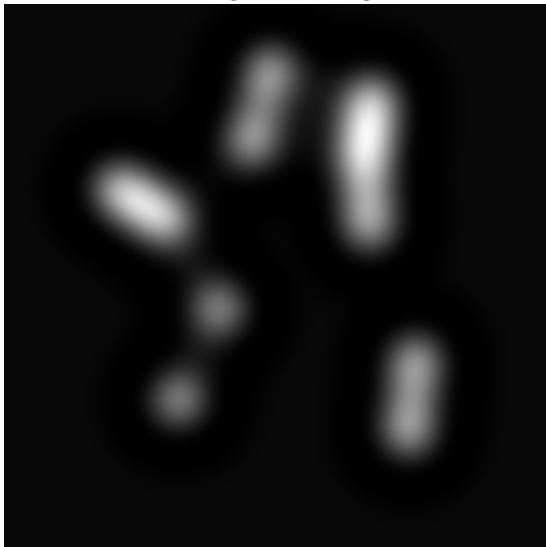
¹from John B. DeVelis and George O. Reynolds, *Theory and Applications of Holography* (Addison-Wesley, 1967).

phase plot (dark: 0, bright: 0.2π)

Zernike phase-contrast image

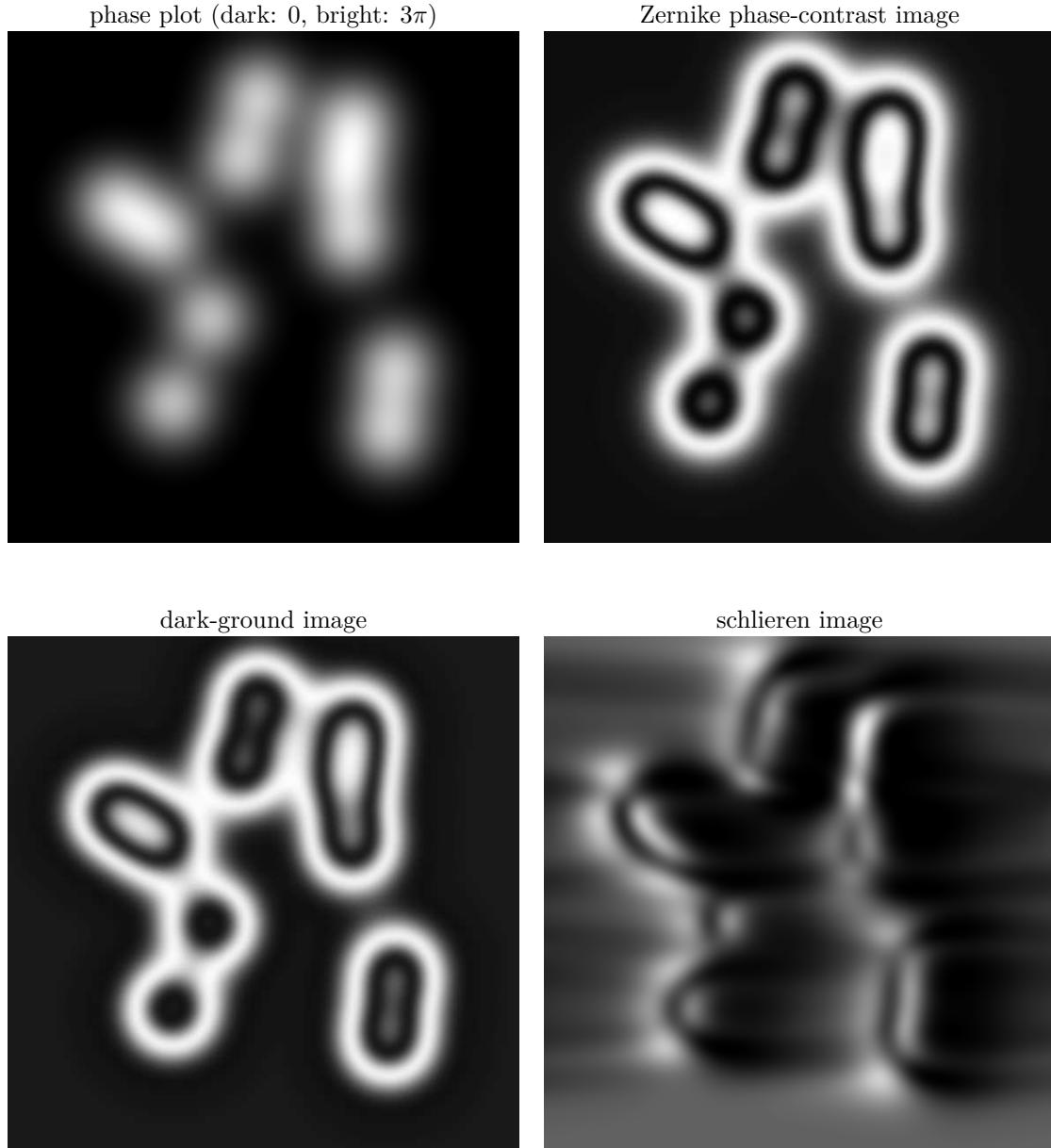


dark-ground image



schlieren image





12.6 Holography

Recall from our earlier discussion of Fourier-optics propagation that an arbitrary monochromatic field $E^{(+)}(x, y, z)$ is *completely determined* by the field in a single plane, say $E^{(+)}(x, y, z = 0)$. This is because we can simply apply the propagation factor $e^{ik_z z}$ to $E^{(+)}(x, y, z = 0)$ (in frequency space) to obtain the field at any location z .

We can exploit this idea that the three-dimensional field can be represented in two dimensions. A **hologram** (at least ideally) is a thin optical element with a complex transmission coefficient

$$t(x, y) \propto E^{(+)}(x, y, 0). \quad (12.113)$$

When uniformly illuminated by a monochromatic field, the holographic mask reconstructs $E^{(+)}(x, y, 0)$, up to an overall constant. Thus the hologram reconstructs the entire three-dimensional field $E^{(+)}(x, y, z)$, at least in the $z > 0$ region past the holographic mask.

12.6.1 Example: Single-Frequency Hologram

Suppose that the hologram has a transmission of

$$t(x, y) = e^{i(k \sin \theta)x}, \quad (12.114)$$

so that it modulates the signal with a single spatial frequency $k \sin \theta$ in the x -direction. Illuminating this with an on-axis plane wave $E^{(+)}(x, y) = E_0^{(+)}$, the transmitted (reconstructed) field is

$$t(x, y)E^{(+)}(x, y) = E_0^{(+)}e^{i(k \sin \theta)x}. \quad (12.115)$$

Then propagating this field forward to arbitrary z , we find

$$t(x, y)E^{(+)}(x, y, z) = E_0^{(+)}e^{i(k \sin \theta)x}e^{i(k \cos \theta)z}, \quad (12.116)$$

which is a plane wave with wave vector

$$\mathbf{k} = k(\hat{x} \sin \theta + \hat{z} \cos \theta). \quad (12.117)$$

The original direction was $\mathbf{k} = k\hat{z}$, so this hologram *deflects* a plane wave by an angle θ . In other words, this hologram acts like a prism. In fact, a phase delay of $\exp[i(k \sin \theta)x] = \exp(ikd)$, where $d = x \sin \theta$, literally *is* a prism.

12.6.2 Film Holograms

But now the question is, how can we make a hologram? Reconstructing the phase information in the last example requires a glass wedge of varying thickness (i.e., a prism). But photographic detectors such as film and ccd cameras respond to intensity, not to the phase of the light. A hologram needs both intensity *and* phase information, so we need to employ some tricks.

The key point in holography is this: we can use *interference* to make a hologram, mixing $E_{\text{obj}}^{(+)}(x, y)$ (the “object field” that we want to record) with a reference plane wave $E_{\text{ref}}^{(+)}(x, y) = E_{\text{ref},0}^{(+)}$. The photograph records the intensity of the superposition:

$$I_{\text{record}} = \frac{2|E_{\text{obj}}^{(+)} + E_{\text{ref}}^{(+)}|^2}{\eta}. \quad (12.118)$$

Expanding out the product,

$$I_{\text{record}} = I_{\text{obj}} + I_{\text{ref}} + \frac{2}{\eta}E_{\text{obj}}^{(+)}E_{\text{ref}}^{(-)} + \frac{2}{\eta}E_{\text{obj}}^{(-)}E_{\text{ref}}^{(+)}. \quad (12.119)$$

(recorded film signal)

Alternately, we may write out the conjugate terms to make the interference more obvious,

$$I_{\text{record}} = I_{\text{obj}} + I_{\text{ref}} + 2\sqrt{I_{\text{obj}}I_{\text{ref}}} \cos[\phi_{\text{obj}}(x, y)], \quad (12.120)$$

(recorded film signal)

where ϕ_{obj} is the object phase, given by

$$E_{\text{obj}}^{(+)} = |E_{\text{obj}}^{(+)}|e^{i\phi_{\text{obj}}(x, y)}. \quad (12.121)$$

The hologram is the developed film that is exposed to this superposition. Both the phase and amplitude information are encoded in the intensity and hence the photograph (like in an interferometer). But now getting the information back out is a bit subtle because of the encoding.

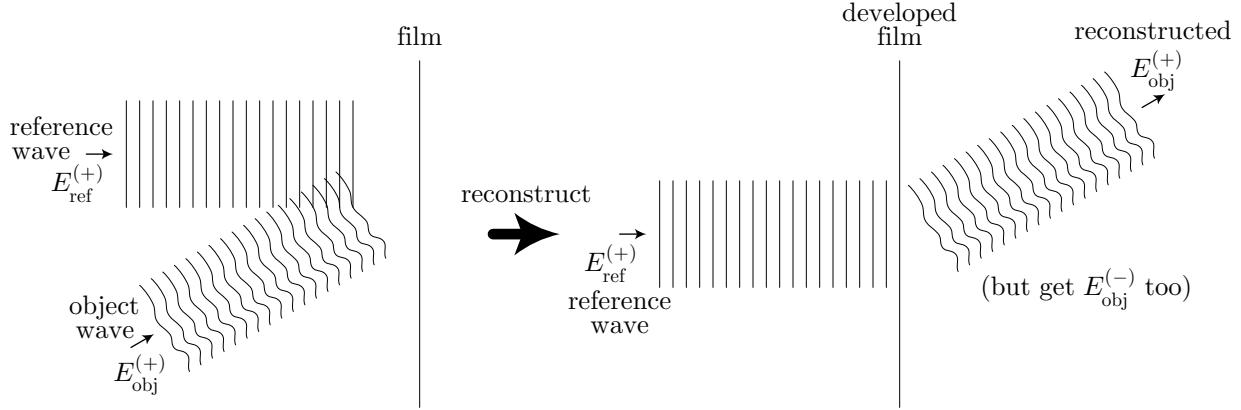
How do we recover the information recorded on the film? The idea is to illuminate the holographic mask with the same reference light $E_{\text{ref}}^{(+)} \propto \sqrt{I_{\text{ref}}}$. Then

$$E^{(+)} = tE_{\text{ref}}^{(+)} \propto E_{\text{ref}}^{(+)} I_{\text{ref}} + E_{\text{ref}}^{(+)} I_{\text{obj}} + I_{\text{ref}} E_{\text{obj}}^{(+)} + \frac{2}{\eta} E_{\text{obj}}^{(-)} [E_{\text{ref}}^{(+)}]^2. \quad (12.122)$$

Dividing through by $E_{\text{ref}}^{(+)}$,

$$\frac{E^{(+)}}{E_{\text{ref}}^{(+)}} \propto I_{\text{ref}} + I_{\text{obj}}(x, y) + \left(\frac{2E_{\text{ref}}^{(-)}}{\eta} \right) E_{\text{obj}}^{(+)}(x, y) + \left(\frac{2E_{\text{ref}}^{(+)}}{\eta} \right) E_{\text{obj}}^{(-)}(x, y). \quad (\text{reconstructed hologram field}) \quad (12.123)$$

The first term is just a constant offset, and the second is the intensity profile of the object. The original field is recovered in the last two terms. The third term is proportional to the original field, which is what we're trying to get. The fourth term is the **conjugate wave**, the complex conjugate of the original wave. This term is a consequence of the fact that the film hologram is fundamentally a real-valued transmission mask.



12.6.3 Hologram of a Plane Wave and Off-Axis Holography

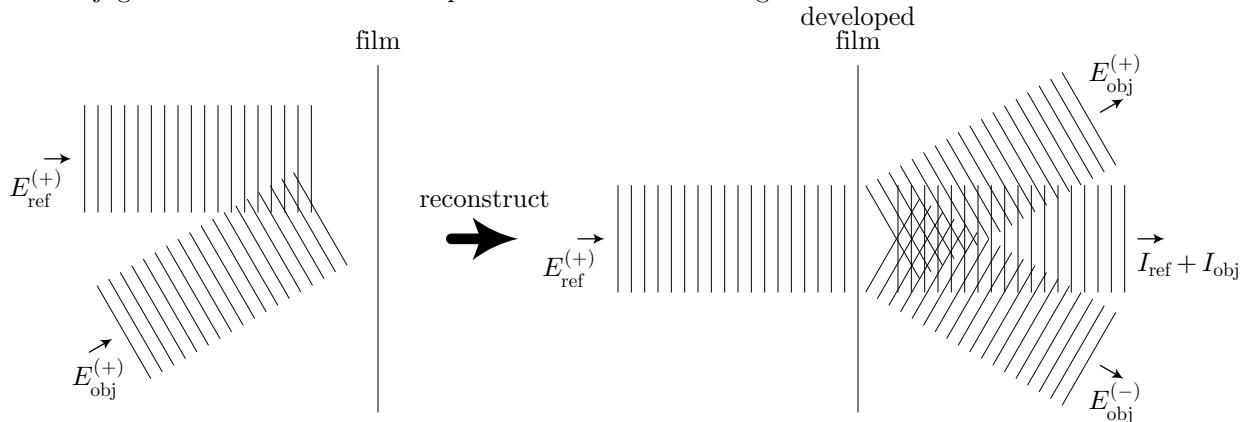
A simple example, similar to the example of Section 12.6.1, is the hologram of a single plane wave,

$$E_{\text{obj}}^{(+)} \sim \sqrt{I_{\text{obj}}} e^{ik(\sin \theta)x}. \quad (12.124)$$

Then the reconstructed wave is

$$\frac{E_{\text{obj}}^{(+)}}{\sqrt{I_{\text{ref}}}} \sim I_{\text{ref}} + I_{\text{obj}} + \sqrt{I_{\text{ref}} I_{\text{obj}}} e^{i(k \sin \theta)x} + \sqrt{I_{\text{ref}} I_{\text{obj}}} e^{-i(k \sin \theta)x}. \quad (12.125)$$

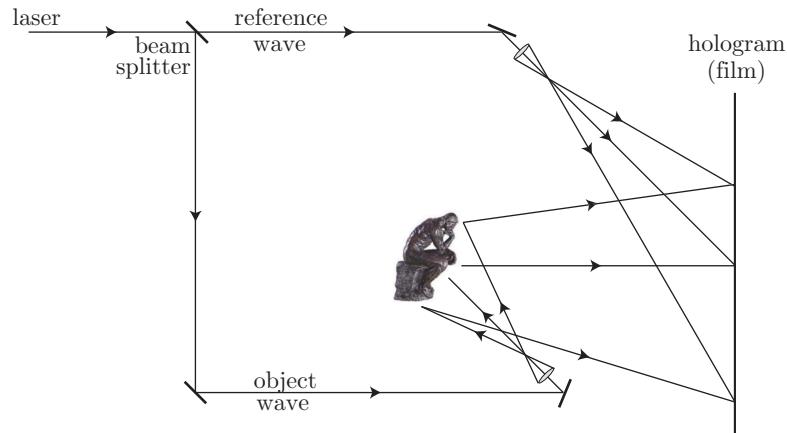
Again, the third term is the reconstructed plane wave, while the fourth term is the extra conjugate wave. The conjugate wave turns out to be a plane wave at the same angle but reflected about the z -axis.



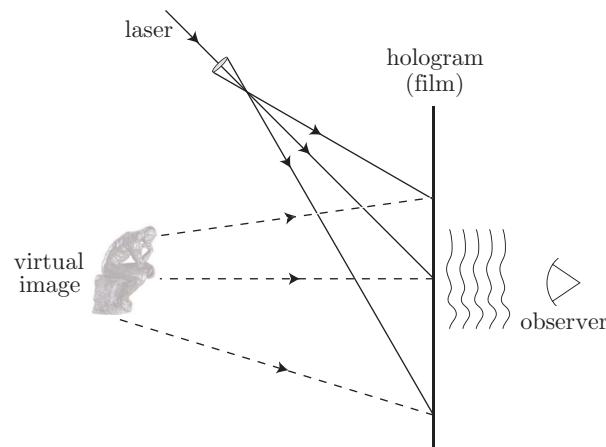
Once again, the reconstruction is not perfect because the film is fundamentally a real-valued transmitter, hence the conjugate wave. This hologram turns out to be just the sinusoidal diffraction grating from Section 12.3.4. However, this example motivates a method to get at the reconstructed object field along, called **off-axis holography**. The idea is to ensure that $E_{\text{obj}}^{(+)}$ is at a sufficiently large angle (larger than its angular spread). Then the forward wave $I_{\text{ref}} + I_{\text{obj}}$ and conjugate wave $E_{\text{obj}}^{(-)}$ are all spatially separated from the reconstructed wave $E_{\text{obj}}^{(+)}$.

12.6.4 Setup: Off-Axis Reflection Hologram

Then the optical setup for an off-axis reflection hologram is as shown here. A laser beam (monochromatic light source) is split, with half illuminating the object to form the object wave, and half going directly to the screen to form the reference wave. Notice that both beams must be expanded appropriately to cover the object and the film. Also, while recording, the relative phase of the two waves must be held very constant, so it is important to minimize the vibrations of the optics. Also, generally holographic film takes the form of glass plates for structural rigidity during exposure. Holographic plates must also have high resolution and are consequently rather expensive.



To reconstruct the hologram, the idea is to illuminate the hologram with the same reference wave (although the exact details of the reference wave aren't too important). Then the wave after the hologram forms a virtual image of the original object in the same location behind the film.



Another interesting thing to note is that it is possible to make *synthetic* holograms of objects that never really existed, for example on a computer. The appropriate pattern can be computed and printed out on a transparency, although usually the pattern will still need to be photographically reduced to achieve the required spatial resolution.

12.7 Exercises

Problem 12.1

Consider the Gaussian field profile $E^{(+)}(x, y, z = 0) = E_0^{(+)} \exp[-(x^2 + y^2)/w_0^2]$. Use the Fourier-transform formalism for propagating the field to derive the form of $E^{(+)}(x, y, z)$ (in the paraxial approximation), and thus show that your results are consistent with the Gaussian beam solution to the paraxial wave equation.

Problem 12.2

We derived a Fourier-transform recipe for propagating a field $E^{(+)}(x, y, z = 0)$ to arbitrary z under the assumption of a *forward*-propagating wave. Write down the analogous recipe for propagating the field to arbitrary z under the assumption of a *backward*-propagating wave. *Explain.*

Problem 12.3

Consider a plane wave incident on a glass–air interface (from the glass side), in total internal reflection. Use the language of Fourier optics and the diffraction limit to explain this phenomenon, and rederive expressions for the critical angle and skin depth from your discussion. (Note: what is the spectrum in k_x for a steeply angled plane wave? Since the spectrum should be continuous across the interface, which acts as a thin optic, what is the difference in k_x before vs. after the interface?)

Problem 12.4

- (a) Compute the far-field (Fraunhofer) diffraction pattern on a screen at location z due to a rectangular aperture of width b at $z = 0$. Assume the aperture is uniformly illuminated by a plane wave at normal incidence to the aperture. You may restrict your calculation to one dimension.
- (b) Write down the far-field diffraction pattern if the aperture is instead a pair of narrow slits separated by a distance a .
- (c) Use the convolution theorem to write down the far-field diffraction of an aperture consisting of two parallel slits of width b with centers separated by distance a .

Problem 12.5

- (a) Show that the far-field (Fraunhofer) diffraction pattern due to a uniformly illuminated (by a normally incident plane wave), *circular* aperture of radius a is

$$E_0^{(+)}(2\pi a^2) \frac{J_1(2\pi ar/\lambda z)}{2\pi ar/\lambda z}, \quad (12.126)$$

where $E_0^{(+)}$ is the amplitude of the illuminating wave, r is the radial distance from the center of the screen, and $J_1(x)$ is an ordinary Bessel function. The Bessel functions are defined by the integral

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(n\phi - x \sin \phi) d\phi, \quad (12.127)$$

and it may help you to know that they satisfy the recurrence relation

$$nJ_n(x) + J'_n(x) = xJ_{n-1}(x), \quad (12.128)$$

which is most useful here for $n = 1$.

- (b) Make an intensity plot of this diffraction pattern. This is a well-known pattern called the *Airy disk*.

Problem 12.6

- (a) Compute the far-field (Fraunhofer) diffraction pattern due to an infinite, regularly spaced array of

narrow slits, where the distance between adjacent slits is a . (This is a simple model of a diffraction grating.) Assume the slits are uniformly illuminated.

- (b) Compute the far-field intensity pattern due to N regularly spaced slits with spacing a .

Problem 12.7

Suppose that when illuminated by a plane wave at normal incidence, an aperture diffracts into an intensity pattern $I(x, z)$ on a screen a long distance z away. We'll ignore the y -direction for simplicity. What is the effect on the intensity pattern if the incident wave makes a small angle θ with respect to the x -axis? The answer should be reasonably obvious but you should prove your answer mathematically.

Problem 12.8

In diffraction theory, **Babinet's principle** is useful for calculating the diffraction pattern of an aperture if you know its *complement*.

More specifically, suppose you illuminate an aperture with a normally incident plane wave

$$E_{\text{in}}^{(+)}(x, y) = E_0^{(+)}. \quad (12.129)$$

We will represent the aperture by the transmission function $t(x, y)$, so that to find the diffraction pattern, we propagate the initial field

$$E^{(+)}(x, y) = t(x, y)E_{\text{in}}^{(+)}(x, y) \quad (12.130)$$

to find the diffraction field $E^{(+)}(x, y, z)$. Let us denote the complement aperture by $\bar{t}(x, y)$, such that

$$t(x, y) + \bar{t}(x, y) = 1, \quad (12.131)$$

and the diffraction field for the complement aperture by $\bar{E}^{(+)}(x, y, z)$. The idea behind Babinet's principle is that if we add the fields for the aperture and complement, we just obtain the incident plane wave:

$$E^{(+)}(x, y) + \bar{E}^{(+)}(x, y) = t(x, y)E_{\text{in}}^{(+)}(x, y) + \bar{t}(x, y)E_{\text{in}}^{(+)}(x, y) = E_{\text{in}}^{(+)}(x, y) = E_0^{(+)}. \quad (12.132)$$

Note that a plane wave propagates to arbitrary z just as a plane wave (it does *not* diffract into a delta function in the far field—there is no way to satisfy the far-field condition for a plane wave!). Thus, the propagation of the plane wave should be

$$E_0^{(+)} \longrightarrow E_0^{(+)}e^{ikz}. \quad (12.133)$$

But the superposition principle (linearity of the wave equation) says that this should be the same as the sum of the diffracted fields of the two apertures:

$$E^{(+)}(x, y, z) + \bar{E}^{(+)}(x, y, z) = E_0^{(+)}e^{ikz}. \quad (12.134)$$

Solving for the diffraction pattern of the complement aperture,

$$\bar{E}^{(+)}(x, y, z) = E_0^{(+)}e^{ikz} - E^{(+)}(x, y, z). \quad (12.135)$$

This says that the diffraction pattern of a particular aperture is equivalent to the *shadow* of the complement aperture in the incident beam.

There is, unfortunately, one complication in applying Babinet's principle, and it relates to the phase factor $(-i)$ that comes up in the coefficient $k/2\pi iz$ in the free-space transfer function (particularly in the far-field limit). (Recall that in a one-dimensional diffraction problem, this factor is $\sqrt{k/2\pi iz}$, so the corresponding phase is $\sqrt{-i}$.) The same phase should be present in the plane wave for consistency. In the far-field limit, this is an artifact of noncommuting limits. For example, suppose we consider

diffraction a Gaussian aperture, which you know should produce a Gaussian beam with longitudinal phase factor

$$e^{ikz} e^{-i \tan^{-1}(z/z_0)}, \quad (12.136)$$

including the Gouy phase factor. If we take the far-field limit $z \rightarrow \infty$, then $\tan^{-1}(z/z_0) \rightarrow \pi/2$, and the longitudinal phase becomes

$$e^{ikz} e^{-i\pi/2} = -ie^{ikz}. \quad (12.137)$$

Then if we take the plane-wave limit $w_0 \rightarrow \infty$ (which implies $z_0 \rightarrow \infty$), the phase factor is the same. On the other hand, if we take the plane-wave limit *first*, the Gouy phase reduces to $\tan^{-1}(z/z_0) \rightarrow 0$, so the longitudinal phase reduces to e^{ikz} , as we expect for a plane wave. Thus, the extra $(-i)$ is essentially a generic Gouy phase, which results from assuming a confined field inside the aperture, and taking the far-field limit. The point is that instead of using the simple plane wave $E_0^{(+)} e^{ikz}$ in Eq. (12.135), you should use the limit of the diffraction pattern as the aperture disappears, so that the diffraction pattern and the reference plane wave have appropriately matching phases. Normally you don't have to worry about this, if you just want the diffraction pattern, because this is an overall phase. But now, we're concerned with *interference* involving the diffraction pattern, so it's important.

Of course, you can avoid all this by just considering the *intensity* of the diffraction pattern, and you should do this for this problem. The same argument above carries through for the *intensities* of the diffraction pattern, with the result

$$\bar{I}^{(+)}(x, y, z) = I_0 - I(x, y, z), \quad (12.138)$$

where I and \bar{I} are the regular and complementary diffraction patterns, respectively, and I_0 is the intensity of the incident plane wave.

Use Babinet's principle to compute the Fraunhofer diffraction pattern for a plane wave around an opaque strip (e.g., wire) of material of width b . (Be careful to get the phase right when you set up the diffraction for the complement aperture.) Verify that your answer makes sense in any appropriate limits.

Problem 12.9

Consider a Gaussian beam, with the focus occurring at $z = 0$, and propagating along the z -axis. Suppose a circular aperture is introduced in the $z = 0$ plane, centered with respect to the Gaussian beam's axis. Describe *qualitatively* the far-field (large z) profile of the apertured beam, paying special attention to the limits of large and small aperture diameter. *Describe* the mathematical form of the far-field profile for *arbitrary* aperture size.

Note: you should mention convolutions in your answer.

Problem 12.10

- (a) Write down an expression for the Fresnel diffraction pattern for a knife edge illuminated by a plane wave at normal incidence. Write your answer in terms of the Fresnel integrals $C(x)$ and $S(x)$.
- (b) Make a few representative plots showing the diffraction pattern in the near, intermediate, and far fields. (Better yet, try to do all this in one plot!)

Problem 12.11

Suppose you have a cruddy Gaussian beam of radius $w_0 = 1$ mm and wavelength $\lambda = 780$ nm that you want to clean up and expand.

- (a) Using a lens of focal length $f = 25.4$ mm to focus the beam, what is the best pinhole diameter according to the "rule of thumb" to spatially filter the beam? You may assume that the lens is at the focus of the original beam.
- (b) If you want to expand the beam to a final radius of $w_0 = 10$ mm, what focal length lens should you use to recollimate the beam after the pinhole?

Problem 12.12

Following the treatment for the Zernike phase-contrast imaging system, work out the output intensity of a central dark-ground imaging system, which is a $4-f$ system with a filter consisting of an absorbing dot that blocks the dc component in the Fourier plane. Assume that the system input is from a phase object, of the form $\exp[i\phi(x)]$.

Problem 12.13

Go through the derivation in the notes of the output intensity of a schlieren imaging system, where the input is a phase object. Do this *in detail*, filling in the missing steps where appropriate. Also go through the argument that the Hilbert transform is “something like” a derivative, so the schlieren image is something like the derivative of the phase shift.

Problem 12.14

Compute the Hilbert transforms of the following functions: $\sin(x)$, $\cos(x)$, and $\exp(ix)$. From your answer, explain the following statement: for narrowband signals, the Hilbert transform *is* proportional to the derivative, without any silly approximations. *Hint:* use the convolution theorem to evaluate the transform integral.

Problem 12.15

Compute the Hilbert transform of the unit box function $\text{rect}(x)$:

$$\text{rect}(x) := \begin{cases} 1 & \text{if } |x| < 1/2 \\ 0 & \text{if } |x| > 1/2. \end{cases} \quad (12.139)$$

Problem 12.16

Recall that the Hilbert transform is defined by

$$\mathcal{H}[f(x)] \equiv \hat{f}(x) := \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(x')}{x' - x} dx'. \quad (12.140)$$

Show that the Hilbert transform is its own inverse, provided you throw in an extra minus sign; that is, show that

$$\mathcal{H}^{-1} = -\mathcal{H}. \quad (12.141)$$

One method to prove this is as follows. Recall from Problem 12.14 that $\mathcal{H}[e^{\pm ix}] = \pm ie^{\pm ix}$. From this you can show that $\mathcal{H}[e^{ikx}] = \text{sgn}(k)ie^{ikx}$. Then compute what the function $\mathcal{H}[f(x)]$ looks like in frequency space, and finally do the same for $\mathcal{H}^2[f(x)]$.

Problem 12.17

Consider a normalized Gaussian of the form

$$G(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}. \quad (12.142)$$

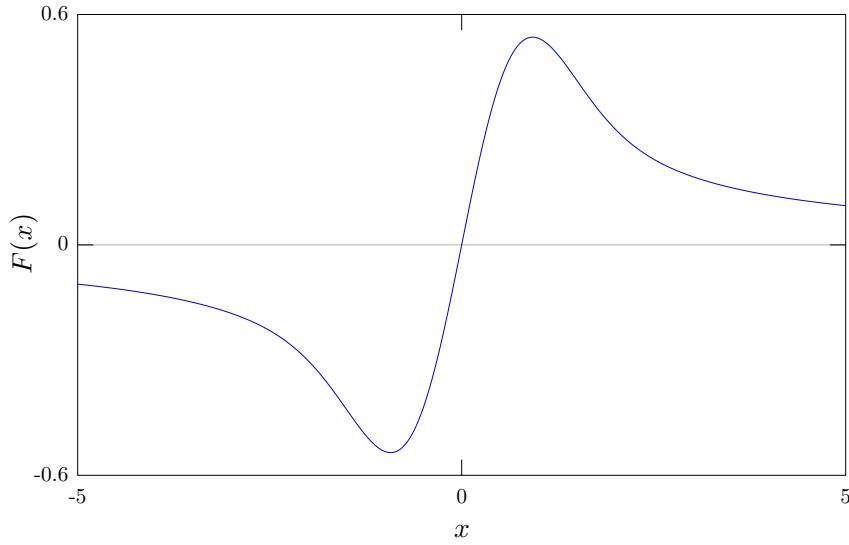
Show that the Hilbert transform $\hat{G}(x)$ may be written

$$\hat{G}(x) = -\frac{2}{\pi} F(x), \quad (12.143)$$

where

$$F(x) := e^{-x^2} \int_0^x dx' e^{x'^2} \quad (12.144)$$

is **Dawson's integral**, plotted below.



Use the following outline to do this.

- (a) Start by assuming the relation between $F(x)$ and $\hat{G}(x)$ holds; we will transform this into the definition for $F(x)$ using reversible steps. Write out the Hilbert transform:

$$F(x) = -\frac{\pi}{2} \hat{G}(x) = -\frac{1}{2} \int_{-\infty}^{\infty} dx' \frac{G(x')}{x' - x} = -\frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} dx' \frac{e^{-x'^2}}{x' - x}. \quad (12.145)$$

Then change variables by letting $x' \rightarrow x' + x$, and use the result to show that $F(x)$ satisfies the differential equation

$$F'(x) = 1 - 2xF(x). \quad (12.146)$$

- (b) Now the idea is to integrate this differential equation. Show that the differential equation

$$e^{-x^2} \frac{d}{dx} \left[e^{x^2} F(x) \right] = 1, \quad (12.147)$$

is equivalent to the one in part (a).

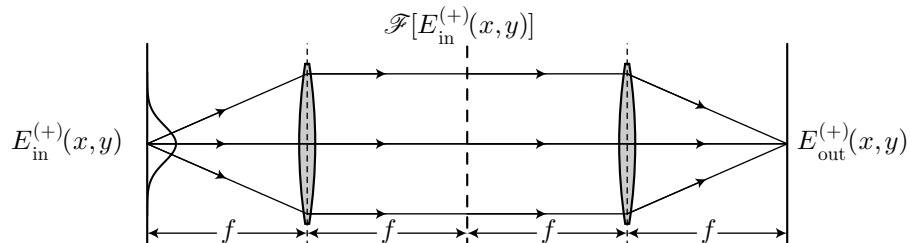
- (c) Now multiply through by e^{-x^2} and integrate both sides from 0 to x , to obtain the definition of $F(x)$.

- (d) Plot $\hat{G}(x)$ and $G'(x)$, and comment on the statement that the Hilbert transform is “something like a derivative.”

- (e) To make the distinction more formal, obtain a large- $|x|$ asymptotic form for $\hat{G}(x)$ [and thus for $F(x)$], and show that it is a much slower decay than the tails of $G'(x)$.

Problem 12.18

Consider the 4- f imaging system shown here, and suppose that there is nothing between the two lenses of focal length f .



Recall that the field in the focal plane is given by

$$E_{\text{focalplane}}^{(+)}(x, y) = \frac{k}{2\pi f} \tilde{E}_{\text{in}}^{(+)} \left(\frac{kx}{f}, \frac{ky}{f} \right). \quad (12.148)$$

Show, by computing the Fourier transform (*not* the inverse transform) of the focal-plane field, that the field in the output plane is

$$E_{\text{out}}^{(+)}(x, y) = E_{\text{in}}^{(+)}(-x, -y). \quad (12.149)$$

Strictly speaking, this is only true in the paraxial approximation. This is the image inversion of the 4- f system that we ignored.

Problem 12.19

- (a) What is the diffraction limit that occurs in optical-wave propagation? Describe the effect mathematically, using the Fourier-transform propagation recipe.
- (b) Consider the 4- f imaging system from Problem 3, and suppose again that there is nothing between the two lenses of focal length f . Suppose also that the lenses are of diameter D . What is the new diffraction limit for this situation? For simplicity, you may assume that the image object represented by $E_{\text{in}}^{(+)}(x, y)$ is much smaller than D , and you can also ignore the effect of the right-hand lens.

Hint: Start by sketching some rays from the object through the system, to see how the image forms in the output plane. Then recall the argument for why the Fourier transform of the input image appears in the focal plane of the first lens. How is the field in the transform plane affected by the finite lens diameter?

Problem 12.20

The Hilbert transform, being a convolution with $1/x$, is “something like” a derivative. But a true derivative, written as a convolution, involves a delta function. Prove (by induction) that

$$f^{(n)}(x) = (-1)^n (f * \delta^{(n)})(x), \quad (12.150)$$

where $f^{(n)}(x)$ is the n th derivative of $f(x)$, and $\delta^{(n)}(x)$ is the n th derivative of $\delta(x)$.

Problem 12.21

Consider the integral

$$\int_{-\infty}^{\infty} x \, dx. \quad (12.151)$$

- (a) Explain why this integral is problematic.
- (b) Give a definition of the Cauchy principal value of this integral (suitably generalized from the $1/x$ example from class),

$$\int_{-\infty}^{\infty} x \, dx. \quad (12.152)$$

What is the value of this integral?

- (c) Give an alternate definition of this integral (let’s call it the “Schmauchy principal value”) such that the value of the integral is 1. (Thus, you are giving another reason why the original integral is not well-defined without being careful.)

Problem 12.22

- (a) Write down a formal definition for the Cauchy principal value integral in the special case of the integrand

$$\int_{-\infty}^{\infty} \frac{dx}{(x+a)(x-a)}. \quad (12.153)$$

(b) Show that

$$\int_{-\infty}^{\infty} \frac{dx}{(x+a)(x-a)} = 0 \quad (a \in \mathbb{R}, a \neq 0). \quad (12.154)$$

Hint: write out the partial-fraction decomposition of the integrand; that is, set the integrand equal to $c_+/(x+a) + c_-/(x-a)$ for undetermined constants c_{\pm} .

Problem 12.23

Compute the Hilbert transform of $f(x) = 1/x$.

Hint: play with the transform integral to try to guess the answer, up to a constant factor. This is most of the way there. Then verify your answer using the *inverse* transform. You may also find the following formula useful (Problem 12.22):

$$\int_{-\infty}^{\infty} \frac{dx}{(x+a)(x-a)} = 0 \quad (a \in \mathbb{R}, a \neq 0). \quad (12.155)$$

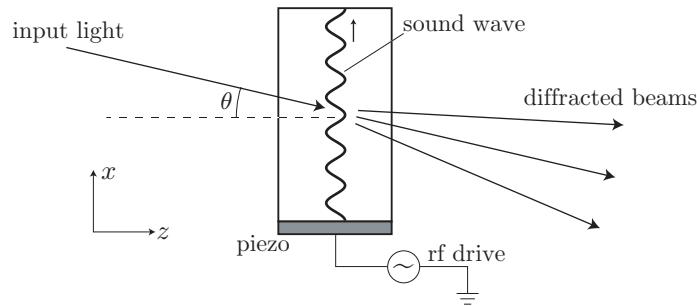
Problem 12.24

Consider a hologram of an object, recorded on a photographic film plate with light of wavelength λ . Compare the reconstructed image when using a reference wave of wavelength λ vs. a reference wavelength of some *other* wavelength λ' . For concreteness, take $\lambda' < \lambda$, and be *quantitative* in your answer. (*Hint:* think of the photographic plate as basically a complex diffraction grating, and use diffraction theory to analyze this problem.)

Chapter 13

Acousto-Optic Diffraction

One important application of diffraction occurs when you can make a *dynamical* diffraction grating. The **acousto-optic effect** is one way to accomplish a dynamical grating. The idea is that a piezoelectric transducer driven by a radiofrequency (rf) wave (typically in the 10 MHz to 1 GHz range) causes an acoustic wave to propagate through an optical crystal.



The sound wave leads to local compressions in the crystal and thus local density variations. This in turn leads to a spatial variation in the refractive index due to the acoustic strain. Generally, the other end of the crystal is cut at an angle to avoid a backreflected acoustic wave and thus a standing wave. For a monochromatic acoustic traveling wave, the resulting refractive index variation is

$$n(x) = n_0 + \delta n \cos(k_{\text{rf}}x - \omega_{\text{rf}}t), \quad (13.1)$$

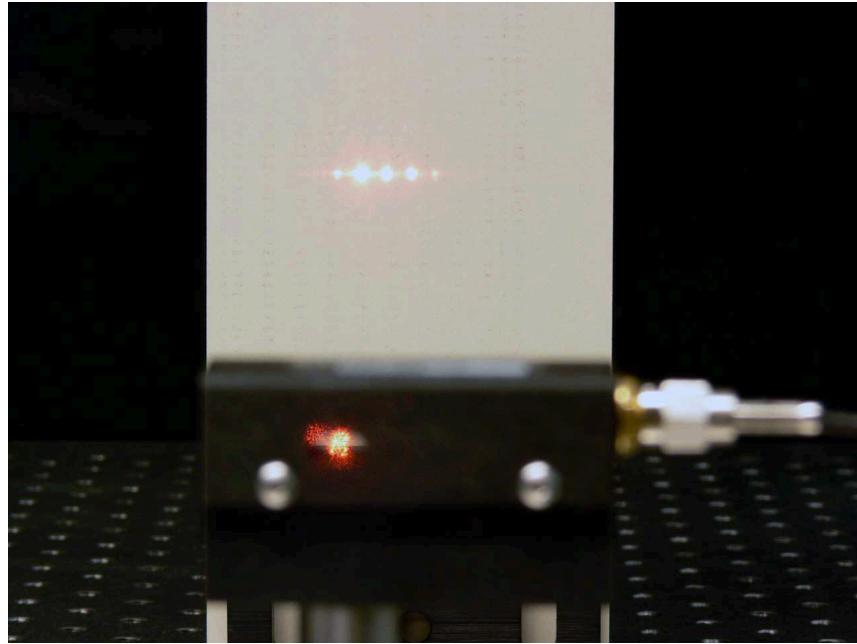
where n_0 is the natural refractive index of the crystal and δn is the modulation amplitude induced by the acoustic strain. Typically, $\delta n \ll n_0$.

As implied by the name, this device is useful for modulating the amplitude of a wave (since the “amplitude” of the diffraction grating can be controlled electronically), and it turns out that this device can also change the *frequency* of a laser beam.

A commercial acousto-optic modulator is shown below. This is an 80 MHz, TeO_2 modulator designed for 780 nm light. The crystal is the chunk of transparent material on the left-hand side of the modulator. Notice the red piezo transducer on the right-hand side of the crystal, and notice the angle-cut of the left-hand side of the crystal to suppress the reflected acoustic wave. The crystal is surrounded by aluminum to help it dissipate heat from the acoustic wave. The rf power enters via the coaxial cable on the right, and the coils in the modulator form an impedance-matching circuit to maximize the coupling efficiency into the piezo.



This photo shows the same modulator passing a He-Ne laser beam. In the far-field, the beam is diffracted into several spots. This modulator is designed to maximize the power in a single diffraction order (operating in the Bragg regime, as we will discuss below), but in this regime and alignment multiple diffracted orders are visible.



13.1 Raman–Nath Regime

There are distinct regimes of acousto-optic diffraction, essentially depending on the optical path length of the crystal. We will first treat the simpler **Raman–Nath regime**, where the acoustic wave is confined to a small width Δz . What this means is that we can treat the effect of the modulator as simply leaving a *phase imprint* on the beam, as if it were an aperture. That is, we don't have to worry about the fact that the beam

diffracts a bit from the front of the crystal before getting to the back of the crystal. Thus, this problem is essentially that of computing the Fraunhofer diffraction pattern due to a sinusoidal phase grating, although we will be a bit more careful and go beyond the paraxial approximation.

The phase imprint due to the acoustic wave is just $e^{i\mathbf{k} \cdot \Delta\mathbf{r}}$, where $\Delta\mathbf{r}$ is the path length through the acoustic wave and the dependence on the refractive index is implicit in the longitudinal spatial frequency k_z . We can write this out explicitly:

$$\exp(i\mathbf{k} \cdot \Delta\mathbf{r}) \approx \exp\left[\frac{ik\Delta z}{n_0 \cos \theta}[n_0 + \delta n \cos(k_{\text{rf}}x - \omega_{\text{rf}}t)]\right], \quad (13.2)$$

where θ is the incident angle (with respect to the z -axis), and the factor of $\cos \theta$ accounts for the fact that the acoustic beam “looks” wider to the optical wave when it propagates through at angle θ from the normal (in the past, we ignored this in the paraxial approximation). Again, this is an approximation assuming that Δz is small, so that the diffraction grating acts as a thin optic. It is also important to note the convention here: we are doing all the diffraction analysis *inside* the crystal, so k and λ refer to the spatial frequency and wavelength inside the crystal, so that they are changed by a factor of n_0 and $1/n_0$ from their respective vacuum values. Also, θ refers to the incident angle *inside* the crystal. Generally it is most useful to analyze the incident and diffracted beams *outside* the crystal, which requires a simple application of Snell’s law.

We can now write Eq. (13.2) in two parts as

$$\exp(i\mathbf{k} \cdot \Delta\mathbf{r}) \approx \exp\left[\frac{ik\Delta z}{\cos \theta}\right] \exp\left[\frac{ik\Delta z \delta n}{n_0 \cos \theta} \cos(k_{\text{rf}}x - \omega_{\text{rf}}t)\right]. \quad (13.3)$$

The first factor is just an overall phase, which we can discard. We can define

$$K := \frac{k\Delta z \delta n}{n_0 \cos \theta} \quad (13.4) \quad (\text{phase-shift amplitude})$$

as the amplitude of the phase imprint, so that

$$\exp(i\mathbf{k} \cdot \Delta\mathbf{r}) \approx \exp[iK \cos(k_{\text{rf}}x - \omega_{\text{rf}}t)]. \quad (13.5)$$

Now we need to break this phase imprint into simpler components.

One important identity is the generating function for the ordinary Bessel functions:

$$\exp\left[\frac{x}{2}\left(t - \frac{1}{t}\right)\right] = \sum_{j=-\infty}^{\infty} J_j(x)t^j. \quad (13.6)$$

This function generates the Bessel functions $J_n(x)$ in the sense that $J_n(x)$ is the coefficient of t^n in the power-series (Laurent) expansion. Letting $t \rightarrow ie^{ix}$ and $x \rightarrow K$, we obtain a useful identity:

$$e^{iK \cos x} = \sum_{j=-\infty}^{\infty} i^j J_j(K) e^{ijx}. \quad (13.7) \quad (\text{modulation identity})$$

Essentially, we just computed the Fourier series of $e^{iK \cos x}$.

If the input is the plane wave $E_{\text{in}}^{(+)} \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$, then after the phase imprint, we can use Eqs. (13.5) and (13.7) to write the effect of the diffraction as

$$E_{\text{in}}^{(+)} \exp[-(\mathbf{k} \cdot \mathbf{r} - \omega t)] \longrightarrow E_{\text{in}}^{(+)} \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)] \sum_{j=-\infty}^{\infty} i^j J_j(K) \exp[i(k_{\text{rf}}x - \omega_{\text{rf}}t)]. \quad (\text{Raman–Nath diffraction pattern}) \quad (13.8)$$

Again, we have technically just computed the Fraunhofer pattern, taking into account a different incident angle, but we will improve on this in a moment.

13.1.1 Diffraction Amplitudes: Bessel Functions

Each component in the sum corresponds to a different diffracted wave. The j th wave, or the j th *order* diffracted wave, has an amplitude of $J_j(K)$, ignoring an overall phase. Each wave has a different wave vector and thus propagates in a different direction. We will analyze this fact more carefully below, but for now we will simply note that this means that it makes sense to compute the intensities of each diffracted wave separately. We are treating the waves as plane waves, but really they are beams with finite diameter, and we will generally be interested in the intensities of the waves in the far field when they are spatially separated.

The intensity of the j th diffracted wave is thus given by

$$\frac{I_j}{I_{\text{in}}} = \frac{|E_j^{(+)}|^2}{|E_{\text{in}}^{(+)}|^2} = J_j^2(K). \quad (13.9)$$

To get a feeling for what this means, we need to explore the Bessel functions a bit more.

For our purposes here, we can define the Bessel function of order ν via the integral

$$J_\nu(x) = \frac{(x/2)^\nu}{\sqrt{\pi}\Gamma(\nu + 1/2)} \int_0^\pi \cos(x \cos \theta) \sin^{2\nu} \theta d\theta, \quad (13.10)$$

for $\nu > -1/2$, and where $\Gamma(\nu)$ is the Gamma function (which satisfies $\Gamma(n+1) = n!$ for integer arguments). In Eq. (13.9), we only need the Bessel functions of integer order n , where the integral simplifies to

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin \theta - n\theta) d\theta. \quad (13.11)$$

The Bessel functions satisfy the recurrence relations

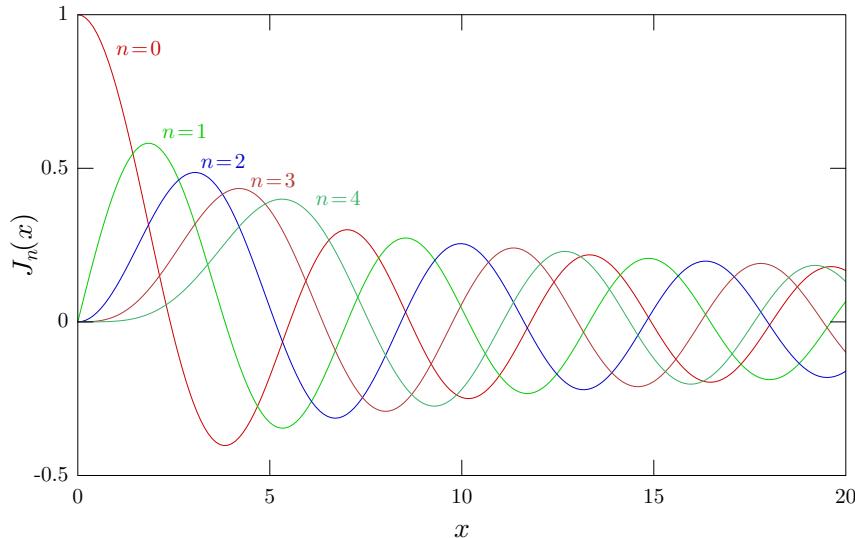
$$J_{\nu+1}(x) = \frac{2\nu}{x} J_\nu(x) - J_{\nu-1}(x); \quad \frac{\nu}{x} J_\nu(x) = J_{\nu-1}(x) - J'_\nu(x). \quad (13.12)$$

For integer orders, it is handy to know that $J_n(x) = (-1)^n J_{-n}(x)$, and for real x , we can also write $J_n(-x) = (-1)^n J_n(x)$.

For small arguments, the Bessel functions have the form

$$J_n(x) \approx \frac{x^n}{2^n n!} \quad (13.13)$$

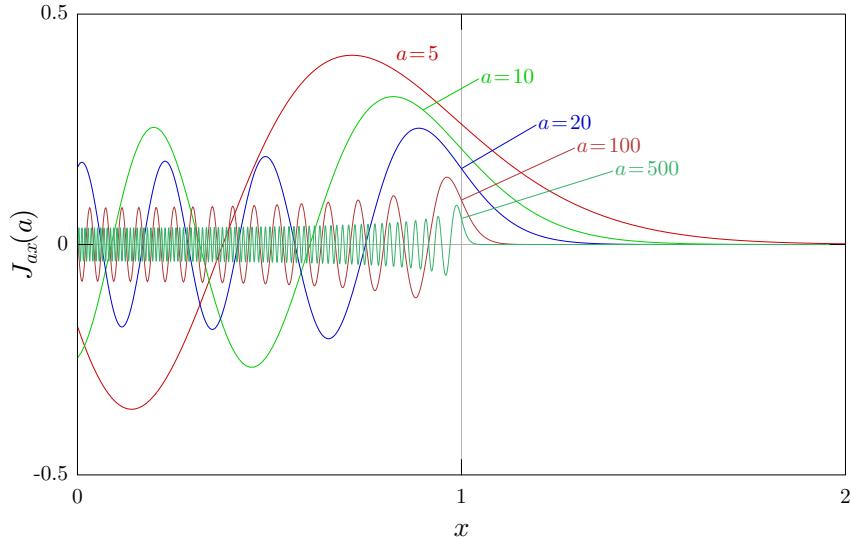
for $n \geq 0$. In particular, $J_0(0) = 1$ and $J_{n \neq 0}(0) = 0$. As x increases from zero, the higher-order Bessel functions “turn on later” compared to their low-order counterparts. The Bessel functions oscillate, and as x becomes large, they decay as $x^{-1/2}$. The first few ordinary Bessel functions are plotted here.



For fixed argument x and large order ν , the Bessel functions are approximately

$$J_\nu(x) \approx \frac{1}{\sqrt{2\pi\nu}} \left(\frac{ex}{2\nu} \right)^\nu. \quad (13.14)$$

In particular, notice that for large order ν , the amplitude *decays exponentially* with ν . We can conclude that as a function of n , $J_n(x)$ has possibly large amplitude (oscillating with n) until n exceeds x . Then, the amplitude of $J_n(x)$ drops quickly off to zero. Here is a plot of $J_{ax}(a)$ for several different values of a . Whenever x exceeds 1, the value of the Bessel function drops rapidly off, as we expect.



The point of all this Bessel stuff is this: Practically, for Raman-Nath diffraction, the maximum diffraction order that contains substantial power is K . So we expect to see about $2K + 1$ spots in the diffraction pattern of the device.

13.1.2 Frequency Shifts

The j th order diffracted wave has a time dependence of the form $\exp(-i\omega_j t)$, where the frequency is

$$\omega_j = \omega + j\omega_{\text{rf}}. \quad (13.15)$$

(diffracted frequencies)

Thus, each diffracted wave also experiences a frequency shift. There are two intuitive interpretations of this shift. The first is the view that the diffraction is a nonlinear scattering process of photons from a stream of phonons. A **phonon** is the sound-wave analog of a photon, or in other words a quantum of the acoustic field. Diffraction into the $+|j|$ order corresponds to the *absorption* of j phonons by the photon, hence raising the photon energy. Diffraction into the $-|j|$ order, on the other hand, corresponds to the *stimulated emission* of j phonons by the photon, hence lowering the photon energy. Thus, we can view the frequency shift as a conservation of energy principle, where part of the energy is carried by the phonon. The other interpretation is purely classical: The diffraction grating is moving, and thus the deflected (diffracted) beams are frequency-shifted due to the Doppler effect.

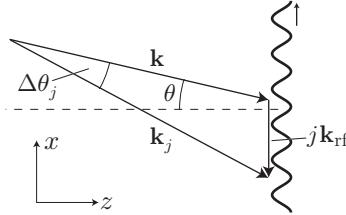
13.1.3 Momentum Conservation

Now we need to examine the spatial dependence of the diffracted waves. From Eq. (13.8), the wave vector of the j th diffracted wave is

$$\mathbf{k}_j = \mathbf{k} + j\mathbf{k}_{\text{rf}}, \quad (13.16)$$

(momentum conservation)

where $\mathbf{k}_{\text{rf}} = k_{\text{rf}}\hat{x}$. Since the momentum of a photon is proportional to its wave vector, we can regard this relation as conservation of momentum. This is a vector relation, though so it is a bit more complicated than conservation of energy.



Note that in this diagram, $j\mathbf{k}_{\text{rf}}$ points opposite \mathbf{k}_{rf} , so this corresponds to some negative diffraction order.

Optical and acoustic frequencies are very different; optical frequencies are around 10^{14} Hz, while acoustic (rf/microwave) frequencies are in practical limited to the few GHz range, and are much more commonly around 100 MHz. Thus,

$$\omega_{\text{rf}} \ll \omega, \quad (13.17)$$

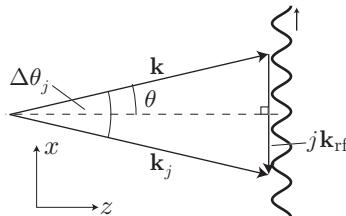
so while relative frequency shifts are observable in very sensitive setups, for our purposes here we can write for absolute frequencies

$$\omega_j \approx \omega, \quad (13.18)$$

which implies that the spatial frequency of the diffracted wave is equal to that of the incident wave to very good approximation:

$$k_j \approx k. \quad (13.19)$$

Thus, the momentum-vector triangle must be isosceles, so that \mathbf{k} and \mathbf{k}_j are symmetrical about the z -axis.



From the diagram, we can see that

$$\sin \frac{\Delta\theta_j}{2} = \frac{k_{\text{rf}}}{2k}, \quad (13.20)$$

or

$$\Delta\theta_j = 2 \sin^{-1} \frac{k_{\text{rf}}}{2k} = 2 \sin^{-1} \frac{\lambda}{2\lambda_{\text{rf}}}, \quad (13.21)$$

(diffraction angles)

where $\lambda_{\text{rf}} = 2\pi/k_{\text{rf}}$ is the wavelength of the rf acoustic wave.

Now, it's important to realize that from the argument we just presented, the statement

$$\mathbf{k}_{\text{rf}} = k_{\text{rf}}\hat{x} \quad (13.22)$$

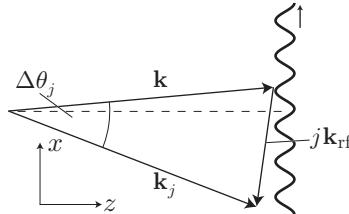
can't be quite true, even though the acoustic wave propagates in the x -direction. If we assume that $k_j \approx k$ to conserve energy, the only way to satisfy momentum conservation (i.e., to get the triangle to close) is to have the isosceles triangle shown above. But if $\mathbf{k}_{\text{rf}} = k_{\text{rf}}\hat{x}$, then for any given input angle θ , there can be at most one diffraction order that closes the momentum triangle. In particular, this can only be true for the j th order if

$$\theta = -\frac{\Delta\theta_j}{2}, \quad (13.23)$$

so that

$$\theta_j = \theta + \Delta\theta_j = \frac{\Delta\theta_j}{2}. \quad (13.24)$$

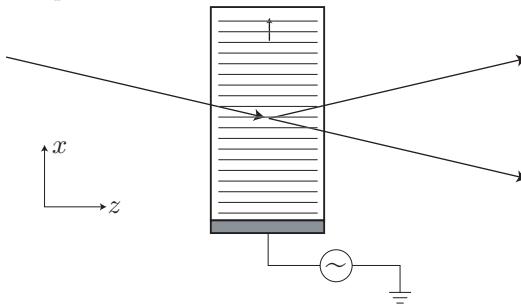
This seems to imply that there is only one diffracted beam. However, in the Raman–Nath regime, many diffraction orders are observed. How is this possible? Well, we started off with the assumption that the acoustic wave was confined to a very narrow beam. The acoustic wave is subject to diffraction effects, so the narrow spatial profile implies that it has a large spread in the direction of k_{rf} , just as it would if it were an optical wave—diffraction patterns from smaller apertures are larger in the far field for precisely this reason, and this is another manifestation of the uncertainty relation $\delta x \delta k_x \sim 1$ that we saw in the discussion of Fraunhofer diffraction. Since there are many different available orientations of k_{rf} available, conservation of energy and momentum “pick out” the right phonon that closes the triangle. It is the thinness of the acoustic beam that makes the proper phonons available for many diffraction orders.



The length of k_{rf} is the same; it's only the orientation that changes, so for “nonresonant” diffraction orders, the momentum triangle has the same shape, but the entire thing is rotated appropriately.

13.2 Bragg Regime

The **Bragg regime** of acousto-optic diffraction is the opposite extreme from Raman–Nath diffraction. Instead of a narrowly confined acoustic wave, in the Bragg regime, we assume that the acoustic wave is very broad but relatively weak. In this case, the acoustic wave is approximately a plane wave, so that k_{rf} has a well-defined direction. The arguments of the last section hold here, so we can already conclude that diffraction can only occur into a single order. This turns out to be a useful fact, since nearly all of the incident wave can be diffracted into a single order. For this reason most acousto-optic modulators in laboratory use are Bragg-regime devices, since the diffracted beam is the intensity-modulated beam. It is thus desirable to have as much modulation range as possible as well as to not waste too much of the incident beam.

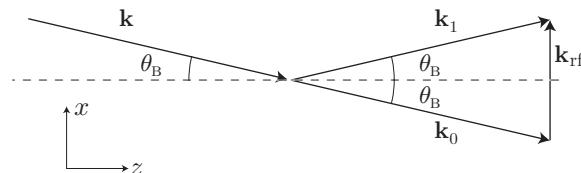


We will thus consider only first-order diffraction. Simplifying the setup from the Raman–Nath discussion above, we see that the incident angle must satisfy the **Bragg condition**,

$$\theta = \theta_B := \frac{\Delta\theta_1}{2} = \sin^{-1} \frac{\lambda}{2\lambda_{\text{rf}}} \quad (13.25)$$

(Bragg condition)

Again, diffraction in this configuration corresponds to absorbing a single phonon. Diffraction to other orders in this geometry is suppressed.



The diffracted wave makes an angle $2\theta_B$ with respect to the incident wave. Notice that the Bragg condition is equivalent to the following condition: the variations in the *optical wave* in the direction of the acoustic wave must have *twice* the wavelength of the acoustic wave. There is an interpretation to this, which is as follows: each of the “bumps” in the refractive-index profile reflects a little bit of the light in the opposite direction. If the small reflections from successive bumps interfere constructively, then a *lot* of the light can be reflected in the opposite direction. The condition for this is that one round-trip between two bumps is 2π in phase—one round trip in the direction of the acoustic wave, that is. Thus, the array of bumps acts as a sort of generalized Fabry–Perot cavity, whereby if the transverse optical wavelength matches the grating wavelength, efficient diffraction occurs. Of course this argument works for larger angles, when the transverse optical wavelength is smaller (e.g., when the transverse optical wavelength exactly matches the acoustic wavelength): these larger angles correspond to higher diffraction orders.

Recall that forward propagation over a distance z is equivalent to multiplication by $\exp(ik_z z)$. We will want to be a little more general than this, so let’s say that when you propagate the wave forward in the z -direction from z to $z + dz$, the field picks up some phase change $d\phi$:

$$E^{(+)}(z + dz) = e^{i d\phi} E^{(+)}(z). \quad (13.26)$$

Expanding the left-hand side, we can write

$$E^{(+)}(z + dz) = E^{(+)}(z) + \frac{\partial E^{(+)}}{\partial z} dz, \quad (13.27)$$

since for infinitesimals, $dz^2 = 0$. Expanding the exponential in the right-hand side, we can similarly write

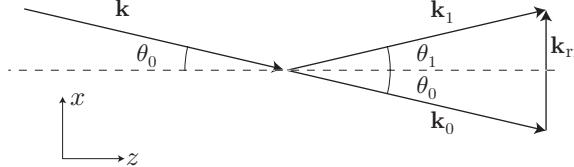
$$e^{i d\phi} E^{(+)}(z) = E^{(+)}(z)[1 + i d\phi], \quad (13.28)$$

since $d\phi^2 = 0$. Thus, we can write Eq. (13.26) in differential form as

$$\frac{\partial E^{(+)}}{\partial z} = i \frac{\partial \phi}{\partial z} E^{(+)}. \quad (13.29)$$

In principle, this evolution equation is the same as the wave equation, but this equation is first-order: it implicitly considers only forward-propagating waves ($k_z > 0$), whereas the full wave equation handles propagation in *any* direction. This is fine as long as there isn’t anything like a dielectric interface that would generate a reverse wave.

Now let’s calculate the phase change $d\phi$ as the wave propagates in the acousto-optic medium. Note that we will want to keep track of the phase evolution *along* each of the diffracted waves, so we need to compute the phase evolution along the wave vectors \mathbf{k}_0 and \mathbf{k}_1 . As $z \rightarrow z + dz$, the path length along $\mathbf{k}_{0,1}$ is $dz / \cos \theta_{0,1}$, where $\theta_{0,1}$ are the angles made by $\mathbf{k}_{0,1}$ from the z -axis.



From our discussion above, the Bragg condition is $|\theta_0| = |\theta_1| = \theta_B$ for efficient diffraction, but we will consider this more general setup to handle angular misalignment. The two angles will still be subject to the constraint $|\theta_0| + |\theta_1| = 2\theta_B$.

Recalling that $k = k_0 \approx k_1$, we can write the phase change as

$$d\phi_{0,1} = \frac{n}{n_0} k dr_{0,1} = \frac{n}{n_0} \frac{k dz}{\cos \theta_{0,1}} = \frac{k dz}{n_0 \cos \theta_{0,1}} [n_0 + \delta n \cos(k_{\text{rf}} x - \omega_{\text{rf}} t)], \quad (13.30)$$

where $dr_{0,1}$ is the path length along $\mathbf{k}_{0,1}$ corresponding to a displacement dz along z . Then

$$\frac{\partial \phi_{0,1}}{\partial z} = \frac{k}{\cos \theta_{0,1}} + \frac{k \delta n}{2n_0 \cos \theta_{0,1}} [e^{i(k_{\text{rf}} x - \omega_{\text{rf}} t)} + e^{-i(k_{\text{rf}} x - \omega_{\text{rf}} t)}]. \quad (13.31)$$

The first term here represents free-space evolution in the absence of the acoustic wave, the second term represents coupling of diffraction order $j \rightarrow j + 1$, and the third term represents coupling of diffraction order $j \rightarrow j - 1$. Let's simplify this by defining the phase-rotation rate

$$\kappa_j := \frac{k}{\cos \theta_j} = \frac{k}{\sqrt{1 - (k_{j,x}/k)^2}} \approx k + \frac{k_{j,x}^2}{2k}, \quad (13.32) \quad (\text{phase-precession rate})$$

where the last expression holds in the paraxial approximation. We will also define the **Rabi frequency**

$$\Omega := \frac{k \delta n}{n_0 \cos \theta_j}. \quad (13.33) \quad (\text{Rabi frequency})$$

In principle the Rabi frequency is different for each wave, so we should keep track of two Rabi frequencies $\Omega_{0,1}$. However, we will assume that $|\theta_0|$ is very close to $|\theta_1|$. We will see that even small differences in the two angles have a substantial impact on the diffraction via the κ_j , but since $\delta n/n_0$ is small, any difference in the Rabi frequencies is a much smaller effect and thus ignorable.

With these definitions, the phase shift becomes

$$\frac{\partial \phi_{0,1}}{\partial z} = \kappa_{0,1} + \frac{\Omega}{2} \left[e^{i(k_{\text{rf}}x - \omega_{\text{rf}}t)} + e^{-i(k_{\text{rf}}x - \omega_{\text{rf}}t)} \right]. \quad (13.34)$$

Now let's consider the evolution of the zeroth and first diffraction orders. We'll ignore the others, since they don't satisfy the Bragg condition and thus they shouldn't be excited. We'll denote the zeroth-order wave by $E^{(+)}(\mathbf{k}_0, z)$ and the first by $E^{(+)}(\mathbf{k}_1, z)$, where the wave vectors are given by $\mathbf{k}_j = \mathbf{k}_0 + j\mathbf{k}_{\text{rf}} = \mathbf{k}_0 + jk_{\text{rf}}\hat{x}$. The total field is the sum of the two waves, and with the explicit time dependence we can write

$$E^{(+)}(z, t) = E^{(+)}(\mathbf{k}_0, z) e^{-i\omega_0 t} + E^{(+)}(\mathbf{k}_1, z) e^{-i\omega_1 t}. \quad (13.35)$$

Note that we've already included the $\exp(i\mathbf{k}_j \cdot \mathbf{r})$ in the z -dependence of $E^{(+)}(\mathbf{k}_j, z)$, so we haven't written it out explicitly here.

Then recalling that the phase evolution ϕ depends on the wave vector, the evolution equation (13.29) becomes

$$\begin{aligned} \frac{\partial}{\partial z} \left[E^{(+)}(\mathbf{k}_0, z) e^{-i\omega_0 t} + E^{(+)}(\mathbf{k}_1, z) e^{-i\omega_1 t} \right] = \\ i\kappa_0 E^{(+)}(\mathbf{k}_0, z) e^{-i\omega_0 t} + i\kappa_1 E^{(+)}(\mathbf{k}_1, z) e^{-i\omega_1 t} \\ + \frac{i\Omega}{2} E^{(+)}(\mathbf{k}_0, z) e^{ik_{\text{rf}}x} e^{-i\omega_1 t} + \frac{i\Omega}{2} E^{(+)}(\mathbf{k}_0, z) e^{-ik_{\text{rf}}x} e^{-i\omega_{-1} t} \\ + \frac{i\Omega}{2} E^{(+)}(\mathbf{k}_1, z) e^{ik_{\text{rf}}x} e^{-i\omega_2 t} + \frac{i\Omega}{2} E^{(+)}(\mathbf{k}_1, z) e^{-ik_{\text{rf}}x} e^{-i\omega_0 t}, \end{aligned} \quad (13.36)$$

where $\omega_j = \omega_0 + j\omega_{\text{rf}}$. Picking out the frequency component of this equation at ω_0 (e.g., by multiplying through by $\exp(i\omega_0 t)$ and integrating over a time interval of length $2\pi/\omega_0$) and discarding the common time dependence, we obtain the equation of motion

$$\frac{\partial}{\partial z} E^{(+)}(\mathbf{k}_0, z) = i\kappa_0 E^{(+)}(\mathbf{k}_0, z) + \frac{i\Omega}{2} E^{(+)}(\mathbf{k}_1, z) e^{-ik_{\text{rf}}x}. \quad (13.37)$$

Similarly, for the component at frequency ω_1 ,

$$\frac{\partial}{\partial z} E^{(+)}(\mathbf{k}_1, z) = i\kappa_1 E^{(+)}(\mathbf{k}_1, z) + \frac{i\Omega}{2} E^{(+)}(\mathbf{k}_0, z) e^{ik_{\text{rf}}x}. \quad (13.38)$$

In these two equations, the factors of $\exp(\pm ik_{\text{rf}}x)$ account for the momentum change on absorbing or emitting a phonon. Notice that $E^{(+)}(\mathbf{k}_j, z) \sim \exp(i\mathbf{k}_j \cdot \mathbf{r})$ and $\mathbf{k}_j = \mathbf{k}_0 + jk_{\text{rf}}\hat{x}$, so all terms in Eq. (13.37) have the

common dependence of $\exp(i\mathbf{k}_0 \cdot \mathbf{r})$, and all terms in Eq. (13.38) have the common dependence of $\exp(i\mathbf{k}_1 \cdot \mathbf{r})$. As long as we keep this in mind, we don't have to write out the explicit phase dependence, since what we're really after is the relative amplitudes of the two waves.

Thus, we've made things much simpler, reducing the wave equation to a pair of ordinary differential equations. Defining the normalized electric fields,

$$c_j(z) := \frac{E^{(+)}(\mathbf{k}_j, z)}{E_{\text{in}}^{(+)}} , \quad (13.39)$$

Eqs. (13.37) and (13.38) become

$$\begin{aligned} \partial_z c_0 &= i\kappa_0 c_0 + i\frac{\Omega}{2} c_1 \\ \partial_z c_1 &= i\kappa_1 c_1 + i\frac{\Omega}{2} c_0, \end{aligned} \quad (13.40)$$

where $\partial_z \equiv \partial/\partial z$. We can solve these coupled equations by decoupling them. Differentiating the $\partial_z c_0$ equation,

$$\partial_z^2 c_0 = i\kappa_0 \partial_z c_0 + i\frac{\Omega}{2} \partial_z c_1. \quad (13.41)$$

Using the other equation for $\partial_z c_1$,

$$\partial_z^2 c_0 = i\kappa_0 \partial_z c_0 + i\frac{\Omega}{2} (i\kappa_1 c_1) + \left(i\frac{\Omega}{2}\right)^2 c_0, \quad (13.42)$$

and using the $\partial_z c_0$ equation to eliminate c_1 , we find

$$\partial_z^2 c_0 - i(\kappa_0 + \kappa_1) \partial_z c_0 - \left(\kappa_0 \kappa_1 - \frac{\Omega^2}{4}\right) c_0 = 0. \quad (13.43)$$

Repeating this procedure yields the same equation of motion for c_1 :

$$\partial_z^2 c_1 - i(\kappa_0 + \kappa_1) \partial_z c_1 - \left(\kappa_0 \kappa_1 - \frac{\Omega^2}{4}\right) c_1 = 0. \quad (13.44)$$

Let's concentrate on obtaining an explicit solution for $c_1(z)$. We can factor Eq. (13.44) and write it as

$$\left(\partial_z - i\frac{\Omega_+}{2}\right) \left(\partial_z - i\frac{\Omega_-}{2}\right) c_1 = 0 \quad (13.45)$$

where the quadratic roots are given by

$$\Omega_{\pm} = (\kappa_0 + \kappa_1) \pm \tilde{\Omega}, \quad (13.46)$$

where

$$\tilde{\Omega} := \sqrt{(\kappa_0 - \kappa_1)^2 + \Omega^2}. \quad (13.47)$$

The general solution to Eq. (13.45) is

$$c_1(z) = A_+ e^{i\Omega_+ z/2} + A_- e^{i\Omega_- z/2}, \quad (13.48)$$

where A_{\pm} are constants to be determined by the initial conditions. Note that given the form of Ω_{\pm} , we can equivalently write

$$c_1(z) = e^{i(\kappa_0 + \kappa_1)/2} \left[A_1 \cos\left(\frac{\tilde{\Omega}z}{2}\right) + A_2 \sin\left(\frac{\tilde{\Omega}z}{2}\right) \right], \quad (13.49)$$

where $A_{1,2}$ are (other) constants to be determined by the initial conditions.

Initially, upon entering the sound field, all of the incident light is undiffracted, so $c_0(z = 0) = 1$ and $c_1(z = 0) = 0$. Thus, we can see that $A_1 = 0$. Also, from the $\partial_z c_1$ equation in Eq. (13.40) gives

$$\partial_z c_1(z = 0) = i \frac{\Omega}{2}, \quad (13.50)$$

while differentiating Eq. (13.49) gives

$$\partial_z c_1(z = 0) = \frac{\tilde{\Omega} A_2}{2}, \quad (13.51)$$

so that $A_2 = i\Omega/\tilde{\Omega}$. Thus, the solution is

$$c_1(z) = ie^{i(\kappa_0 + \kappa_1)/2} \frac{\Omega}{\tilde{\Omega}} \sin\left(\frac{\tilde{\Omega}z}{2}\right). \quad (13.52) \quad (\text{Bragg-diffraction amplitude})$$

Since we're only interested in relative amplitudes, we can obtain $|c_1|$ from the conservation constraint $|c_0|^2 + |c_1|^2 = 1$.

The intensity of the diffracted beam is given by

$$\frac{I_1}{I_{\text{in}}} = |c_1(z)|^2 = \left(\frac{\Omega}{\tilde{\Omega}}\right)^2 \sin^2\left(\frac{\tilde{\Omega}z}{2}\right), \quad (13.53) \quad (\text{Bragg-diffraction intensity})$$

where of course $I_1 + I_0 = I_{\text{in}}$. We can also write this in the form

$$\frac{I_1}{I_{\text{in}}} = |c_1(z)|^2 = \left(\frac{\Omega}{\tilde{\Omega}}\right)^2 \left(\frac{1}{2} - \frac{1}{2} \cos \tilde{\Omega}z\right). \quad (13.54) \quad (\text{Bragg-diffraction intensity})$$

In this form we can see that the intensity oscillates sinusoidally with frequency $\tilde{\Omega}$.

Now we are in a position to make a few observations about acousto-optic Bragg diffraction

1. **Rabi flopping.** The intensity oscillates between the incident and diffracted wave at rate $\tilde{\Omega}$. This behavior (oscillation of energy between two coupled oscillators) occurs often in physics, but in particular this is equivalent to a two-level atom (the two optical modes) driven by a classical laser field (the sound wave, in this case). With this analogy, the intensity oscillations are **Rabi oscillations**, and $\tilde{\Omega}$ is the **generalized Rabi frequency**.
2. **Resonance condition.** The (peak-to-peak) amplitude of the intensity oscillations from Eq. (13.53) is

$$\frac{\Omega^2}{\tilde{\Omega}^2} = \frac{\Omega^2}{\Omega^2 + (\kappa_0 - \kappa_1)^2}. \quad (13.55)$$

Since this amplitude ≤ 1 , there are *incomplete* Rabi oscillations unless the **resonance condition**

$$\kappa_0 = \kappa_1 \quad (13.56)$$

is satisfied. Putting in the definitions of the κ_j , this condition is equivalent to

$$|k_{0,x}| = |k_{1,x}|. \quad (13.57)$$

But $\Delta k_x = k_{\text{rf}}$ for first-order diffraction, so the resonance condition occurs when

$$-k_{0,x} = k_{1,x} = k_{\text{rf}}/2. \quad (13.58)$$

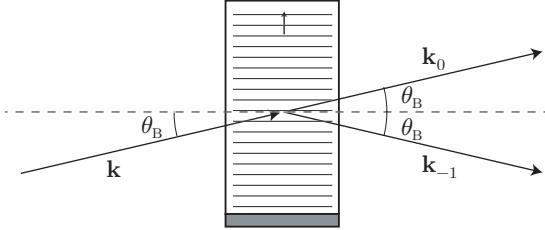
Expanding this further, of course, just leads to the Bragg condition $\theta_0 = \theta_B = \sin^{-1}(\lambda/2\lambda_{\text{rf}})$.

3. **Resonant Rabi flopping.** On resonance ($\kappa_0 = \kappa_1$), the generalized Rabi frequency $\tilde{\Omega}$ reduces to the ordinary Rabi frequency Ω . Then the diffraction efficiency becomes

$$\frac{I_1}{I_{\text{in}}} = \sin^2 \left(\frac{\Omega z}{2} \right) = \sin^2 \left(\frac{k \delta n}{2n_0 \cos \theta_B} z \right). \quad (13.59)$$

Since the index modulation amplitude δn is related to the rf drive power, we can see that for best diffraction efficiency there is an optimum rf drive power for a given crystal length.

4. **Mirror image.** We did the analysis for the +1 diffraction order, but the analysis works just as well for diffraction into the -1 order, the differences being the orientation and that the diffracted beam is downshifted in frequency instead of upshifted like the +1 order.



5. **Resonant selection.** Only one order can be on resonance (i.e., satisfy the Bragg condition) at a time. Again \mathbf{k}_{rf} is well-defined, so we need a particular input angle θ_0 . Diffraction to the j th order requires that $|k_{0,x}| = |k_{j,x}|$, which in turn requires that the zeroth and j th order diffracted waves to subtend equal angles about the normal to \mathbf{k}_{rf} .
6. **Higher-order scattering.** We only considered first-order diffraction, but higher-order Bragg diffraction is also possible by absorbing or emitting multiple phonons. The intermediate waves are not resonantly coupled, so this process proceeds exponentially more slowly (in terms of interaction length for fixed δn) than first-order diffraction. Thus, practical devices usually stick to first-order diffraction.

13.2.1 Efficiency

We can manipulate the form of the (optimally aligned) diffraction efficiency in the Bragg regime from Eq. (13.59):

$$\frac{I_1}{I_{\text{in}}} = \sin^2 \left(\frac{k \delta n}{2n_0 \cos \theta_B} z \right) = \sin^2 \left(\frac{\pi \delta n z}{n_0 \lambda \cos \theta_B} \right) = \sin^2 \left[\frac{\pi z}{\lambda \sqrt{2} \cos \theta_B} \left(\frac{\delta n \sqrt{2}}{n_0} \right) \right]. \quad (13.60)$$

Now we can translate the index modulation δn into properties of the crystal and acoustic wave. We can write

$$MI_{\text{acoustic}} := \frac{2(\delta n)^2}{n_0^2}, \quad (13.61)$$

(figure of merit)

where I_{acoustic} is the intensity of the acoustic wave and M is the **figure of merit** that characterizes the acousto-optic crystal. Thus, we can write the efficiency as

$$\frac{I_1}{I_{\text{in}}} = \sin^2 \left[\frac{\pi z}{\lambda \sqrt{2} \cos \theta_B} \sqrt{MI_{\text{acoustic}}} \right]. \quad (13.62)$$

(Bragg-diffraction efficiency)

Noting that the dimensions of the acoustic wave are the interaction length z and the crystal height h , we can write the acoustic intensity as

$$I_{\text{acoustic}} = \frac{P_{\text{acoustic}}}{zh}, \quad (13.63)$$

(acoustic intensity)

where P_{acoustic} is the total power in the acoustic wave.

13.2.2 Example: TeO₂ Modulator (Bragg Regime)

To get a feeling for the numbers, we'll look at a commercial TeO₂ ("tellurium dioxide") modulator designed to work with a 150 MHz rf drive at 780 nm (vacuum) optical wavelength. For TeO₂, the figure of merit is $M = 34.5 \times 10^{-15} \text{ s}^3/\text{kg}$, the acoustic-wave velocity is $v_{\text{rf}} = 4.26 \text{ km/s}$, and the refractive index is $n_0 = 2.35$. The crystal dimensions are $z = 15 \text{ mm}$ thickness and $h = 1 \text{ mm}$ height. We will use the paraxial approximation throughout.

The Bragg angle is

$$\theta_B \approx \frac{\lambda \omega_{\text{rf}}}{4\pi v_{\text{rf}}} = 14 \text{ mrad} = 0.79^\circ. \quad (13.64)$$

The deflection angle of the diffracted beam is $2\theta_B = 1.6^\circ$. Recall that all geometrical quantities refer to their values inside the crystal, so after exiting the crystal, the deflection angle is 3.7° .

Maximum diffraction efficiency occurs when

$$\frac{I_1}{I_{\text{in}}} = \sin^2(\pi/2). \quad (13.65)$$

Since $\cos \theta_B \approx 1$, we can use Eq. (13.62) to write an expression for the optimal acoustic intensity,

$$I_{\text{acoustic, optimum}} = \frac{\lambda^2}{2z^2 M}, \quad (13.66)$$

or equivalently for the optimal power

$$P_{\text{acoustic, optimum}} = \frac{\lambda^2 h}{2zM} = 0.1 \text{ W}. \quad (13.67)$$

The actual power applied to the piezo for the best diffraction efficiency is actually 0.8 W, which is of the same order of magnitude, but implies that much power is wasted in the piezo transducer in converting electrical to acoustic energy. We can also use Eq. (13.60) to see that at maximum diffraction efficiency, the index modulation amplitude is given by

$$\frac{\delta n}{n_0} \approx \frac{\lambda}{2z} = 1.1 \times 10^{-5}. \quad (13.68)$$

This is a very small index change, but its effect builds up coherently over the optical path length of the modulator.

13.3 Borderline

We have discussed the operation of acousto-optic modulators in the Raman–Nath and Bragg regimes. But how do we distinguish which regime applies to a particular device? Recall that what distinguishes the two regimes is whether or not the direction of \mathbf{k}_{rf} is well-defined, which says whether or not diffraction to many orders is allowed. One approach is to examine the Klein–Cook parameter¹

$$Q := \frac{(\text{angular spread scale of diffracted beams})}{(\text{angular spread of rf phonons})}, \quad (13.69)$$

which we can evaluate as

$$\begin{aligned} Q &= \frac{(2\theta_B)}{(1/k_{\text{rf}} z)} \\ &\approx \frac{(k_{\text{rf}}/k)}{1/k_{\text{rf}} z} \\ &= \frac{k_{\text{rf}}^2 z}{k}, \end{aligned} \quad (13.70)$$

¹W. Klein and B. Cook, "Unified approach to Ultrasonic Light Diffraction," *IEEE Transaction on Sonics and Ultrasonics*, SU-14, 123 (1967).

where we used the paraxial form for θ_B , and finally if again v_{rf} is the phase velocity of the acoustic wave in the medium,

$$Q = \frac{\lambda\omega_{rf}^2 z}{2\pi v_{rf}^2}. \quad (13.71)$$

(Klein–Cook parameter)

Then we can distinguish the two regimes as follows:

1. **Raman–Nath** (large angular spread of phonons): $Q \ll 1$
2. **Bragg** (small angular spread of phonons): $Q \gg 1$

The borderline is usually taken to be from 1 to 2π , which is some intermediate regime where neither treatment is really appropriate. However, $Q = 1$ is a practical dividing line between the two regimes. For the TeO_2 modulator in the previous example, $Q = 39$, so this device is clearly in the Bragg regime.

13.4 Exercises

Problem 13.1

A *phase-modulated* optical wave has the form

$$E^{(+)}(x, t) = E_0^{(+)} e^{i[kx - \omega t + \delta\phi \sin(\omega_{\text{mod}} t)]}, \quad (13.72)$$

where ω_{mod} is the modulation frequency. Such a wave could result, for example, by running the wave through an *electro-optic crystal*, whose refractive index is modulated by an applied ac signal with frequency ω_{mod} .

- (a) For a wave with time dependence $\exp[-i\phi(t)]$, we can define the *instantaneous frequency* as

$$\omega_{\text{inst}} := \frac{d\phi}{dt}. \quad (13.73)$$

Compute the instantaneous frequency of the phase-modulated wave and thus show that the frequency oscillates about ω . That is, phase modulation is in some sense the same as frequency modulation.

- (b) Write the phase-modulated wave as a sum of plane waves, with the general form

$$\sum_{j=-\infty}^{\infty} c_j e^{i(kx - \omega_j t)}. \quad (13.74)$$

Hint: start by writing down a Bessel series for the function $\exp(iK \sin x)$.

- (c) From your answer to (b), argue that the intensity spectrum (as viewed through a Fabry–Perot spectrum analyzer) consists of a series of peaks with relative intensity $J_j^2(\delta\phi)$. You may assume the response of the Fabry–Perot analyzer is slow compared to the modulation frequency. This phase-modulation technique is commonly used in the laboratory to shift the frequency of a laser or to generate multiple laser frequencies.

Problem 13.2

Suppose you observe that for a particular acousto-optic modulator in the Bragg regime, 780 nm light (wavelength in air) diffracts efficiently into the first order, with an angle of 2° between the undiffracted and first-order beams. The most efficient diffraction is observed for 0.8 W of input rf power.

- (a) If 850 nm light is used instead, what is the new diffraction angle?
 (b) What is the rf power required for most efficient diffraction at the new wavelength?
 (c) The coupled-wave theory of Bragg diffraction predicts that the best diffraction efficiency is 100%. However, in practice the best efficiencies are more like 90%, so obviously we made some assumptions that aren't quite true. Give an example of something we didn't take into account that could explain the lower practical efficiencies of acousto-optic modulators.

Problem 13.3

In the Bragg regime, the optimum diffraction efficiency occurs when the incident angle is θ_B from the normal to the acoustic-wave propagation direction.

- (a) Show that for small θ_B , an angular misalignment of

$$\delta\theta = \frac{\delta n}{2n_0\theta_B} \quad (13.75)$$

causes the diffraction efficiency to drop by a factor of 2. You can use the following general outline:

First, note that the best diffraction efficiency is given by $\Omega^2/\tilde{\Omega}^2$, and this is the part of the efficiency that is sensitive to angle. This drops to 1/2 when $(\kappa_0 - \kappa_1)^2 = \Omega^2$. Now draw out the diffraction geometry

for the misaligned case, noting that the total diffracted angle is still $2\theta_B$, so that $|k_{1,x}/k| \approx |\theta_B - \delta\theta|$ and $|k_{0,x}/k| \approx |\theta_B + \delta\theta|$. Use these to eliminate $\kappa_{0,1}$ and obtain the result above.

- (b) For the 150 MHz TeO₂ modulator that we discussed as an example in class, what is $\delta\theta$?

Problem 13.4

Consider an acousto-optic modulator operating in the Raman-Nath regime at $\lambda_0 = 780$ nm (vacuum) wavelength. Assume the modulator is made from TeO₂.

- (a) If the modulator is 80 MHz, what is the maximum thickness of the crystal in the direction of beam propagation? Assume the beam is normal to the acoustic wave.
- (b) What is the maximum achievable relative intensity in any diffracted order (other than the zeroth)? This is why Bragg modulators are potentially much more efficient at diffracting light into a single order. Assuming $z = 0.1$ mm, what would the index modulation $\delta n/n_0$ need to be to achieve this maximum intensity?
- (c) What would the index modulation $\delta n/n_0$ need to be to completely deplete the input beam (zeroth order)?

For parts (b) and (c) you will need to consult a table of Bessel functions or use some mathematical software.

Problem 13.5

Consider a Gaussian beam focused to a waist w_0 in a Bragg-mode acousto-optic modulator. What is the turn-on time of the diffracted beam, assuming that the piezo transducer is suddenly switched on? Assume that the only contribution to the rise time is due to the finite propagation velocity of the acoustic wave front, and ignore any distortion of the wave front due to dispersive propagation in the crystal. For concreteness, define the rise time as the time it takes the diffracted beam to go from 10% to 90% of its maximum intensity.

Problem 13.6

A glass acousto-optic modulator ($n_0 \approx 1.5$) diffracts He–Ne laser light (vacuum $\lambda \approx 200\pi$ nm). Suppose that you observe that the angle between the zeroth and first diffracted order is 3° (obviously, you are measuring this angle *outside* the modulator). However, some diabolical fiend has arranged it such that you cannot tell if any other diffracted orders are present. In order to save the world, you must figure out whether or not the modulator diffracts substantial power into other orders. How did you get yourself into this situation, anyway?

- (a) What is the Bragg angle θ_B ?
- (b) Under what conditions would you expect there to be many diffracted beams? Give a rough range of optical path lengths of the crystal that would lead to many diffracted beams. (You need not compute any particular parameter, just go through the reasoning for what distinguishes the Raman–Nath and Bragg regimes.)
- (c) Suppose that you observe that

$$\frac{I_0}{I_{\text{in}}} \approx J_0^2(17), \quad (13.76)$$

where I_0 is the the intensity of the zeroth diffraction order and I_{in} is the input intensity. Estimate the number of diffracted beams you would expect. Assume that the modulator operates in the regime where there are many diffracted peaks.

Problem 13.7

Describe how you can use an acousto-optic modulator as a spectrum analyzer for an rf signal. Also

describe how set up the analyzer to maximize the frequency resolution. You may assume that any diffraction angles are small, and that the acoustic-wave velocity is independent of frequency.

Chapter 14

Coherence

When considering interference of monochromatic waves in Chapter 5, we made the comment that assuming a monochromatic wave with time dependence of the form $\exp(-i\omega t)$ is implicitly computing a Fourier transform. Now that we have the proper mathematical machinery in place, we will explore this idea in more detail.

Let's now consider an interference experiment that involves a range of frequencies. Recalling the interference of two monochromatic plane waves, we can write the superposition of the two waves as

$$E_{\text{sum}}^{(+)}(\mathbf{r}, t) = E_{10}^{(+)} e^{i(kz - \omega t)} + E_{20}^{(+)} e^{i(kz - \omega t)} e^{i\phi}. \quad (14.1)$$

Here, ϕ is a relative phase difference between the two waves. Recall that we are just writing the positive-frequency components, so that the physical fields also must include the negative-frequency parts.

The intensity of the superposition is

$$I_{\text{sum}} = \langle |\mathbf{E}_{\text{sum}} \times \mathbf{H}_{\text{sum}}| \rangle = \frac{1}{\eta} \left\langle (E_{\text{sum}})^2 \right\rangle. \quad (14.2)$$

Writing this out explicitly in terms of the component fields,

$$I_{\text{sum}} = \frac{1}{\eta} \left\langle \left(E_{10}^{(+)} e^{i(kz - \omega t)} + E_{20}^{(+)} e^{i(kz - \omega t)} e^{i\phi} + E_{10}^{(-)} e^{-i(kz - \omega t)} + E_{20}^{(-)} e^{-i(kz - \omega t)} e^{-i\phi} \right)^2 \right\rangle. \quad (14.3)$$

The optical terms of the form $\exp(\pm i2\omega t)$ vanish in the time average, so we obtain

$$\begin{aligned} I_{\text{sum}} &= \frac{2}{\eta} E_{10}^{(-)} E_{10}^{(+)} + \frac{2}{\eta} E_{20}^{(-)} E_{20}^{(+)} + \frac{2}{\eta} E_{10}^{(-)} E_{20}^{(+)} e^{i\phi} + \frac{2}{\eta} E_{20}^{(-)} E_{10}^{(+)} e^{-i\phi} \\ &= I_1 + I_2 + \left[\frac{2}{\eta} E_{10}^{(-)} E_{20}^{(+)} e^{i\phi} + \text{c.c.} \right]. \end{aligned} \quad (14.4)$$

Again, the interference is in the terms with the relative phase ϕ .

Suppose that the phase difference represents a difference in optical path length, in the form of a time delay τ . Then $\phi = -\omega\tau$, and so

$$I_{\text{sum}} = I_1 + I_2 + \left[\frac{2}{\eta} E_{10}^{(-)} E_{20}^{(+)} e^{-i\omega\tau} + \text{c.c.} \right]. \quad (14.5)$$

Now let's handle the case of multiple frequencies. To simplify things, we'll assume that the two waves have equal amplitude and come from a common source. Then the intensity *density* at frequency ω is

$$I_{\text{sum}}(\omega) = 2I(\omega) + \left[\frac{2}{\eta} |E_0^{(+)}(\omega)|^2 e^{-i\omega\tau} + \text{c.c.} \right]. \quad (14.6)$$

Note that the notation here is a little funny; the frequency-dependent quantities $I(\omega)$ and $E^{(+)}(\omega)$ don't have the respective dimensions of intensity and electric field; rather, $I(\omega) d\omega$ and $|E^{(+)}(\omega)|^2 d\omega$ are the intensity and (squared) electric field, respectively, in the frequency interval between ω and $\omega + d\omega$.

Now we can sum over all frequencies to find the total intensity:

$$\begin{aligned} I_{\text{total}} &= \int_0^\infty I_{\text{sum}}(\omega) d\omega \\ &= 2 \int_0^\infty I(\omega) d\omega + \left[\frac{2}{\eta} \int_0^\infty |E_0^{(+)}(\omega)|^2 e^{-i\omega\tau} d\omega + \text{c.c.} \right]. \end{aligned} \quad (14.7)$$

Note that the frequency integral ranges only over *positive* frequencies; we've already accounted for the *negative* frequencies by including the complex conjugate terms. Thus the intensity spectrum $I_{\text{sum}}(\omega)$ is a *one-sided* spectrum, which is common when working with intensities and powers. We can now recognize the second integral in the last expression as a “one-sided” Fourier transform of $|E_0^{(+)}(\omega)|^2$, where we recall the normalization convention for ω - t Fourier transforms:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^\infty \tilde{f}(\omega) e^{-i\omega t} d\omega, \quad \tilde{f}(\omega) = \int_{-\infty}^\infty f(t) e^{i\omega t} dt. \quad (14.8) \quad (\omega\text{-}t \text{ Fourier transform})$$

But what is this, when we don't know the general form of the electric field?

14.1 Wiener–Khinchin Theorem

Recall the convolution theorem for functions f and g (see Section 11.3):

$$\mathcal{F}[f * g] = \mathcal{F}[f]\mathcal{F}[g]. \quad (14.9)$$

Writing out the convolution integral explicitly,

$$(f * g)(t) = \int_{-\infty}^\infty f(t') g(t - t') dt' = \mathcal{F}^{-1}[\mathcal{F}[f]\mathcal{F}[g]]. \quad (14.10)$$

If we make the particular choice $g(t) = f^*(-t)$, then

$$\mathcal{F}[g(t)] = \mathcal{F}[f^*(-t)] = \int_{-\infty}^\infty f^*(-t) e^{i\omega t} dt = \int_{-\infty}^\infty f^*(t) e^{-i\omega t} dt = (\mathcal{F}[f(t)])^*. \quad (14.11)$$

Thus, Eq. (14.10) becomes

$$\int_{-\infty}^\infty f(t') f^*(t' - t) dt' = \mathcal{F}^{-1}\left[\left|\mathcal{F}[f]\right|^2\right]. \quad (14.12)$$

Inverting the transform and letting $t' \rightarrow t' + t$, we obtain the **Wiener–Khinchin theorem**:

$$\mathcal{F}\left[\int_{-\infty}^\infty f^*(t') f(t + t') dt'\right] = |\mathcal{F}[f]|^2. \quad (14.13) \quad (\text{Wiener–Khinchin theorem})$$

The function on the left-hand side,

$$\int_{-\infty}^\infty f^*(t) f(t + \tau) dt, \quad (14.14) \quad (\text{autocorrelation function})$$

is the **autocorrelation function** of $f(t)$. Essentially, it compares f to itself but shifted by an amount τ by computing an overlap integral. We can understand the right-hand side by noting that $\mathcal{F}[f]$ is the *spectrum* of f , and so $|\mathcal{F}[f]|^2$ is the **energy spectral density** of f . Essentially, the energy spectrum is the square of the usual spectrum, with the phase information removed. This is consistent with the notion of energy going

as the square of a signal amplitude. Thus, the Wiener–Khinchin theorem states that the Fourier transform of the autocorrelation function gives the energy spectrum.

There is one subtle point to these definitions: for some signals, such as steady optical signals, the correlation integral diverges:

$$\int_{-\infty}^{\infty} f^*(t) f(t + \tau) dt \rightarrow \infty. \quad (14.15)$$

In this case, we should consider a time average instead of the normal integral. For an averaging time of T ,

$$\langle f^*(t) f(t + \tau) \rangle_T := \frac{1}{T} \int_{-T/2}^{T/2} f^*(t) f(t + \tau) dt. \quad (14.16)$$

For bounded signals, this integral is guaranteed to converge. To be physically sensible, T should be a suitably long observation time (e.g., long enough to resolve the frequency spectrum). For such signals, we can write the Wiener–Khinchin theorem as

$$\mathcal{F}[\langle f^*(t) f(t + \tau) \rangle_T] = \mathcal{F}^*[f] \mathcal{F}_T[f]. \quad (14.17)$$

Here we have defined a finite-time Fourier transform by

$$\mathcal{F}_T[f] := \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{i\omega t} dt. \quad (14.18)$$

Defining this seems a bit funny, but it avoids problems with singular spectra. Now the Wiener–Khinchin theorem says that the Fourier transform of the (time-averaged) correlation function is the **power spectral density**, or the energy spectral density per unit time. For a *stationary* process, the correlation function is independent of t (generally for a sufficiently long averaging time T). Then we can extend the averaging time $T \rightarrow \infty$. Denoting this long-time averaging limit as

$$\langle f^*(t) f(t + \tau) \rangle = \langle f^*(t) f(t + \tau) \rangle_{T \rightarrow \infty}, \quad (14.19)$$

we can thus write the Wiener–Khinchin theorem as

$$\mathcal{F}[\langle f^*(t) f(t + \tau) \rangle] = \lim_{T \rightarrow \infty} \frac{1}{T} \left| \int_{-T/2}^{T/2} f(t) e^{i\omega t} dt \right|^2. \quad (\text{Wiener–Khinchin theorem, time-average form}) \quad (14.20)$$

Again, the right-hand side is the power spectral density, and in this form it is more clear that this is the energy density per unit time.

14.2 Optical Wiener–Khinchin Theorem

In terms of stationary optical fields, the Wiener–Khinchin theorem (14.20) becomes

$$\int_0^\infty I(\omega) e^{-i\omega\tau} d\omega = \frac{2}{\eta} \langle E^{(-)}(t) E^{(+)}(t + \tau) \rangle. \quad (\text{optical Wiener–Khinchin theorem}) \quad (14.21)$$

This is because the intensity density $I(\omega) = \langle |\mathbf{S}| \rangle$ is the time-averaged power spectral density of the optical field. Note that from the inverse Fourier transform, there is conventionally a factor of $1/2\pi$, but it is missing here because it is already implicitly included in $I(\omega)$. We can see this from the boundary condition at $\tau = 0$, which gives the total intensity:

$$\int_0^\infty I(\omega) d\omega = \frac{2}{\eta} \langle E^{(-)}(t) E^{(+)}(t) \rangle = \frac{2}{\eta} \langle |E^{(+)}(t)|^2 \rangle. \quad (14.22)$$

In optics, as in statistics and other fields, it is conventional to define a normalized correlation function

$$g^{(1)}(\tau) := \frac{\langle E^{(-)}(t)E^{(+)}(t+\tau) \rangle}{\langle E^{(-)}(t)E^{(+)}(t) \rangle}, \quad (\text{degree of first-order temporal coherence}) \quad (14.23)$$

so that

$$\frac{2}{\eta} \langle E^{(-)}(t)E^{(+)}(t+\tau) \rangle = \left(\int_0^\infty I(\omega) d\omega \right) g^{(1)}(\tau). \quad (14.24)$$

The normalization is such that $g^{(1)}(\tau = 0) = 1$. (Also, the “(1)” superscript is a notation that indicates this is the first in a hierarchy of increasingly complex correlation functions that we won’t consider here.) That is, a correlation value of unity indicates perfect correlation. Note that we could just as well have written the correlation function as $\langle E^{(+)}(t+\tau)E^{(-)}(t) \rangle$, but it turns out that the order of $E^{(+)}$ and $E^{(-)}$ matters in quantum mechanics, so we’ll be careful to stick to the form in Eq. (14.23). Notice also that from the definition of $g^{(1)}(\tau)$, the correlation function is subject to the constraint

$$g^{(1)}(-\tau) = [g^{(1)}(\tau)]^*. \quad (\text{time symmetry of correlation function}) \quad (14.25)$$

In optics, $g^{(1)}(\tau)$ is called the **degree of first-order temporal coherence**. The light is said to be **coherent** if $|g^{(1)}(\tau)|^2 = 1$, **incoherent** if $|g^{(1)}(\tau)|^2 = 0$, and **partially coherent** otherwise (for $\tau \neq 0$).

Returning to the interference result of Eq. (14.7), we find

$$\begin{aligned} I_{\text{total}} &= 2 \int_0^\infty I_{\text{sum}}(\omega) d\omega + \left[\frac{2}{\eta} \int_0^\infty |E_0^{(-)}(\omega)|^2 e^{-i\omega\tau} d\omega + \text{c.c.} \right] \\ &= \left(\int_0^\infty I(\omega) d\omega \right) \left\{ 2 + \left[g^{(1)}(\tau) + \text{c.c.} \right] \right\}, \end{aligned} \quad (14.26)$$

and thus

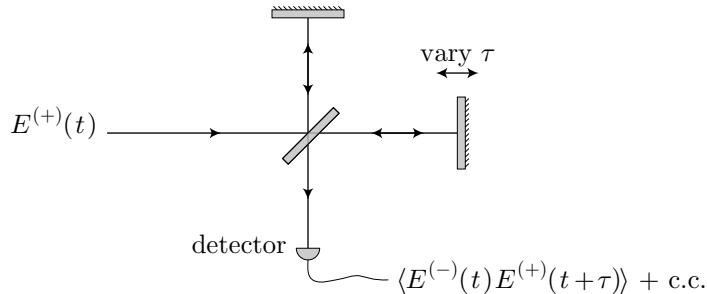
$$I_{\text{total}} = 2 \left(\int_0^\infty I(\omega) d\omega \right) \left\{ 1 + \text{Re} [g^{(1)}(\tau)] \right\}.$$

(interferometer signal in terms of $g^{(1)}(\tau)$) (14.27)

It is worth keeping in mind that the form (14.13) is still relevant for pulsed fields, in which case the result here is modified so that the coefficient in front is the integrated *power* rather than intensity.

14.2.1 Application: FTIR Spectroscopy and the Michelson Interferometer

One example of where the optical Wiener–Khinchin theorem is very useful is in **Fourier-transform infrared (FTIR) spectroscopy**. The idea in FTIR spectroscopy is to use a Michelson interferometer to monitor a stationary input signal $E^{(+)}(t)$. The interferometer splits and interferes the beam with itself, and the path-length difference τ varies with the displacement of one of the mirrors.



The photodetector at the output measures the time average of the product of the fields, up to an overall constant, and thus measures $g^{(1)}(\tau)$, just as in Eq. (14.27). The idea is then to digitize the output of the

photodetector as a function of the mirror displacement, effectively recording $\text{Re}[g^{(1)}(\tau)]$. Computing the Fourier transform of the signal on the computer gives the spectrum $I(\omega)$. (Note that $I(\omega)$ is real, and so the imaginary parts of $g^{(1)}(\tau)$ don't contribute to the spectrum.) This technique is common in molecular spectroscopy, because it turns out that obtaining a spectrum this way gives better rejection of thermal detector noise compared to, say, a recording the same spectrum with the same detector on a diffraction-grating spectrometer.

14.2.2 Example: Monochromatic Light

As a simple example, let's compute the correlation function for monochromatic light and verify that the Wiener-Khinchin theorem makes sense. Let's take a monochromatic wave of the form

$$E^{(+)}(t) = E_0^{(+)} e^{-i\omega_0 t}. \quad (14.28)$$

The correlation function is

$$\langle E^{(-)}(t)E^{(+)}(t + \tau) \rangle = |E_0^{(+)}|^2 e^{-i\omega_0 \tau}. \quad (14.29)$$

In normalized form, this function becomes

$$g^{(1)}(\tau) = e^{-i\omega_0 \tau}. \quad (14.30)$$

Thus, a wave with harmonic time dependence leads to a harmonic correlation function. This is true *independent* of the phase of the input field; that is, the correlation function does not reflect any extra phase in $E_0^{(+)}$.

The power spectrum is easy to calculate via the Wiener-Khinchin theorem:

$$\mathcal{F} \left[\langle E^{(-)}(t)E^{(+)}(t + \tau) \rangle \right] = |E_0^{(+)}|^2 \mathcal{F} [e^{-i\omega_0 \tau}] = |E_0^{(+)}|^2 \delta(\omega - \omega_0). \quad (14.31)$$

Again, the harmonic time dependence produces a single frequency in the power spectrum. Of course, with our convention of positive and negative frequencies, there would be a matching $\delta(\omega + \omega_0)$ component, but this is already included in the *one-sided* spectrum $I(\omega)$.

Let's now compute the power spectrum directly, using Eq. (14.17). The normal spectrum is

$$\mathcal{F} [E^{(+)}(t)] = E_0^{(+)} \delta(\omega - \omega_0). \quad (14.32)$$

The finite-time transform is

$$\mathcal{F}_T [E^{(+)}(t)] = \frac{1}{T} \int_{-T/2}^{T/2} E_0^{(+)} e^{i(\omega - \omega_0)t} dt = E_0^{(+)} \text{sinc}[(\omega - \omega_0)T/2]. \quad (14.33)$$

Note that the value of the sinc function is 1 at $\omega = \omega_0$. Thus, the spectrum is

$$\mathcal{F}^* [E^{(+)}(t)] \mathcal{F}_T [E^{(+)}(t)] = E_0^{(-)} \delta(\omega - \omega_0) E_0^{(+)} \text{sinc}[(\omega - \omega_0)T/2] = |E_0^{(+)}|^2 \delta(\omega - \omega_0). \quad (14.34)$$

This result is consistent with Eq. (14.31). Note that without being careful with finite-time transforms, we would run into something bad involving the square of a δ -function.

14.2.3 Normalized One- and Two-Sided Spectra

Now we will be a bit more precise about the nature of the spectrum, to avoid confusion with different possible conventions for the power spectrum. First, it's convenient to define a *normalized* spectral density (lineshape function)

$$s(\omega) := \frac{I(\omega)}{\int_0^\infty I(\omega) d\omega}. \quad (14.35)$$

(normalized spectral density)

Note that this spectrum extends only over *positive* frequencies, as the intensity spectrum $I(\omega)$ corresponded to *physical* frequency components. Examining the inverse Fourier transform, we combine Eqs. (14.21) and (14.24) to obtain

$$g^{(1)}(\tau) = \int_0^\infty s(\omega) e^{-i\omega\tau} d\omega.$$

(first-order coherence in terms of normalized, one-sided spectral density) (14.36)

However, the usual inverse Fourier transform has an integral extending over both positive and negative frequencies. We can obtain something of this form by considering the *real* part of the correlation function,

$$\begin{aligned} \operatorname{Re}[g^{(1)}(\tau)] &= \frac{g^{(1)}(\tau) + [g^{(1)}(\tau)]^*}{2} = \frac{g^{(1)}(\tau) + g^{(1)}(-\tau)}{2} \\ &= \frac{1}{2} \int_0^\infty s(\omega) e^{-i\omega\tau} d\omega + \frac{1}{2} \int_0^\infty s(\omega) e^{i\omega\tau} d\omega, \end{aligned} \quad (14.37)$$

so that if we define a **two-sided spectrum** $s_{\leftrightarrow}(\omega)$ (for all $\omega \in \mathbb{R}$) via

$$s_{\leftrightarrow}(\omega) := \begin{cases} s(\omega)/2 & (\omega > 0) \\ s(\omega) & (\omega = 0) \\ s(-\omega)/2 & (\omega < 0), \end{cases}$$

(two-sided spectrum in terms of one-sided spectrum) (14.38)

which satisfies

$$s_{\leftrightarrow}(-\omega) = s_{\leftrightarrow}(\omega), \quad (14.39)$$

(symmetry of two-sided spectrum)

we obtain

$$\operatorname{Re}[g^{(1)}(\tau)] = \int_{-\infty}^\infty s_{\leftrightarrow}(\omega) e^{-i\omega\tau} d\omega.$$

(first-order coherence in terms of two-sided spectrum) (14.40)

We have lost the imaginary part, but *only* the real part can contribute to a physical result: $g^{(1)}(\tau)$ must always be accompanied by its conjugate, as we saw in the interference experiment.

Inverting this last relation, we have

$$\begin{aligned} s_{\leftrightarrow}(\omega) &= \frac{1}{2\pi} \int_{-\infty}^\infty \operatorname{Re}[g^{(1)}(\tau)] e^{i\omega\tau} d\tau \\ &= \frac{1}{2\pi} \int_{-\infty}^\infty g^{(1)}(\tau) \cos \omega\tau d\tau \\ &= \frac{1}{2\pi} \int_0^\infty g^{(1)}(\tau) \cos \omega\tau d\tau + \text{c.c.} \quad (\omega \in \mathbb{R}). \end{aligned}$$

(two-sided spectral density in terms of first-order coherence) (14.41)

The one-sided spectrum can then be written in terms of the double-sided spectrum as

$$s(\omega) = \begin{cases} s_{\leftrightarrow}(\omega) + s_{\leftrightarrow}(-\omega) = 2s_{\leftrightarrow}(\omega) & (\omega > 0) \\ s_{\leftrightarrow}(\omega) & (\omega = 0) \\ 0 & (\omega < 0), \end{cases}$$

(two-sided spectrum in terms of one-sided spectrum) (14.42)

so that the one-sided spectrum simply concentrates all the power at positive and negative frequencies on the positive side. Of course, this cannot be done for *any* spectrum, but the power spectral density—the Fourier transform of the coherence function—has a special structure due to the symmetry of the correlation

function. Thus, we can write the one-sided spectrum as

$$\begin{aligned} s(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \operatorname{Re} [g^{(1)}(\tau)] e^{i\omega\tau} d\tau \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} g^{(1)}(\tau) \cos \omega\tau d\tau \\ &= \frac{1}{\pi} \int_0^{\infty} g^{(1)}(\tau) \cos \omega\tau d\tau + \text{c.c.} \quad (\omega \geq 0) \end{aligned}$$

(one-sided spectral density in terms of first-order coherence) (14.43)

in terms of the coherence function. Combining Eqs. (14.40) and (14.42), we then find

$$\operatorname{Re} [g^{(1)}(\tau)] = \int_0^{\infty} s(\omega) \cos \omega\tau d\omega$$

(first-order coherence in terms of one-sided spectrum) (14.44)

for the reverse transformation.

Finally, note that it can be more convenient to associate spectra separately with $g^{(1)}(\tau)$ and its conjugate, by defining

$$s_{\pm}(\omega) := \frac{1}{4\pi} \int_{-\infty}^{\infty} g^{(1)}(\tau) e^{\pm i\omega\tau} d\tau = \frac{1}{4\pi} \int_0^{\infty} g^{(1)}(\tau) e^{\pm i\omega\tau} d\tau + \text{c.c.} \quad (\omega \in \mathbb{R}),$$

(component spectral densities in terms of first-order coherence) (14.45)

so that

$$s_+(\omega) = s_-(-\omega). \quad (14.46)$$

(symmetry of component spectra)

Then the relations

$$\begin{aligned} s_{\leftrightarrow}(\omega) &= s_+(\omega) + s_-(\omega) = s_+(\omega) + s_+(-\omega) \quad (\omega \in \mathbb{R}) \\ s(\omega) &= 2[s_+(\omega) + s_-(\omega)] = 2[s_+(\omega) + s_+(-\omega)] \quad (\omega \geq 0) \end{aligned}$$

(component spectral densities in terms of first-order coherence) (14.47)

recover the total spectra (14.41) and (14.43), respectively.

14.3 Visibility

The example in Section 14.2.2 of the correlation function for a monochromatic wave is special in the sense that the correlation function does not decay with τ . This is important, because the correlation function is the magnitude of the interference terms. To quantify this better, we can define the *fringe visibility* as

$$\mathcal{V} := \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \quad (14.48)$$

(fringe visibility)

where I_{\max} and I_{\min} are the maximum and minimum intensities achieved for phase variations on the order of several π . For example, complete interference results in intensity variation from 0 to some maximum, and so $\mathcal{V} = 1$. For no interference, the intensity does not vary with phase and so $\mathcal{V} = 0$. Partial coherence is represented by intermediate values of the visibility.

Writing out the explicit phase of the correlation function,

$$g^{(1)}(\tau) = |g^{(1)}(\tau)|e^{i\phi(\tau)}, \quad (14.49)$$

and so Eq. (14.27) becomes

$$I_{\text{total}} = 2 \left(\int_0^{\infty} I(\omega) d\omega \right) \left[1 + |g^{(1)}(\tau)| \cos \phi(\tau) \right]. \quad (14.50)$$

The cosine varies from -1 to 1 , so the visibility is just the magnitude of the correlation function:

$$\mathcal{V} = |g^{(1)}(\tau)|. \quad (14.51)$$

(visibility in terms of coherence)

For monochromatic light, $\mathcal{V} = 1$. Actually, this is only true if the amplitudes of the input waves are equal. For two monochromatic waves of *unequal* amplitude, the visibility becomes

$$\mathcal{V} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}. \quad (14.52)$$

(monochromatic, unbalanced interference)

The important thing to note is that for interference of monochromatic light, the visibility is independent of τ .

14.4 Coherence Time, Coherence Length, and Uncertainty Measures

As we have just seen, a peculiarity of monochromatic light is that the coherence does not decay with τ . In the generic case, where light is composed of a range of frequencies, the visibility drops as the phase difference increases, since $g^{(1)}(\tau) \rightarrow 0$ as $\tau \rightarrow \infty$. Intuitively, this is because as light waves with different wavelengths propagate, the monochromatic components tend to dephase. Mathematically, we can express this as an uncertainty relationship. Recall from Section 14.2.3 that the coherence $\text{Re}[g^{(1)}(\tau)]$ and the normalized (two-sided) spectral density (lineshape function) form a Fourier-transform pair. Actually, $2\pi s_{\leftrightarrow}(\omega)$ is the second half of the pair, due to the form of the transforms (14.40) and (14.41): recall that in the ω - t convention, there is normally a factor of $1/2\pi$ in front of the inverse-transform integral, as in Eqs. (14.8).

Thus, if we define the root-mean-square (rms) widths of these two functions, regarding $\text{Re}[g^{(1)}(\tau)]^2$ and $[s_{\leftrightarrow}(\omega)]^2$ as (unnormalized) probability distributions, we can write down an **uncertainty relation**, as in quantum mechanics:

$$\delta\omega_{\text{rms}} \delta\tau_{\text{rms}} \geq \frac{1}{2}. \quad (14.53)$$

(rms uncertainty relation)

For many distributions, we can say that the equality is more or less satisfied to

$$\delta\omega_{\text{rms}} \delta\tau_{\text{rms}} \sim \frac{1}{2}, \quad (14.54)$$

and thus see that the “widths” of $g^{(1)}(\tau)$ and $s_{\leftrightarrow}(\omega)$ are inversely related. The problem is that for some useful functions in optics, these uncertainties can diverge (e.g., for Lorentzian functions) or vary by orders of magnitude. The uncertainty inequality is always satisfied, of course, but as a practical relation for the temporal and frequency widths, this is less useful.

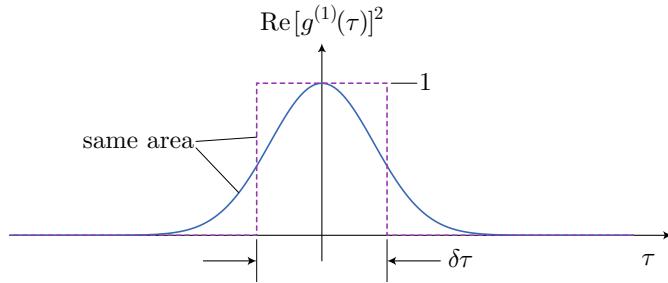
Thus we will adopt other uncertainty conventions in time and frequency¹ We can define the **coherence time** as the power-equivalent width of the correlation function:

$$\delta\tau := \int_{-\infty}^{\infty} \left| \text{Re} [g^{(1)}(\tau)] \right|^2 d\tau. \quad (14.55)$$

(coherence time, power-equivalent width)

The idea here is to first note that $g^{(1)}(\tau = 0) = 1$. Thus the width $\delta\tau$ is the width of a box of unit height, with the same area as $|\text{Re}[g^{(1)}(\tau)]|^2$. That is, $\delta\tau$ is the width of a unit-height box signal with the same *power* as $|\text{Re}[g^{(1)}(\tau)]|^2$.

¹as in Max Born and Emil Wolf, *Principles of Optics*, 7th (expanded) ed. (Cambridge, 1999).



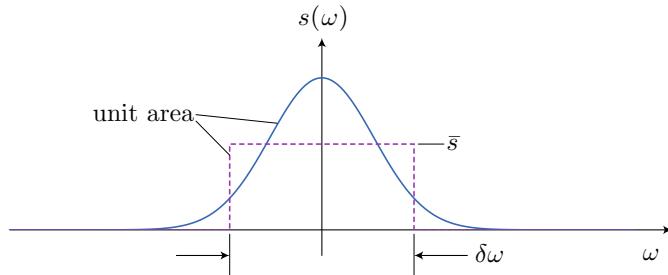
We'll take a slightly different convention for the width of the frequency spectrum. Let's define the *average value of the one-sided normalized spectrum* as

$$\bar{s} = \int_0^\infty s^2(\omega) d\omega = 2 \int_{-\infty}^\infty s_{\leftrightarrow}^2(\omega) d\omega. \quad (14.56)$$

That is, we're regarding $s(\omega)$ as a probability distribution (because it has the appropriate normalization), and then we're calculating the expectation value of $s(\omega)$ with respect to itself. Then we'll define the effective frequency width by the reciprocal of the average value:

$$\delta\omega := (\bar{s})^{-1} = \left[\int_0^\infty s^2(\omega) d\omega \right]^{-1}. \quad (14.57) \quad (\text{spectral width})$$

Thus, $\delta\omega$ is the width of the box function of unit area and height \bar{s} .



Note that we have constructed the diagram as if $s(\omega)$ is two-sided, but the argument carries through for the one-sided case as well, being careful to keep track of factors of 2. The definitions for $\delta\tau$ and $\delta\omega$ look like inverses, but they're really quite different because of the different normalizations of the two functions.

The big advantage of these definitions is that they are related in a simple way. We can write

$$\delta\tau \delta\omega = \frac{\int_{-\infty}^\infty |\operatorname{Re}[g^{(1)}(\tau)]|^2 d\tau}{2 \int_{-\infty}^\infty s_{\leftrightarrow}^2(\omega) d\omega}. \quad (14.58)$$

We can use **Parseval's theorem** to evaluate this ratio, which states that the signal power is equivalently measured in either the time or frequency basis:

$$\int_{-\infty}^\infty |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^\infty |\tilde{f}(\omega)|^2 d\omega. \quad (14.59) \quad (\text{Parseval's theorem})$$

Noting again that $\operatorname{Re}[g^{(1)}(\tau)]$ and $2\pi s_{\leftrightarrow}(\omega)$ form a Fourier-transform pair, and recalling the factor of two for using the one-sided spectrum, we can use this to write

$$\delta\tau \delta\omega = \pi. \quad (14.60) \quad (\text{uncertainty relation})$$

This “uncertainty relation” is a strict equality, valid for any functions as long as the measures exist and are finite. Neat, eh?

The point of all this is, for time (optical path length) delays larger than the coherence time $\delta\tau$, the fringe visibility is mostly gone. This coherence time corresponds to a physical path length difference

$$\ell_c := c \delta\tau, \quad (14.61)$$

(coherence length)

which is called the **coherence length**.

Here are some examples. For a He-Ne laser, the laser line width is $\delta\nu = \delta\omega/2\pi \sim 1$ GHz. This corresponds to a coherence time of $\delta\tau \sim 1$ ns, or a coherence length $\ell_c \sim 30$ cm. On the other hand for a light bulb that spans the visible wavelength range of 400-700 nm, the line width is

$$\delta\nu \sim \frac{\nu \delta\lambda}{\lambda} = \frac{c \delta\lambda}{\lambda^2} = 300 \text{ THz.} \quad (14.62)$$

This gives a coherence time $\delta\tau \sim 3$ fs and a coherence length $\ell_c \sim 1 \mu\text{m}$. So in fact it is possible to see interference of white light in a Michelson, but it's very difficult because the path lengths must be matched to μm accuracy. On the other hand, it's much easier to observe interference or record a hologram with light from a He-Ne laser, because it remains coherent on the scale of about a foot.

14.5 Interference Between Two Partially Coherent Sources

In general, we can now look at the interference pattern between two partially coherent sources, represented by the two fields $E_1^{(+)}(t)$ and $E_2^{(+)}(t)$. The second field has an adjustable time delay of τ . Then the intensity of the superposition of these waves is

$$\begin{aligned} I &= \frac{2}{\eta} \left\langle \left| E_1^{(+)}(t) + E_2^{(+)}(t + \tau) \right| \right\rangle \\ &= \frac{2}{\eta} \left\langle \left| E_1^{(+)}(t) \right| \right\rangle + \frac{2}{\eta} \left\langle \left| E_2^{(+)}(t + \tau) \right| \right\rangle + \left[\frac{2}{\eta} \left\langle \left| E_1^{(-)}(t) E_2^{(+)}(t + \tau) \right| \right\rangle + \text{c.c.} \right] \\ &= I_1 + I_2 + 2\sqrt{I_1 I_2} \operatorname{Re} \left[g_{12}^{(1)}(\tau) \right], \end{aligned} \quad (14.63)$$

where $g_{12}^{(1)}(\tau)$ is the **normalized cross-correlation function**:

$$g_{12}^{(1)}(\tau) := \frac{\left\langle \left| E_1^{(-)}(t) E_2^{(+)}(t + \tau) \right| \right\rangle}{\left\langle \left| E_1^{(-)}(t) E_2^{(+)}(t) \right| \right\rangle}. \quad (14.64)$$

(normalized cross-correlation)

Again, the visibility is

$$\mathcal{V} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} \left| g_{12}^{(1)}(\tau) \right|. \quad (14.65)$$

(visibility for two different fields)

This tells us that very different beams, resulting in little correlation, don't interfere very well.

14.6 Exercises

Problem 14.1

Prove Parseval's Theorem: if $f(x)$ and $\tilde{f}(k)$ form a Fourier transform pair, then

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\tilde{f}(k)|^2 dk \quad (14.66)$$

Problem 14.2

Go through the derivation in the notes and fill in the details of the “uncertainty relation” $\delta\tau\delta\omega = \pi$, where the coherence time $\delta\tau$ is the power-equivalent width of $g^{(1)}(\tau)$ and $\delta\omega$ is the effective spectral width.

Problem 14.3

Find the coherence times $\delta\tau$ according to the power-equivalent width definition for the following coherence functions:

- (a) Lorentzian: $g^{(1)}(\tau) = 1/[1 + (\tau/\Delta\tau)^2]$, where $\Delta\tau$ is the full width at half maximum.
- (b) Gaussian: $g^{(1)}(\tau) = \exp[-\tau^2/(\Delta\tau)^2]$, where $\Delta\tau$ is the $1/e$ time.
- (c) Exponential: $g^{(1)}(\tau) = \exp[-|\tau/\Delta\tau|]$, where $\Delta\tau$ is the $1/e$ time.

Problem 14.4

Consider a Michelson interferometer, where the input is bichromatic with air wavelengths λ_1 and λ_2 . Assume that the splitting is small, $|\Delta\lambda| \ll \lambda$, where $\Delta\lambda := \lambda_2 - \lambda_1$ and $\lambda := (\lambda_1 + \lambda_2)/2$.

Show that the visibility of the output fringes varies periodically as one mirror is displaced, with period

$$\Delta d = \frac{\lambda^2}{2\Delta\lambda} \quad (14.67)$$

in terms of the mirror displacement.

Chapter 15

Laser Physics

15.1 Overview

Now we'll explore in a fair amount of detail the operation of a **laser**, which is very important as a source of intense and/or highly coherent light. The name "laser" is an acronym for "light amplification by the stimulated emission of radiation," emphasizing that the laser is an active, amplifying device.

There are three key components to any laser:

1. The **pump** is the energy source that is converted to laser light.
2. The **gain medium** performs the actual conversion of energy to coherent light, by *amplifying* light passing through it.
3. The **optical resonator**, which we have already studied in detail, recycles the light through the gain medium to achieve better amplification. Also, since an optical resonator is a frequency-selective filter, it narrows the frequency spectrum of the output and thus increases the coherence of the laser light.

Put together to form a laser, these are the ingredients for an *active optical oscillator*,¹ in contrast to the passive optical elements that we have studied so far.

Laser light has a number of potentially desirable properties. Among its key features are:

1. **Monochromaticity.** A laser tends to have a relatively narrow frequency spectrum, either in the absolute sense or in the sense of being limited by the time-frequency uncertainty principle (i.e., for "transform-limited" laser pulses).
2. **Directionality.** The output of a laser has small beam divergence, typically diffraction-limited or nearly so. The output of a laser is commonly a Gaussian beam.
3. **High power.** Not all lasers have high output powers, but many of the most intense man-made optical sources are lasers.

Now let's examine the main components of a laser in more detail.

15.1.1 Laser Pumps

The laser pump is the energy reservoir for the laser. Often, the energy comes from electricity, although sometimes indirectly. Examples of pumps that directly involve electricity are gas discharges and injection currents in semiconductor lasers. Less direct are optical pumps, in the form of flashlamps or even another laser. A more bizarre example of a pump is a nuclear device, for the "pop-up x-ray laser" defense system.

¹in fact, to emphasize the importance of the oscillator nature of the laser, one could use the name "light *oscillation* by the stimulated emission of radiation," but this would be a *loser*, and not nearly as cool. (I stole this joke from Leno Pedrotti, from way back when I learned laser physics from him at the University of Dayton.)

15.1.2 Gain Media

The gain medium converts the pump energy into the desired light. Invariably the gain medium involves quantum systems, where the transition energies determine the output wavelength.

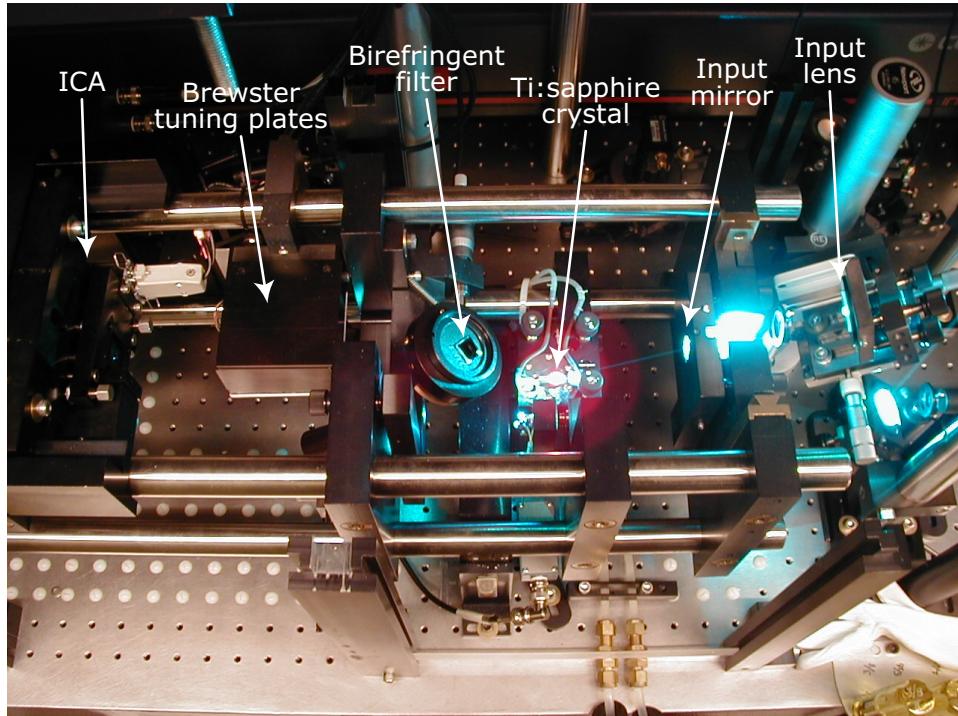
15.1.2.1 Gas-Phase Atoms

One important class of gain media is atoms in the gas phase. Examples include the He–Ne laser and the argon-ion laser. These gain media typically lase in the visible to ultraviolet range, and they are pumped by electrical discharge. Commonly these lasers operate in CW (continuous-wave) mode, although pulsed gas lasers are possible (e.g., nitrogen and copper-vapor lasers). The efficiencies of gas lasers tend to be appalling, in the neighborhood of 0.05% for an Ar^+ laser. Output powers vary from 1 mW (He–Ne) to 20 W (Ar^+).

15.1.2.2 Atoms Embedded in Transparent Solids

Atoms can also be suspended in (doped into) transparent solids. Important examples are Nd:YAG, Nd:glass, Ti:sapphire (sapphire is crystalline Al_2O_3), ruby (Cr:sapphire), Er:fiber, and Yb:fiber. Here, the notation is “dop-ant:medium.” These lasers operate mostly in the near-infrared (in the range of 0.7–1.6 μm), and can operate as either CW or pulsed lasers. The efficiencies of these solid-state lasers are much higher than gas lasers, in the neighborhood of 1%. The efficiency varies widely depending on the details of the pump; for example, a laser-pumped Ti:sapphire laser can have an efficiency approaching 10%. Output powers are in the range of 100 mW (CW lasers) to PW peak pulse powers for some truly huge lasers ($P = \text{peta} = 10^{15}$).

Here is a closeup view of part of a Ti:sapphire ring-cavity laser. The blue-green pump light (from an Ar^+ laser) is clearly visible as it enters through the input mirror (with a special coating to transmit blue-green light but reflect the 850 nm laser light).



The input lens aligns and matches the pump-beam profile to the laser cavity mode. The Ti:sapphire crystal is Brewster-cut to suppress reflection loss at the surfaces and glows bright red due to **amplified spontaneous emission**. The other components (birefringent filter, Brewster etalons, ICA) all help force the laser to oscillate in a single cavity mode. The output mirror is to the left of the area shown. For overall scale, the holes in the table surface are spaced at 1" intervals.

15.1.2.3 Molecules

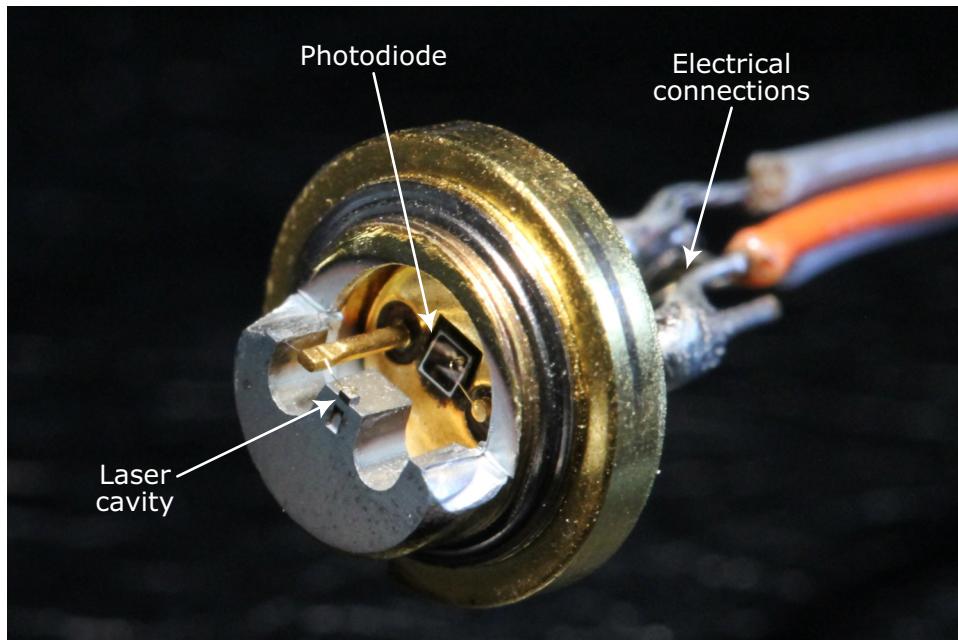
Molecular lasers come in two main flavors. A gas-phase molecular laser is CO₂, and this is a particularly important laser in industrial cutting and welding. Molecules can also be contained in liquid solvent, as in the manifold organic laser dyes. The wavelengths of these lasers range from the infrared (10.6 μm for CO₂) to the ultraviolet/visible/near-infrared, the wide range covered by the various available laser dyes. The efficiencies of molecular lasers are also relatively high, in the range of 1-10%, depending on the details of the pump. Output powers range from around 100 mW (the low end of dye lasers) to 10 kW for CO₂ lasers.

One of the most amusing (and nearly edible) lasers falls in this category: the gelatin laser, where the molecules included sodium fluorescein (marginally edible) and rhodamine 6G (much less edible). The “normal slight tremor” when holding the gain medium by hand was observed to enhance the operation of the medium.²

15.1.2.4 Semiconductor Lasers

Diode lasers have become enormously important in recent years. Materials include GaAs, GaAlAs, and InGaN/GaN/AlGaN. Most diodes lase in the red- to near-infrared, although InGaN/GaN/AlGaN lasers in the 400-420 nm range. Diode lasers are pumped via injection current flowing across the p-n junction. Efficiencies of diode lasers are as high as it gets (up to 65%), and output powers are in the range of 1 mW to 50 W (for diode arrays).

A single-mode, low-power laser diode is shown in the photo below. This diode would normally have a cylindrical, protective metal “can” surrounding the front end of the diode, with a window for the laser output, which has been removed here. For overall scale, the outer, circular flange of the diode is 9 mm in diameter.



The laser cavity itself is a tiny dot in this photo; the active wave-guiding region has transverse dimensions in the range of about 1–6 μm, though the transverse extent of the chunk of semiconductor is larger than this. The length of the cavity is of the order of 100 μm. The front of the laser emits out of the page, down, and to the left of the photo. The back of the cavity is also transparent and emits some light, which struck a monitor photodiode (photodetector). Electrical connections come in through three pins in the back. The leftmost pin (as viewed in the photo) protrudes out next to the semiconductor, and a thin wire is visible connecting the pin to the top of the laser cavity. The bottom of the laser cavity is connected to case ground.

²T. W. Hänsch, M. Pernier, A. L. Schawlow, “Laser Action of Dyes in Gelatin,” *IEEE Journal of Quantum Electronics* QE-7, 45 (1971).

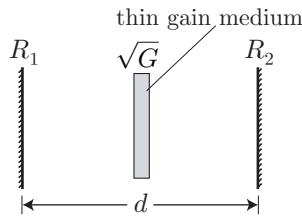
The rightmost pin is connected to the photodiode by another thin wire, visible to the right of the photodiode. The photodiode shares a common ground connection with the laser.

15.1.3 Optical Resonator

The laser cavity or resonator is often a spherical-mirror resonator. Typically, the laser resonator is stable for CW operation, but sometimes unstable resonators are used in pulsed-laser designs.

15.1.4 A Simple Model of Laser Oscillation: Threshold Behavior

We can get a great deal of insight about laser operation just by considering a very simple model.³ Consider a laser, composed of a two-mirror resonator, with mirror reflectances $R_{1,2}$. The laser contains a thin gain medium, with single-pass intensity gain \sqrt{G} .



The effect of the mirrors after one round trip on the intracavity intensity is

$$I_{\text{after}} = R_{1,2} I_{\text{before}}. \quad (15.1)$$

The effect of the gain medium after a single pass is

$$I_{\text{after}} = \sqrt{G} I_{\text{before}}. \quad (15.2)$$

Note that G is the *round-trip* gain, where $G > 1$ for net gain and $G < 1$ for net loss (e.g., due to scattering loss). The reflectances $R_{1,2} < 1$, representing loss. Thus, the intensity after a round trip is

$$I_{\text{after}} = G R_1 R_2 I_{\text{before}}. \quad (15.3)$$

If $\tau_{\text{rt}} = 2d/c$ is the cavity round-trip time (ignoring any contribution due to the thin gain medium), we can write

$$I(t + \tau_{\text{rt}}) = G R_1 R_2 I(t). \quad (15.4)$$

Let's assume that the intensity varies slowly on the time scale of τ_{rt} . Then we can expand to lowest order in τ_{rt} :

$$I(t) + \frac{dI(t)}{dt} \tau_{\text{rt}} + O(\tau_{\text{rt}}^2) = G R_1 R_2 I(t). \quad (15.5)$$

Rearranging gives the differential equation

$$\frac{dI(t)}{dt} = \frac{(G R_1 R_2 - 1)}{\tau_{\text{rt}}} I(t). \quad (15.6) \quad (\text{simple laser rate equation})$$

Note that this is a very simple model: we are ignoring the fact that the gain G should depend on the pump, and we are assuming that the laser light is resonant with the cavity. But with the assumption of constant G , we can write the explicit solution to Eq. (15.6) as

$$I(t) = I_0 \exp \left[\frac{(G R_1 R_2 - 1)}{\tau_{\text{rt}}} t \right]. \quad (15.7) \quad (\text{laser evolution})$$

³In this and the next subsection we are following the excellent treatment in Frank L. Pedrotti, Leno M. Pedrotti, and Leno S. Pedrotti, *Introduction to Optics*, 3rd ed. (Benjamin Cummings, 2006).

Note that there is a **threshold behavior** represented in this solution, depending on the value of the gain:

$$\begin{aligned} G > \frac{1}{R_1 R_2} &\implies I(t) \rightarrow \infty \\ G < \frac{1}{R_1 R_2} &\implies I(t) \rightarrow 0. \end{aligned} \tag{15.8}$$

(laser threshold behavior)

That is, the gain must be larger than the *threshold gain* of $1/R_1 R_2$ (i.e., the gain must exceed the loss) in order for the intensity to start growing. But an exponentially growing (divergent) solution is unphysical! We've arrived at this funny solution because the *linear* model with constant G is too simple.

15.1.5 A Less-Simple Model of Laser Oscillation: Steady-State Oscillation

A better (but still quite simplified) model for the gain is

$$G(I) = 1 + \frac{g_0}{1 + \frac{I}{I_{\text{sat}}}}, \tag{15.9}$$

(gain-saturation model)

where g_0/ℓ_g is the **small-signal gain coefficient** (with ℓ_g the length of the thin gain medium), and I_{sat} is the **saturation intensity**. For small intensities ($I \ll I_{\text{sat}}$),

$$G(I) \approx 1 + g_0 =: G_0, \tag{15.10}$$

and so the small-signal gain G_0 is a constant, and is also the maximum possible value of G . For large intensities ($I \gg I_{\text{sat}}$), the gain becomes

$$G(I) \approx 1 + \frac{g_0 I_{\text{sat}}}{I}, \tag{15.11}$$

and thus the gain decreases as I increases, with $G \rightarrow 1$ as $I \rightarrow \infty$. That is, there is no gain in the limit of large input intensity. This nonlinear effect is called **gain saturation**, and is a universal feature of physical amplifiers, which have finite energy reservoirs.

The saturation intensity I_{sat} marks the crossover between these two regimes. If we write the gain as

$$G(I) = 1 + g(I), \tag{15.12}$$

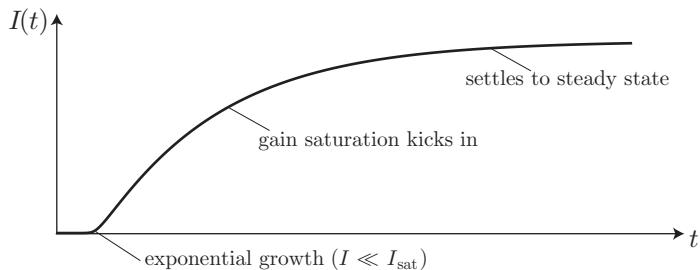
where the integrated gain coefficient $g(I)$ is related to the *small-signal* gain coefficient by

$$g(I) = \frac{g_0}{1 + \frac{I}{I_{\text{sat}}}}. \tag{15.13}$$

Then the saturation intensity is defined such that when $I = I_{\text{sat}}$, the gain drops to half its small-signal value:

$$g(I_{\text{sat}}) = \frac{g_0}{2}. \tag{15.14}$$

Now let's look at the solution to Eq. (15.6) when we include gain saturation, assuming we start with some small, positive intensity.



Assuming we are above threshold, we only see the exponential rise in intensity while $I \ll I_{\text{sat}}$. Once I becomes comparable to I_{sat} , gain saturation begins to kick in, tempering the growth of the intensity. The gain continues to decrease due to gain saturation until it just balances the losses, and the intensity settles into a steady state.

We can calculate the steady-state intensity I_{ss} by setting $dI/dt = 0$. This happens when

$$G(I_{\text{ss}})R_1R_2 = 1. \quad (15.15)$$

Putting in the gain from Eq. (15.9) and solving for I_{ss} ,

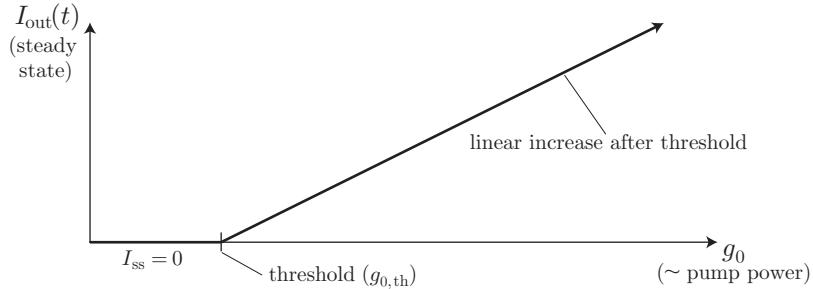
$$I_{\text{ss}} = \left[\frac{(g_0 + 1)R_1R_2 - 1}{1 - R_1R_2} \right] I_{\text{sat}}. \quad (15.16)$$

(steady-state intensity)

Again, this is only true if the small-signal gain G_0 is above threshold ($G_0R_1R_2 > 1$). If the small-signal gain is below threshold ($G_0R_1R_2 < 1$), then any initial intensity decays away, and so $I_{\text{ss}} = 0$. Roughly speaking, I_{sat} is a fixed property of a given gain medium, and g_0 is proportional to the pump power (at least for small g_0). Noting that the steady-state laser output is

$$I_{\text{out}} = (1 - R_2)I_{\text{ss}}, \quad (15.17)$$

where the second mirror is the output coupler, we can see the more physical threshold behavior of a laser.



Above the threshold value $g_{0,\text{th}}$ of the small-signal integrated gain coefficient, the steady-state output intensity rises linearly with slope

$$\left(\frac{R_1R_2}{1 - R_1R_2} \right) (1 - R_2). \quad (15.18)$$

For a given pump, and thus gain g_0 , there is an optimum output coupler reflectance R_2 that maximizes the output power, because this slope and the threshold gain are coupled together. However, we will defer this calculation until we have treated the laser in more detail.

15.2 Light–Atom Interactions

15.2.1 Quantization

Here we are going to adopt a fairly simplistic view of how the electromagnetic field interacts with atoms. In particular, we will be using the language of photons, but we will stick to a strictly semiclassical treatment, not really treating the atoms *or* the field quantum mechanically.

But we will start with the observation that the energy in an electromagnetic field is *quantized*. This means that a monochromatic field of frequency ω (typically restricted to some “quantization volume” such as an optical cavity) has possible energies given by

$$E_n = \left(n + \frac{1}{2} \right) \hbar\omega, \quad (15.19)$$

(allowed field energies)

where n is a nonnegative integer, representing the number of **photons** in the field. This may be familiar as the energy-level structure of the quantum harmonic oscillator. The photon number is always defined with respect to a particular mode (fixing the direction, polarization, and frequency characteristics).

The energies for atoms and molecules are also quantized, although the exact energy-level structure depends on the specific atom or molecule. If we denote the quantized energies by E_n , then the *differences* in energy levels correspond to frequencies via

$$\Delta E_{mn} := E_n - E_m = \hbar\omega_{mn}. \quad (15.20)$$

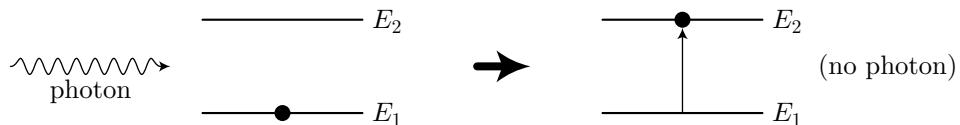
(transition energies)

The idea is that atoms with an energy difference ΔE_{mn} prefer to interact with *resonant* fields of frequency ω_{mn} . In this case, the energy of a single photon matches the atomic energy difference, and energy is conserved. There are different types of transitions, generally corresponding to different types of radiation. Electronic transitions in atoms are the most energetic of the type we will consider, and they correspond to visible optical frequencies. Vibrational transitions in a molecule correspond to different amplitudes and types of motion *internal* to the molecule, and generally correspond to radiation in the infrared. Rotational transitions in molecules have yet lower energy, and they correspond to microwave radiation (which enables the **maser**, the microwave predecessor to the laser).

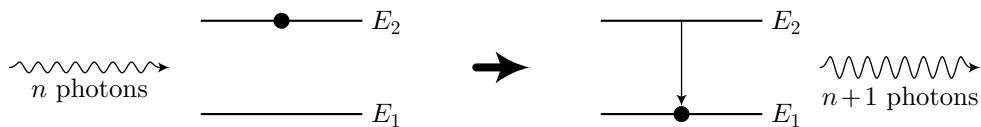
15.2.2 Fundamental Light-Atom Interactions

There are three fundamental interactions between light and atoms. In all cases we will consider only a two-level atom with ground-state energy E_1 and excited-state energy E_2 . We will also assume resonant light, $\omega = (E_2 - E_1)/\hbar$.

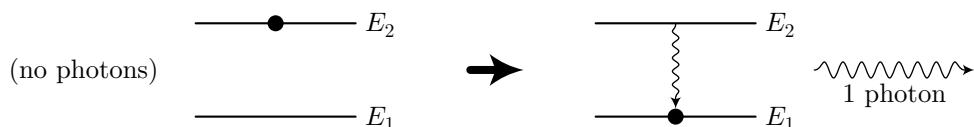
- Absorption (stimulated).** In the absorption process, a photon is destroyed and the atom is promoted to the excited state. More generally, if there are n photons to start with in some resonant mode, then there are $n - 1$ photons after the absorption process.



- Stimulated Emission.** This process involves the atom initially being in the excited state, in the presence of n photons in some resonant mode. After the stimulated-emission event, the atom is demoted to the ground state and the field is left with $n + 1$ photons. In some sense, this process is the opposite of stimulated absorption, although absorption ending with 0 photons is possible while stimulated emission beginning with 0 photons is not.



- Spontaneous Emission.** This process is much like stimulated emission, but when the atom is demoted, a photon is created in some mode that is initially unpopulated. Thus, a photon can go into a wide range of possible modes by spontaneous emission. It is possible to view spontaneous emission as stimulated emission due to quantum vacuum fluctuations in addition to classical radiation reaction (the bound electron radiates because it is accelerating).



Generally, we can associate stimulated absorption and emission with the laser mode, while spontaneous emission is additionally associated with all other modes. Absorption and stimulated emission are the most important once the laser gets going—absorption *attenuates* the laser mode, while stimulated emission *amplifies* the laser mode.

15.2.3 Einstein Rate Equations

Now let's consider an *ensemble* of two-level atoms interacting with light. Let $N_{1,2}$ denote the number density of atoms with energy $E_{1,2}$. Then the Einstein rate equation for the excited state is

$$\frac{dN_2}{dt} = -A_{21}N_2 - B_{21}\rho(\omega)N_2 + B_{12}\rho(\omega)N_1.$$

(Einstein rate equation, excited state) (15.21)

Here, $\rho(\omega)$ is the energy density of the electromagnetic field (the energy density in the frequency interval ω to $\omega + d\omega$). The first term corresponds to spontaneous emission, and we can see that it reduces the excited-state population, even in the absence of any field. The second and third terms are proportional to $\rho(\omega)$, and correspond to stimulated emission and absorption, respectively, as we can see from their overall signs. By convention, the constant A_{21} is called the **Einstein A coefficient**, while B_{21} and B_{12} are called the **Einstein B coefficients**.

To be consistent, $N_1 + N_2$ must add up to some constant, assuming that we really have two-level atoms and that the atoms stay in place (something that even works fairly well for gas lasers as long as we modify A_{21} appropriately). Thus, $dN_2/dt = -dN_1/dt$, and so it is easy to write down the rate equation

$$\frac{dN_1}{dt} = A_{21}N_2 + B_{21}\rho(\omega)N_2 - B_{12}\rho(\omega)N_1.$$

(Einstein rate equation, ground state) (15.22)

for the ground-state population N_1 .

It is most convenient to look at the steady-state behavior to see if we can have laser operation. Steady state occurs when $dN_2/dt = 0$, whence it follows from Eq. (15.21) that

$$\frac{N_2}{N_1} = \frac{B_{12}\rho(\omega)}{A_{21} + B_{21}\rho(\omega)}. \quad (15.23)$$

If the energy levels are not degenerate, it turns out that $B_{12} = B_{21}$, as we will see in the next section. That is, stimulated emission and absorption are exactly symmetric from the rate-equation point of view. Then we can rewrite the steady-state solution as

$$\frac{N_2}{N_1} = \frac{1}{\frac{A_{21}}{B_{21}\rho(\omega)} + 1}. \quad (15.24)$$

(steady-state solution, two-level atom)

We can see from this that $N_2 < N_1$ in steady state. Thus, *there is no steady-state population inversion in a two-level system*. The spontaneous emission breaks the symmetry of the stimulated emission and absorption processes and guarantees that there will be always more atoms in the ground state than in the excited state. In the limit of large intensity, $\rho(\omega) \rightarrow \infty$, the best we can do is to *equalize* the population. Thus, we can't use a two-level system for a laser gain medium. We need a population inversion for stimulated emission (amplification) to win out over absorption (attenuation).

15.2.4 Relations Between the Einstein Coefficients

Now we briefly outline Einstein's derivation of the relation between the A and B coefficients.⁴ If the energy levels are degenerate, we can define the degeneracy factors $g_{1,2}$ as the number of ways of having energy $E_{1,2}$.

⁴ A. Einstein, "Zur Quantentheorie der Strahlung," *Physikalische Zeitschrift* **18**, 121 (1917). For an English translation, see A. Einstein, "On the Quantum Theory of Radiation," in *The World of the Atom*, H. A. Boorse and L. Motz, Eds. (Basic Books, 1966), vol. 2, p. 888. Einstein's main aim in this paper was in fact to derive the Planck blackbody distribution by solving the rate equations.

For example $g_{1,2} = 2J_{1,2} + 1$ for atomic angular-momentum states. Then the steady-state population ratio from Eq. (15.22) can be written also via Boltzmann statistics as

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} e^{-\hbar\omega/k_B T} = \frac{B_{12}\rho(\omega)}{A_{21} + B_{21}\rho(\omega)}. \quad (15.25)$$

Solving for $\rho(\omega)$,

$$\rho(\omega) = \frac{A_{21}}{B_{21}} \frac{1}{\left(\frac{B_{12}g_1}{B_{21}g_2} e^{\hbar\omega/k_B T} - 1 \right)}. \quad (15.26)$$

This is equivalent to the Planck blackbody distribution⁵

$$\rho(\omega) = \frac{8\pi\hbar}{\lambda^3} \frac{1}{e^{\hbar\omega/k_B T} - 1} \quad (15.27)$$

if we make the identifications

$$g_2 B_{21} = g_1 B_{12}. \quad (15.28)$$

(relation between B coefficients)

and

$$\frac{A_{21}}{B_{21}} = \frac{8\pi\hbar}{\lambda^3}. \quad (15.29)$$

(relation between A and B coefficients)

Recall that λ here is the wavelength *within* the gain medium.

15.2.5 Line Shape and Spectral Distributions

So far, we've considered only monochromatic light and two-level atoms with sharply defined energy levels. Now it's time to improve our model of the two-level atom and its interaction with light.

We will first introduce a **line-shape function** $s(\omega)$ to model the fact that the energy levels have some width. The line shape is defined such that $s(\omega)d\omega$ is the probability that a spontaneously emitted photon will have frequency between ω and $\omega + d\omega$. We can also interpret this as the *relative* probability of stimulated emission or absorption of a photon with frequency between ω and $\omega + d\omega$. Since $s(\omega)$ represents a probability density, it is appropriately normalized:

$$\int_0^\infty s(\omega) d\omega = 1. \quad (15.30)$$

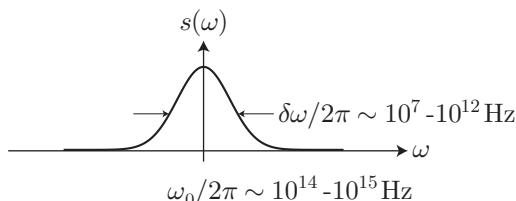
Note that as in our discussion of coherence in Chapter 14, we are using a "one-sided spectrum" that ranges only over positive frequencies. In terms of the "two-sided spectrum" $s_{\leftrightarrow}(\omega)$ with both positive and negative frequencies, the one-sided spectrum satisfies $s(\omega) := s_{\leftrightarrow}(\omega) + s_{\leftrightarrow}(-\omega)$ for $\omega \geq 0$ and $s(\omega) = 0$ for $\omega < 0$.

When we apply the line shape and sum over all frequencies, the rate equation becomes

$$\frac{dN_2}{dt} = -A_{21}N_2 - B_{21}N_2 \int_0^\infty \rho(\omega)s(\omega) d\omega + B_{12}N_1 \int_0^\infty \rho(\omega)s(\omega) d\omega. \quad (15.31)$$

(rate equation including spectra)

Qualitatively, we can picture the line shape function as a relatively sharply peaked distribution centered around the resonant optical frequency ω_0 .



⁵P. W. Milonni and M.-L. Shih, "Zero-point energy in early quantum theory," *American Journal of Physics* **59**, 684 (1991).

Often, $s(\omega)$ turns out to be a Lorentzian, a Gaussian, or a convolution of the two (a *Voigt* profile). The line-shape function models transition width due to spontaneous emission, collisions, Doppler shifts in gas lasers, and local crystal structure effects on the dopant atoms. Note that in the absence of radiation, the rate equation is $dN_2/dt = -A_{21}N_2$, which has an exponentially damping solution. The Fourier transform of an exponential is a Lorentzian, so the line shape for spontaneous emission (the “natural line shape”) is Lorentzian, with a half-width at half maximum of A_{21} . Collisions are often modeled by a spontaneous-emission-like term, and thus also lead to Lorentzian line shapes. Doppler shifts lead to Gaussian line shapes because the Maxwell–Boltzmann velocity distribution is Gaussian. If multiple, independent broadening effects contribute, their combined effect can be modeled by the convolution of the individual line shapes.

Now we will consider two limiting cases for the light spectrum. Both are important in understanding laser operation, but the second is the more useful case for understanding laser gain.

15.2.5.1 Broadband Light

Light is broadband (relative to the transition) if $\rho(\omega)$ is much broader than $s(\omega)$. Then we can evaluate the integral in Eq. (15.31) by noting that $\rho(\omega_0)$ varies slowly over the width of $s(\omega)$, so that we can pull it out of the integral:

$$\int_0^\infty \rho(\omega)s(\omega) d\omega \approx \rho(\omega_0) \int_0^\infty s(\omega) d\omega = \rho(\omega_0). \quad (15.32)$$

Thus, we recover the previous rate equations, corresponding to Eq. (15.21), with sharp energy levels.

15.2.5.2 Nearly Monochromatic Light

For nearly monochromatic light, the field spectrum is narrow, so $s(\omega)$ is much broader than $\rho(\omega)$. Thus, we can evaluate the integral with the same slowly varying approximation as for the broadband case:

$$\int_0^\infty \rho(\omega)s(\omega) d\omega \approx s(\omega_{\text{field}}) \int_0^\infty \rho(\omega) d\omega. \quad (15.33)$$

The integral on the right-hand side is the total field energy density, summed over all frequencies. Let’s denote this by ρ_{total} . Then the rate equation becomes

$$\frac{dN_2}{dt} = -A_{21}N_2 - B_{21}N_2s(\omega_{\text{field}})\rho_{\text{total}} + B_{12}N_1s(\omega)\rho_{\text{total}}. \quad (15.34)$$

The energy density is related to the intensity by $\rho_{\text{total}} = I_{\text{total}}/c$ (see Problem 4.4), so

$$\frac{dN_2}{dt} = -A_{21}N_2 - \sigma(\omega_{\text{field}}) \frac{I_{\text{total}}}{\hbar\omega_{\text{field}}} \left[N_2 - \frac{g_2}{g_1} N_1 \right].$$

(rate equation for monochromatic light) (15.35)

Here, we have defined the **laser cross-section**

$$\sigma(\omega) = A_{21} \frac{\lambda^2}{4} s(\omega). \quad (15.36)$$

(laser cross-section)

The cross-section has the dimensions of area, and is defined such that $\sigma(\omega)I(\omega)$ is the power absorbed by a single atom when irradiated by intensity $I(\omega)$ (in the weak-excitation limit). Note that for a Lorentzian line shape $s(\omega)$,

$$s(\omega) = \frac{\Delta\omega}{2\pi [(\omega_0 - \omega)^2 + (\Delta\omega/2)^2]}, \quad (15.37)$$

the *resonant* cross section $\sigma(\omega_0)$ is given by

$$\sigma(\omega_0) = \frac{A_{21}}{\Delta\omega} \frac{\lambda^2}{2\pi}. \quad (15.38)$$

For homogenous broadening, $\Delta\omega$ is the **natural line width** given by $\Delta\omega = A_{21}$, so that the natural cross section is

$$\sigma(\omega_0) = \frac{\lambda_0^2}{2\pi}. \quad (15.39)$$

This answer is consistent with a fully quantum-mechanical calculation. In fact, this answer assumes an average over all possible atomic orientations, since the blackbody distribution of Eq. (15.27) assumes isotropic radiation. For atomic dipole moments aligned with the field polarization, the resonant cross section is

$$\sigma(\omega_0) = \frac{3\lambda_0^2}{2\pi}, \quad (15.40)$$

since the coupling that would normally be “distributed” among three orthogonal directions is concentrated into one.

15.3 Light Amplification

In writing down the atomic rate equations, we have developed a quantitative model for the influence of light on atoms. To complete our picture of light-matter interactions to model a laser, we have to quantify the influence of *atoms* on *light*—the other aspect of the fundamental radiation processes of absorption, stimulated emission, and spontaneous emission. This will allow us to quantify the gain of the laser gain medium in terms of atomic (or molecular) parameters. To do so, we will consider the amplification (or absorption) of a monochromatic input intensity component $I(\omega)$ by a thin gain medium of thickness dz . Note that we will assume the input light to be in a single mode; however, the gain medium can in principle radiate into many different modes, and we will need to physically distinguish input photons from amplification photons or spontaneously emitted photons.⁶

15.3.1 Gain Coefficient

Now we can compute the change in intensity through the gain medium and then through the filtering and detection system above. We will treat the contribution of each of the three fundamental processes separately. In each case, we will consider an infinitesimal intensity change of the form

$$dI(\omega) = \left(\frac{\text{photon-interaction rate}}{\text{atom}} \right) \left(\frac{\text{energy}}{\text{photon}} \right) \left(\frac{\text{atoms}}{\text{cross-sectional area}} \right) \left(\frac{\text{line-shape}}{\text{weight}} \right) \quad (15.41)$$

due to the gain medium of thickness dz . We have simply converted the rate at which photons are absorbed or emitted by a single atom to an intensity, essentially by changing units: we multiplied by the energy per photon to give the rate at which energy is absorbed or emitted per atom; then we multiplied by the atoms per unit area of the gain medium to give an intensity; then finally, we multiplied by the spectral line shape to account for the frequency dependence (resonance behavior) of the atom, producing an intensity density as we expect for $I(\omega)$.

15.3.1.1 Stimulated Emission

The intensity change for stimulated emission, according to the general template of Eq. (15.41), becomes

$$dI(\omega) = \left(B_{21} \frac{I(\omega)}{c} \right) \left(+\hbar\omega \right) \left(N_2 dz \right) \left(s(\omega) \right). \quad (\text{stimulated emission}) \quad (15.42)$$

Let's go through these factors again in a bit more detail. From the atomic rate equation, stimulated-emission events occur at the rate of $B_{21}\rho(\omega) = B_{21}I(\omega)/c$ per atom, and every emission event adds the photon energy $\hbar\omega$ to the atom. This must be weighted by the relative emission probability $s(\omega)$ at the incident optical

⁶Here we are following the construction, and most of the notation, of Joseph T. Verdeyen, *Laser Electronics*, 3rd ed. (Prentice Hall, 1995).

frequency. We then multiply this rate per atom by the number of excited-state atoms per unit area, $N_2 dz$. Due to the nature of stimulated emission, the emitted photons are in the same mode as the incoming photons, and so we count every stimulated-emission event as contributing to the amplification of the input mode.

15.3.1.2 Absorption

Absorption leads to a similar expression for the intensity change,

$$dI(\omega) = \left(B_{12} \frac{I(\omega)}{c} \right) \left(-\hbar\omega \right) \left(N_1 dz \right) \left(s(\omega) \right), \quad (\text{absorption}) \quad (15.43)$$

with the only two changes being the sign of the photon energy (and hence the overall sign), since absorption *attenuates* the wave, and the number of atoms per unit area is now $N_1 dz$, since only the ground-state atoms are relevant here.

15.3.1.3 Spontaneous Emission

For spontaneous emission, we have a more complicated expression, because spontaneous emission can occur into *any* mode, but we want to count only those in the *input* mode:

$$dI(\omega) = \left(A_{21} \right) \left(+\hbar\omega \right) \left(N_2 dz \right) \left(s(\omega) \right) \left(\frac{d\Omega}{4\pi} \right) \left(\frac{\Delta\omega}{\Delta\omega_{\text{gain}}} \right) \left(\frac{1}{2} \right). \quad (\text{spontaneous emission}) \quad (15.44)$$

The emission rate is similar to the stimulated-emission version, but the overall rate per atom is now given in terms of the Einstein A coefficient, and of course the energy density of the input mode does not enter here. The other differences lie in the fraction of spontaneously emitted photons that are emitted into the input mode (i.e., the laser-resonator mode if we are considering amplification within the laser). Only a small fraction of the photons have the right direction to match the input mode, which we assume subtends the solid angle $d\Omega$, so we include the factor $d\Omega/4\pi$. The fraction of photons with the proper frequency in the width $\Delta\omega$ of the input mode is $\Delta\omega/\Delta\omega_{\text{gain}}$, where $\Delta\omega_{\text{gain}}$ is the width of $s(\omega)$. Finally, we include a factor of $1/2$, since spontaneous emission can occur into either of 2 possible polarizations, and only one will match the input mode. Generally, the angle and frequency factors are the most restrictive, and hence spontaneous emission accounts for only a small fraction of the detected intensity. Thus, we can ignore it for the purposes of computing gain once the laser is going, but it is still important to get the laser going, since it always provides a source of power, even for an initially empty cavity.

15.3.1.4 Combined Effects

Now we can combine the contributions of Eqs. (15.42) and (15.43), while neglecting spontaneous emission. We can use the relations of Eqs. (15.28) and (15.28) between the Einstein coefficients and the definition of the laser cross section, Eq. (15.36), to write

$$\frac{dI(\omega)}{dz} = \frac{\hbar\omega B_{21}}{c} I(\omega) \left[N_2 - \frac{g_2}{g_1} N_1 \right] = \sigma(\omega) I(\omega) \left[N_2 - \frac{g_2}{g_1} N_1 \right], \quad (15.45)$$

which becomes

$$\frac{dI(\omega)}{dz} = \gamma(\omega) I(\omega), \quad (15.46)$$

(intensity change by medium)

if we define the **gain coefficient** (gain per unit length)

$$\gamma(\omega) := \sigma(\omega) \left[N_2 - \frac{g_2}{g_1} N_1 \right]. \quad (15.47)$$

(gain coefficient)

Thus, we see that in terms of the atomic parameters, the gain is related only to the cross section and the population inversion. How does gain saturation fit into this? Without any optical intensity, the population

relaxes to some steady state depending on other pumping fields and decay mechanisms. In the limit of small intensity, the light does not change the populations, and so the gain coefficient keeps the value of the **small-signal gain coefficient** $\gamma_0(\omega)$. Strong light, however, depletes the population inversion, since each added photon reduces the inversion. Thus, $N_2 - (g_2/g_1)N_1 \rightarrow 0$ for strong intensities, and the gain coefficient reduces to zero. This is the fundamental root of the gain saturation that we explored earlier.

15.3.2 Threshold Behavior and Single-Mode Operation

Again, for small signals, the gain coefficient $\gamma(\omega) \approx \gamma_0(\omega)$, and is thus constant. The solution to Eq. (15.46) is therefore just an exponential,

$$I(\omega, z) = I_0(\omega) e^{\gamma(\omega)z}, \quad (15.48)$$

and so we can write the total (small-signal) round-trip power gain as

$$G_0 = \exp[2\gamma_0(\omega)\ell_g], \quad (15.49)$$

(small-signal, round-trip gain)

where ℓ_g is the length of the gain medium. The factor of two here arises because we are assuming a linear laser cavity, so that a round trip entails two passes through the medium.

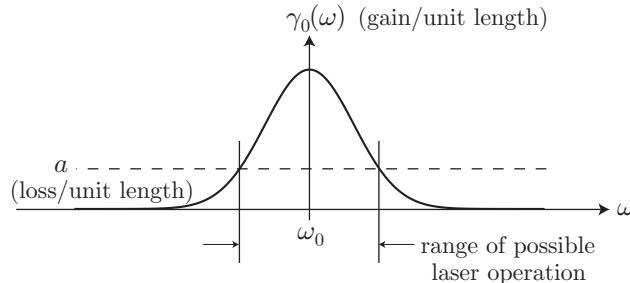
Recall that the threshold condition is that the gain matches the loss. For oscillation in a two-mirror linear cavity, we should have $G_0 R_1 R_2 \geq 1$. We can rewrite this as

$$\exp[2\gamma_0(\omega)\ell_g] R_1 R_2 \geq 1, \quad (15.50)$$

or solving for γ_0 ,

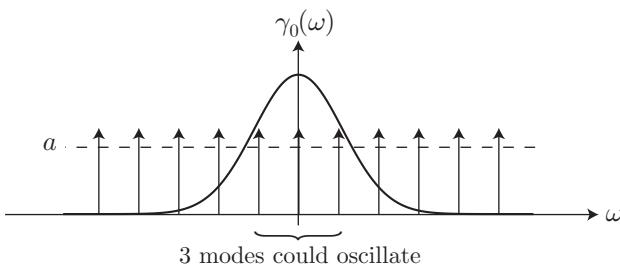
$$\gamma_0(\omega) \geq \frac{1}{2\ell_g} \log\left(\frac{1}{R_1 R_2}\right) =: a, \quad (\text{threshold condition for gain coefficient}) \quad (15.51)$$

where a is the effective “loss per unit length” per round trip, distributed along the length of the gain medium. In terms of this coefficient, the threshold condition is $\gamma_0(\omega) \geq a$ for oscillation to occur. Depicting this graphically, we can gain great insight into the frequency dependence of laser operation.

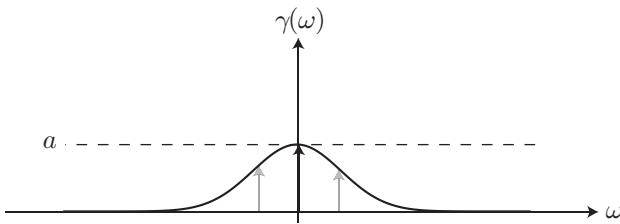


The small-signal gain profile $\gamma_0(\omega)$ reflects the line shape $s(\omega)$, and so with a sufficient pump rate, there will be a range of frequencies that will be above threshold. Any cavity modes that fall within this range are candidates for laser operation.

However, for many lasers, the laser operates in only a single cavity mode, even when there are several candidate modes. This occurs if $s(\omega)$ is the same for all the atoms in the medium (where the width of $s(\omega)$ is said to be due to “homogenous broadening”), and all the atoms interact with the cavity modes in the same way. Under these conditions, we get **single-mode operation**: *only the mode with the largest gain will oscillate*. To see how this works, consider this graphical example.

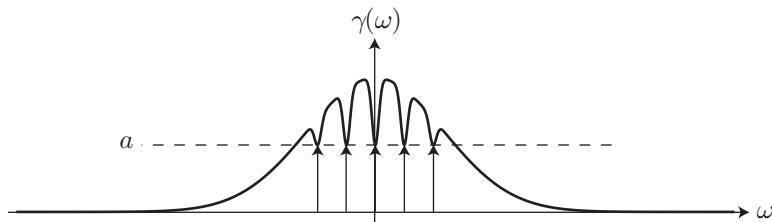


Three modes are above threshold and can potentially oscillate. Initially, they all start to oscillate. As gain saturation kicks in, the gain profile $\gamma(\omega)$ is reduced in magnitude. In equilibrium, the gain must balance the loss, so the equilibrium condition is that $\gamma(\omega) = a$. Thus, the gain profile will continue to decrease in magnitude *until the equilibrium condition is satisfied at the laser frequency*.



The other modes with smaller gain end up below threshold ($\gamma < a$), and are thus extinguished.

This argument does not hold for certain lasers such as gas lasers (e.g., He-Ne), where “Doppler broadening” is a major component to the width of $s(\omega)$. Doppler broadening arises because atoms with different velocities experience different Doppler shifts of the laser light, and thus effectively have different resonant frequencies. Thus, different atomic velocity classes can contribute to different modes, and so the gain profile is not necessarily reduced uniformly by gain saturation.



In fact, it is possible for many modes to oscillate simultaneously. In **multi-mode operation**, the gain profile is reduced only locally around each of the laser frequencies.

15.4 Pumping Schemes

Now that we have established how the laser gain coefficient is related to the atomic inversion, we need to develop some details of specific gain media to compute the inversion. Let’s simplify things by assuming equal degeneracies, $g_1 = g_2$. (This is equivalent to setting $N'_1 = (g_2/g_1)N_1$, so there isn’t any loss of generality in doing this.)

Recall that for a two-level atom, *without* a pump, has the rate equation [from Eq. (15.21) with $B_{21} = 0$]

$$\frac{dN_2}{dt} = -A_{21}N_2 = -N_2/\tau_{21}, \quad (15.52)$$

where $\tau_{21} = 1/A_{21}$ is the **lifetime** of the excited state. This is so named because this rate equation has the exponential solution

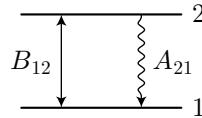
$$N_2(t) = N_2(0) e^{-A_{21}t} = N_2(0) e^{-t/\tau_{21}}, \quad (15.53)$$

and thus τ_{21} is the time scale for exponential decay from the excited level.

When we include the pump, we can write the rate equation (15.21) as

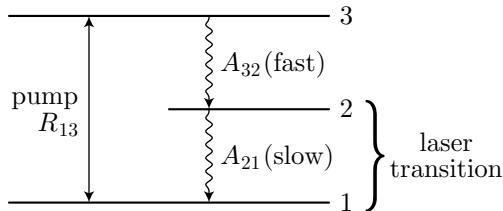
$$\frac{dN_2}{dt} = -A_{21}N_2 - \frac{\sigma(\omega)I}{\hbar\omega}(N_2 - N_1), \quad (15.54)$$

where ω is the pump frequency, and I is the total optical intensity (i.e., $I(\omega)$ integrated over all frequencies), if we assume a nearly monochromatic pump. The basic point that we made earlier is that because the pump enters the rate equation with a term of the form $(N_2 - N_1)$, the pump tries to *equalize* N_1 and N_2 . Since spontaneous decay is asymmetric, there is no possibility for population inversion in the two-level atom. Without inversion, there is no gain, and thus no laser oscillation.



15.4.1 Three-Level Laser

The simplest change that we can make to achieve a population inversion is to add a third level. The ruby laser is one example of a three-level laser. The level scheme is shown here.



The new level (with highest energy) decays quickly, while the laser ($2 \rightarrow 1$) transition decays slowly. That is, we will assume $A_{21} \ll R_{13}, A_{32}$. Also, for a monochromatic pump (e.g., the pump is another laser), we can use Eq. (15.35) to write

$$R_{13} = \frac{\sigma(\omega)I}{\hbar\omega}, \quad (15.55) \quad (\text{pumping rate, monochromatic pump})$$

where again I is the total intensity, while if the pump is broadband (e.g., a flash lamp), we can use Eq. (15.32) to write

$$R_{13} = \frac{I(\omega_0)}{\hbar\omega_0} \int_0^\infty \sigma(\omega) d\omega = A_{21} \frac{\lambda^2 I(\omega_0)}{4\hbar\omega_0}, \quad (15.56) \quad (\text{pumping rate, broadband pump})$$

where $\omega_0 = (E_3 - E_1)/\hbar$ is the pump transition frequency. Then we can write the rate equations as

$$\begin{aligned} \frac{dN_3}{dt} &= -R_{13}(N_3 - N_1) - A_{32}N_3 \\ \frac{dN_2}{dt} &= A_{32}N_3 - A_{21}N_2 \\ \frac{dN_1}{dt} &= A_{21}N_2 + R_{13}(N_3 - N_1), \end{aligned} \quad (15.57) \quad (\text{rate equations, three-level atom})$$

where of course one of the equations is redundant since $N_1 + N_2 + N_3$ must be a constant of the motion.

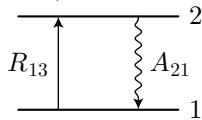
Now we will make the *adiabatic approximation* to eliminate energy level 3. The idea is that the A_{32} term in the dN_3/dt equation is a *damping* term (i.e., like a friction term) for N_3 . In the limit of large A_{32} ($A_{32} \gg R_{13}$), the adiabatic approximation says that N_3 damps to equilibrium very quickly, and so we can assume that it is always in quasiequilibrium. Thus, the equilibrium condition for N_3 gives

$$\frac{dN_3}{dt} \approx 0 \implies N_3 \approx \frac{R_{13}}{A_{32} + R_{13}} N_1 \approx \frac{R_{13}}{A_{32}} N_1. \quad (15.58)$$

The population N_3 is said to *adiabatically follow* the N_1 population, and it is now a redundant quantity since it is completely determined by N_1 . Furthermore, $R_{13}/A_{32} \ll 1$, so $N_3 \ll N_1$, so level 3 is essentially unpopulated. Then we can use $N_3 - N_1 \approx -N_1$ and write effective rate equations (15.57) for the remaining two levels as

$$\begin{aligned}\frac{dN_2}{dt} &= R_{13}N_1 - A_{21}N_2 \\ \frac{dN_1}{dt} &= A_{21}N_2 - R_{13}N_1.\end{aligned}\quad (\text{rate equations, adiabatic approximation}) \quad (15.59)$$

These are the rate equations for a two-level atom, but *with an asymmetric pump*.



The presence of the third level breaks the symmetry of the pump, and now we can get a population inversion if the pump rate exceeds the spontaneous decay rate.

In steady state, $dN_2/dt = dN_1/dt = 0$, which gives the equilibrium population ratio

$$\frac{N_2}{N_1} = \frac{R_{13}}{A_{21}}. \quad (15.60)$$

Again, we see that we have a population inversion if the ratio exceeds unity and thus if the pump rate exceeds the spontaneous emission rate from level 2, $R_{13} > A_{21}$. To compute the inversion, note that

$$N_2 - N_1 = \left(\frac{N_2}{N_1} - 1 \right) N_1 = \left(\frac{R_{13}}{A_{21}} - 1 \right) N_1. \quad (15.61)$$

Similarly,

$$N_2 + N_1 = N = \left(\frac{R_{13}}{A_{21}} + 1 \right) N_1, \quad (15.62)$$

where N is the total atomic number density, and we are neglecting the population N_3 . Using Eq. (15.62) to eliminate N_1 in the right-hand side of Eq. (15.61), we find an inversion of

$$N_2 - N_1 = \left(\frac{R_{13} - A_{21}}{R_{13} + A_{21}} \right) N = \left(\frac{R' - 1}{R' + 1} \right) N,$$

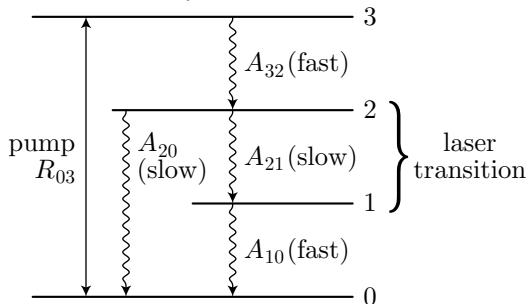
(population inversion, three-level atom, adiabatic approximation) (15.63)

where $R' := R_{13}/A_{21}$. Again, we can see that a pump rate $R' > 1$ corresponds to a population inversion.

15.4.2 Four-Level Laser

Another important level scheme is the four-level laser, where we add yet another extra level. The Nd:YAG laser is one example of a four-level laser. Four-level media are particularly important since, as we will see, they are more efficient in general than three-level media.

The level scheme is shown here. The new level 0 is now the ground state, with the laser transition ($2 \rightarrow 1$) sandwiched between the two auxiliary levels.



Other decay paths besides the ones we have included, but we assume them to be slow, in which case they do not affect our basic results. The rate equations are

$$\begin{aligned}\frac{dN_3}{dt} &= -R_{03}(N_3 - N_0) - A_{32}N_3 \\ \frac{dN_2}{dt} &= A_{32}N_3 - (A_{21} + A_{20})N_2 \\ \frac{dN_1}{dt} &= A_{21}N_2 - A_{10}N_1 \\ \frac{dN_0}{dt} &= A_{20}N_2 + A_{10}N_1 + R_{03}(N_3 - N_0).\end{aligned}$$

(rate equations, four-level atom) (15.64)

Again, we will assume that all decay rates except those from level 2 are fast: $A_{32}, A_{10} \gg R_{03}, A_{21}, A_{20}$.

Now *both* levels 1 and 3 have fast damping terms, so we can adiabatically eliminate both of them. Setting $dN_3/dt \approx 0$ leads to

$$N_3 \approx \frac{R_{03}}{A_{32}}N_0 \ll N_0, \quad (15.65)$$

while setting $dN_1/dt \approx 0$ leads to

$$N_1 \approx \frac{A_{21}}{A_{10}}N_2 \ll N_2. \quad (15.66)$$

Again, we can neglect the small populations in levels 1 and 3, and in particular $N_3 - N_0 \approx -N_0$. Thus, the effective rate equations for the remaining two levels are

$$\begin{aligned}\frac{dN_2}{dt} &= R_{03}N_0 - (A_{21} + A_{20})N_2 \\ \frac{dN_0}{dt} &= (A_{21} + A_{20})N_2 - R_{03}N_0.\end{aligned}$$

(rate equations, four-level atom, adiabatic approximation) (15.67)

Again, this corresponds to a two-level atom with an asymmetric pump. However, the difference is that the ground level 0 of this effective atom is *not* involved in the laser transition. The ground level of the laser transition is level 1, which is almost completely depleted, and this is why it is easier to develop an inversion in the four-level laser than in the three-level version.

In steady state,

$$\frac{N_2}{N_0} \approx \frac{R_{03}}{A_{21} + A_{20}}. \quad (15.68)$$

Using $N_0 + N_2 \approx N$ and a procedure similar to the three-level calculation, we can eliminate N_0 to get an expression for N_2 . Since $N_1 \approx 0$, this is the population inversion:

$$N_2 - N_1 \approx N_2 \approx \left(\frac{R'}{1 + R'} \right) N,$$

(population inversion, four-level atom, adiabatic approximation) (15.69)

where $R' := R_{03}/(A_{21} + A_{20})$. Note that there is an inversion for *any* pump $R' > 0$, whereas we saw that in the three-level laser, the inversion was of the form $N(R' - 1)/(R' + 1)$, so that a certain minimum pump was required.

Again, the whole point of this is: **auxiliary levels break the symmetry of the pump and allow for population inversion**. Other auxiliary levels might make the pumping process easier.

15.5 Gain Coefficient

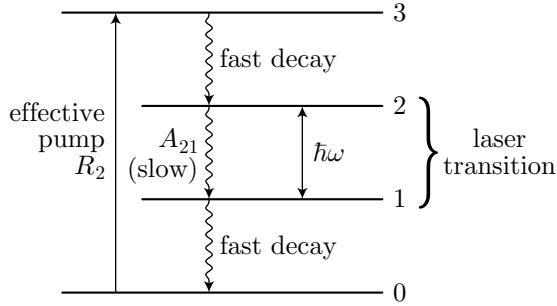
With the specifics of the atomic media in hand, we can proceed to explore the gain coefficient in more detail. The complication is that now we will have to consider the interaction of the gain medium with *two* optical

fields: the pump field and the laser field. For simplicity, we will only do this for the four-level laser, but similar results hold for the three-level laser.

Going back to the four-level laser scheme, we will consider it to be pumped by a monochromatic laser at a rate

$$R_2 = \frac{\sigma_{\text{pump}} I_{\text{pump}}}{\hbar \omega_{\text{pump}}}, \quad (15.70)$$

and we can regard this to be a unidirectional pump since we assume level 3 will decay quickly to level 2.



Recall from Eq. (15.67) that the effective rate equation is

$$\frac{dN_2}{dt} = R_2 N_0 - A_2 N_2 - \frac{\sigma I}{\hbar \omega} (N_2 - N_1), \quad (15.71)$$

where $A_2 := A_{21} + A_{20}$. The last term accounts for the interaction of the atom with the laser field, so that σ , I , and ω all refer to the laser field. Again, for the four-level laser we can neglect N_1 compared to N_2 , with the result

$$\frac{dN_2}{dt} \approx R_2 N_0 - A_2 N_2 - \frac{\sigma I}{\hbar \omega} N_2. \quad (15.72)$$

Of course, the stimulated absorption/emission terms for the pump and laser light have the same form, except that they interact with different levels, and they have different effective signs after the adiabatic elimination of levels 1 and 3.

Steady state occurs when $dN_2/dt = 0$, which gives the steady-state population

$$N_2 = \frac{R_2 N_0}{A_2 + \frac{\sigma I}{\hbar \omega}}. \quad (15.73)$$

Following the algebraic steps of the last section, when we have an equation of the form $N_2 = (\text{stuff})N_0$ with $N_0 + N_2 \approx N$, the solution is $(N_2/N) = (\text{stuff})/(1 + \text{stuff})$. Thus,

$$\frac{N_2}{N} = \frac{\frac{R_2}{A_2 + \frac{\sigma I}{\hbar \omega}}}{1 + \frac{R_2}{A_2 + \frac{\sigma I}{\hbar \omega}}} = \frac{R_2}{A_2 + \frac{\sigma I}{\hbar \omega} + R_2} = \frac{R_2 \tau_2}{1 + R_2 \tau_2 + \frac{I}{I_{\text{sat}}}}, \quad (15.74)$$

where

$$I_{\text{sat}} = \frac{\hbar \omega}{\sigma \tau_2}, \quad (15.75)$$

and $\tau_2 = A_2^{-1}$.

Thus, the gain coefficient for the four-level laser is

$$\gamma(\omega) = \sigma(\omega)[N_2 - N_1] = \frac{\sigma(\omega) R_2 \tau_2 N}{1 + R_2 \tau_2 + \frac{I}{I_{\text{sat}}}}. \quad (\text{gain coefficient, four-level laser}) \quad (15.76)$$

From this expression, we can see that there are two types of saturation in the gain medium.

1. **Gain saturation.** Consider the limit of a weak pump ($R_2\tau_2 \ll 1$). Then the gain coefficient (15.74) becomes

$$\gamma(\omega) = \frac{\sigma(\omega)R_2\tau_2 N}{1 + \frac{I}{I_{\text{sat}}}}. \quad (15.77)$$

In the small-signal regime ($I \ll I_{\text{sat}}$), the gain coefficient takes on the small-signal value

$$\gamma_0(\omega) = \sigma(\omega)R_2\tau_2 N, \quad (\text{small-signal gain coefficient, weak pump}) \quad (15.78)$$

so that

$$\gamma(\omega) = \frac{\gamma_0(\omega)}{1 + \frac{I}{I_{\text{sat}}}}. \quad (\text{gain coefficient, weak pump}) \quad (15.79)$$

Again, the saturation intensity is defined such that at the saturation intensity, the gain coefficient drops to half its small-signal value:

$$\gamma(I = I_{\text{sat}}) = \frac{\gamma_0}{2}. \quad (15.80)$$

In the limit of large intensity, $\gamma(I \rightarrow \infty) = 0$. This is precisely the gain-saturation effect that we discussed before in the context of the simple laser model of Section 15.1.5.

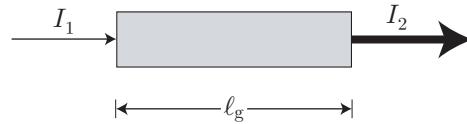
2. **Pump saturation.** In the limit of a very strong pump, $R_2\tau_2 \rightarrow \infty$, the gain coefficient saturates at a value

$$\gamma(\omega) \rightarrow \sigma(\omega)N. \quad (\text{strong-pump gain coefficient}) \quad (15.81)$$

This limit has a nice physical interpretation: there is only a finite number N of atoms, and each can only add an energy $\hbar\omega$ of energy as soon as it is stimulated by an incoming laser photon. Thus, the gain can also be limited by the finite number of atoms in the gain medium, so that eventually the gain coefficient does not continue to increase with an increasing pump intensity.

15.5.1 Gain in a Medium of Finite Length

Now let's revisit the gain medium, and compute the *gain* of the medium in terms of the gain coefficient. The medium has length ℓ_g , and the input and output intensities are I_1 and I_2 , respectively.



The gain coefficient is defined such that

$$\frac{dI(\omega)}{dz} = \gamma(\omega)I(\omega), \quad (15.82)$$

so we can write this explicitly in terms of the small-signal gain coefficient as

$$\frac{1}{I(\omega)} \frac{dI(\omega)}{dz} = \frac{\gamma_0(\omega)}{1 + \frac{I}{I_{\text{sat}}}}. \quad (15.83)$$

Recall that the small-signal gain coefficient $\gamma_0(\omega)$ is defined by Eq. (15.78) in the weak-pump regime, and the saturation intensity I_{sat} is given by Eq. (15.75). Since the right-hand side of this differential equation involves the intensity, the solution is more complicated than a simple growing exponential.

To find the solution, we solve for dz :

$$\left(\frac{1}{I(\omega)} + \frac{1}{I_{\text{sat}}} \right) dI(\omega) = \gamma_0(\omega) dz. \quad (15.84)$$

Integrating over the length of the gain medium,

$$\int_{I_1}^{I_2} \left(\frac{1}{I(\omega)} + \frac{1}{I_{\text{sat}}} \right) dI(\omega) = \int_0^{\ell_g} \gamma_0(\omega) dz, \quad (15.85)$$

and thus

$$\log \left(\frac{I_2}{I_1} \right) + \frac{I_2 - I_1}{I_{\text{sat}}} = \gamma_0(\omega) \ell_g. \quad (15.86)$$

Defining the *single-pass* gain as

$$G := \frac{I_2}{I_1}, \quad (15.87)$$

we can rewrite Eq. (15.86) as

$$\log G + \frac{I_1}{I_{\text{sat}}} (G - 1) = \gamma_0(\omega) \ell_g. \quad (15.88) \quad (\text{single-pass gain-factor equation})$$

This is the central result of this section. We cannot get a general expression for G in terms of γ_0 , but rather we can get a transcendental equation that relates them.

In the small-signal regime ($I_1, I_2 \ll I_{\text{sat}}$), we can ignore the $(G - 1)$ term in Eq. (15.88). In this case, we recover the small-signal result:

$$G_0 = \exp[\gamma_0(\omega) \ell_g]. \quad (15.89) \quad (\text{small-signal gain factor})$$

In the high-intensity limit ($I_1 \gg I_{\text{sat}}$), we can ignore the $\log G$ term in Eq. (15.88), with the result

$$I_2 - I_1 = [\gamma_0(\omega) I_{\text{sat}}] \ell_g. \quad (15.90) \quad (\text{high-intensity gain})$$

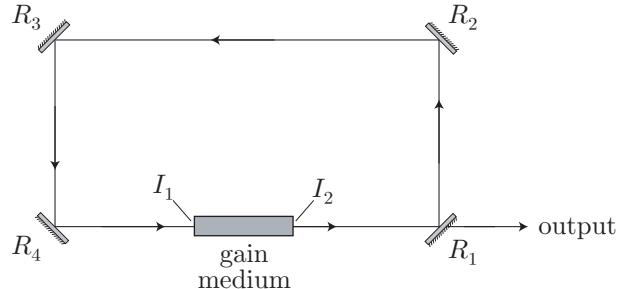
The interpretation of this relation is that the gain medium *adds* an amount $\gamma_0 I_{\text{sat}} \ell_g$ to the input intensity. Writing out the gain coefficient and the saturation intensity,

$$\gamma_0 I_{\text{sat}} \ell_g = [\sigma(\omega) R_2 \tau_2 N] \left[\frac{\hbar \omega}{\sigma(\omega) \tau_2} \right] \ell_g = (N \ell_g)(\hbar \omega) R_2. \quad (15.91)$$

Thus, the gain medium adds an intensity given by the number of atoms per unit of transverse area, multiplied by the photon energy, multiplied by the pumping rate per atom. We can conclude from this that we can only extract energy from the medium at the rate that energy is pumped into it. That is, an intense input signal extract all possible energy from the gain medium.

15.6 Laser Output: CW

Now we'll examine in more detail the output characteristics of a continuous-wave (CW) laser, and in particular look at its efficiency and how to optimize it. Let's consider a ring-cavity laser as shown, so that our above analysis of the single-pass gain remains valid (in a linear cavity, the two counterpropagating waves saturate each other, making things more complicated).



In addition, we will assume that there is an “optical diode” in the cavity to force the laser to operate in only one direction (typically, an optical diode consists of a Faraday rotator and a birefringent plate, which leaves the polarization unaffected in only one direction). If all the losses are due to partial mirror transmission, then the survival probability is

$$P_s = R_1 R_2 R_3 R_4. \quad (15.92)$$

This expression can easily be generalized to include other cavity losses (e.g., due to reflections at the gain medium surfaces). In steady state, the gain equals the loss, and so

$$GP_s = 1. \quad (15.93)$$

Thus, the gain equation (15.88) becomes

$$\log\left(\frac{1}{P_s}\right) + \frac{I_2}{I_{\text{sat}}} (1 - P_s) = \gamma_0(\omega) \ell_g. \quad (15.94)$$

Solving for I_2 , we find

$$I_2 = \frac{I_{\text{sat}}[\gamma_0 \ell_g - \log(1/P_s)]}{1 - P_s}. \quad (15.95)$$

Ideally, there is no loss in the output coupler (mirror), so that $T_1 = 1 - R_1$. Then the output intensity is simply $I_2 T_1$.

15.6.1 Optimum Output

How can we optimize the output of the laser? In other words, given a certain gain and loss in the laser, what reflectance should we choose for the output coupler to maximize the output power? Let’s work in the low-loss approximation where $(1 - P_s)$ is small. Then

$$\log\left(\frac{1}{P_s}\right) = -\log[1 - (1 - P_s)] \approx 1 - P_s. \quad (15.96)$$

Also,

$$1 - P_s = 1 - (1 - T_1) R_2 R_3 R_4 = (1 - R_2 R_3 R_4) + T_1 R_2 R_3 R_4 = P_l + T_1 (1 - P_l) \approx L + T_1, \quad (15.97)$$

where we have defined the loss probability

$$P_l := 1 - R_2 R_3 R_4, \quad (15.98)$$

and we have made the low-loss approximation $1 - P_l \approx 1$. Then from Eq. (15.95),

$$I_{\text{out}} = \frac{T_1 I_{\text{sat}} [\gamma_0 \ell_g - \log(P_l + T_1)]}{P_l + T_1} = T_1 I_{\text{sat}} \left(\frac{\gamma_0 \ell_g}{P_l + T_1} - 1 \right). \quad (15.99)$$

This is an optimum when $\partial I_{\text{out}} / \partial T_1 = 0$, which gives an optimum transmission

$$T_1(\text{opt}) = \sqrt{\gamma_0 \ell_g P_l} - P_l. \quad (15.100)$$

The corresponding optimal output intensity is

$$I_{\text{out}}(\text{opt}) = I_{\text{sat}} \left(\sqrt{\gamma_0 \ell_g} - \sqrt{P_l} \right)^2. \quad (15.101)$$

Sometimes, in real lasers it is necessary to choose an output coupler that has lower reflectance than this theoretical minimum. This can enhance the stability of the laser at the expense of some output power.

15.6.2 Quantum Efficiency

In the ideal case, there is no loss ($P_l = 0$). Then the optimal output intensity is

$$I_{\text{out}}(\text{opt}) = \gamma_0 \ell_g I_{\text{sat}} = \hbar \omega R_2 \ell_g N. \quad (15.102)$$

If A is the area of the gain medium, then the corresponding output power is

$$P_{\text{out}}(\text{opt}) = A I_{\text{out}}(\text{opt}) = \hbar \omega R_2 N (\ell_g A), \quad (15.103)$$

where $\ell_g A$ is the volume of the gain medium, and thus $N \ell_g A$ is the total number of atoms in the gain medium. Thus the output power is $P_{\text{out}} = \hbar \omega \cdot (\text{total pump rate})$. In other words, the output power consists of one output photon energy per input photon. This sounds like perfect efficiency, but it isn't quite: the pump photon is more energetic than the laser photon because they correspond to different energy transitions.

Suppose that η_q is the laser efficiency in this ideal case. Then we can write η_q as the ratio of the output power to the input power:

$$\eta_q = \frac{\hbar \omega \cdot (\text{total pump rate})}{\Delta E_{\text{pump}} \cdot (\text{total pump rate})} = \frac{\hbar \omega}{\Delta E_{\text{pump}}}. \quad (15.104)$$

Here, ΔE_{pump} is the energy of the pumping transition. This ideal efficiency is called the **quantum efficiency** of the laser, and represents the fundamental limit to efficiency for any given laser gain medium. Other limits to efficiency include cavity losses (extraction efficiency) and the fact that not all pump photons lead to stimulated emission (pumping efficiency).

For example, a Nd:YAG laser has a pump wavelength of $\lambda_{\text{pump}} = 810 \text{ nm}$ when pumped from a diode laser (these lasers can also be flashlamp-pumped in the blue with correspondingly less efficiency). The laser wavelength is $\lambda_{\text{laser}} = 1064 \text{ nm}$, giving a relatively high quantum efficiency of $\eta_q \sim 0.8$. A He–Ne laser, on the other hand, operates by electric discharge to a state of $\sim 20 \text{ eV}$ in He. The energy is transferred via collision to a state of $\sim 20 \text{ eV}$ in Ne. The laser is on a Ne transition, with a 632.8 nm wavelength corresponding to $\sim 2 \text{ eV}$. The quantum efficiency in this case is much lower, around $\eta_q \sim 0.1$. This is one reason why gas-discharge lasers tend to be less efficient than optically pumped lasers: the discharge leads to very high excitation, and most of this excitation energy is not extracted by the laser field.

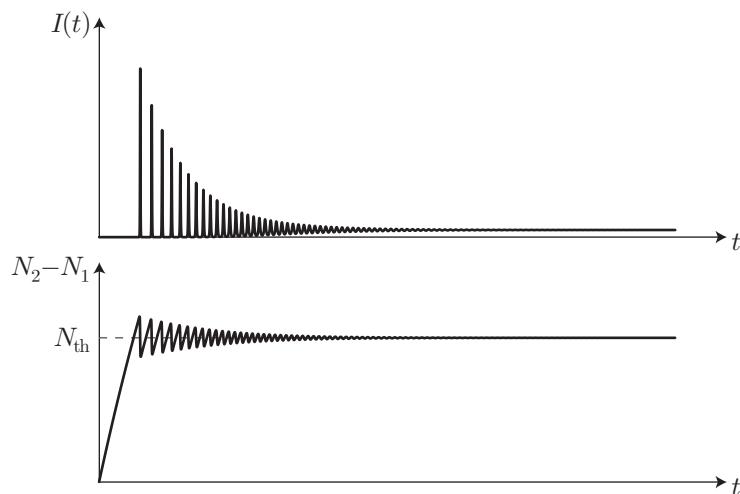
15.7 Laser Output: Pulsed

Many lasers don't work in CW mode, but instead work in pulsed mode. There are many reasons why pulsed lasers are preferred in some applications. For example, pulsed lasers are capable of much higher peak powers than CW lasers (with state-of-the-art lasers reaching the petawatt range). With high powers or certain gain media, it might not be possible to sustain the pump or to cool the laser system efficiently enough. Also, when the laser is used to probe or visualize a fast process, short pulses may be needed to strobe and effectively freeze the dynamics during the probe.

15.7.1 Laser Spiking

In the treatment of CW lasers above, we focused on the steady-state behavior of lasers. To treat pulses, though we must treat the laser *dynamics*. One kind of pulsed laser output occurs in the initial transient behavior of a CW laser. From the simple laser model of Section 15.1.5 (see the plot on p. 287), we might expect a CW laser to turn on smoothly, especially for a weak or slowly activated pump where the laser intensity adiabatically follows the inversion until steady state is reached.

However, especially with high-gain lasers such as flashlamp-pumped Nd:YAG lasers, there is another possibility, called **laser spiking**, where the dynamics of the inversion $N_2 - N_1$ and the intensity I conspire to produce intensity oscillations. The laser spikes are also called **relaxation oscillations**, although this term often refers to the low-amplitude oscillations after the gain spikes settle down to near equilibrium (or when the laser is only slightly perturbed).

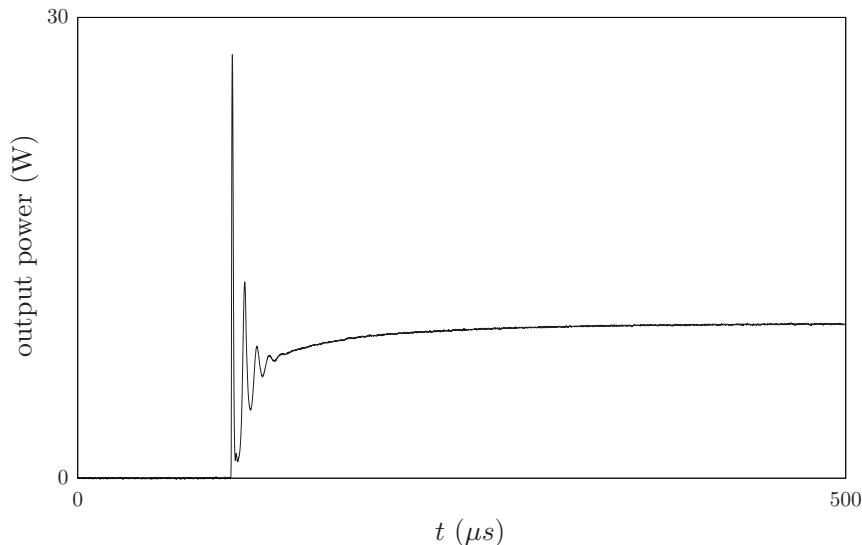


As in the simulation shown here, the initial transient can look like a periodic sequence of short intensity pulses. This process occurs due to the following process.

1. A fast pump rapidly excites the gain medium far above the threshold inversion. This happens on a fast time scale, before the cavity intensity can respond.
2. The intensity, as a result, builds up rapidly, quickly depleting the gain medium as the laser oscillates. The cavity intensity becomes so large that the gain medium drops below threshold.
3. We are more or less at the initial condition again, so the process repeats.

This process causes the intensity oscillations or laser spikes. Since both the laser output and spontaneous emission act as damping (friction-like) processes, the oscillations damp away and the laser eventually relaxes to steady state.

Below is the measured intensity from a diode-laser pumped Yb:fiber laser operating at 1085 nm that shows laser spiking.

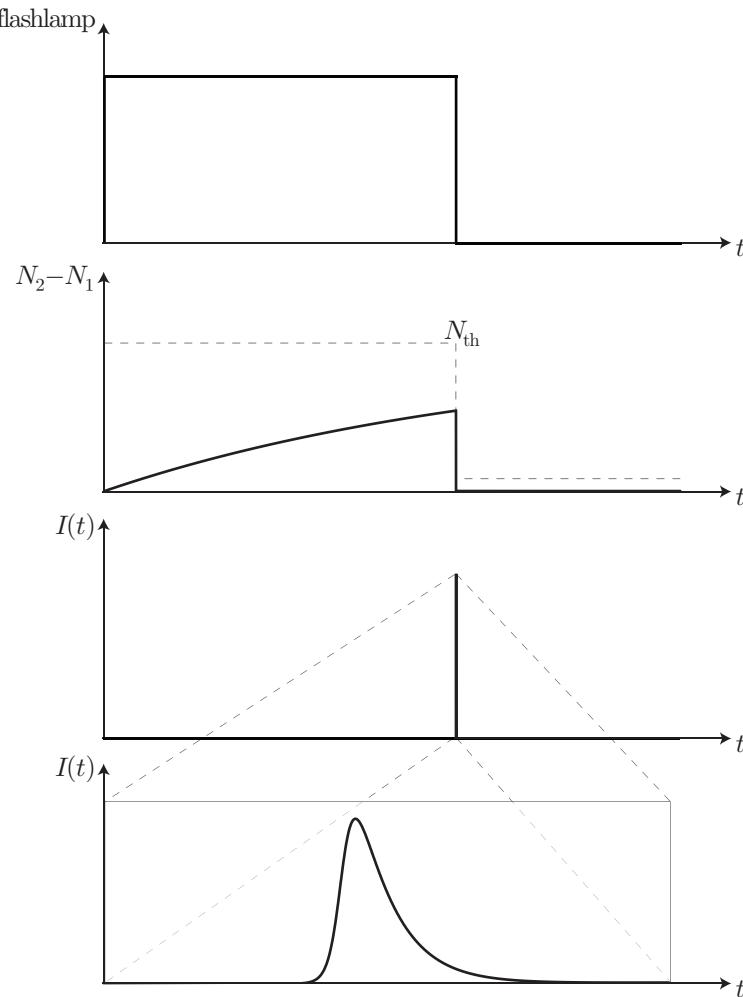


The pump diodes are turned on quickly, and the resulting laser spikes are clearly visible before the laser power settles down to the steady-state value of about 10 W. During the first spike, the laser jumps to well above the steady-state power.

15.7.2 Q-Switching

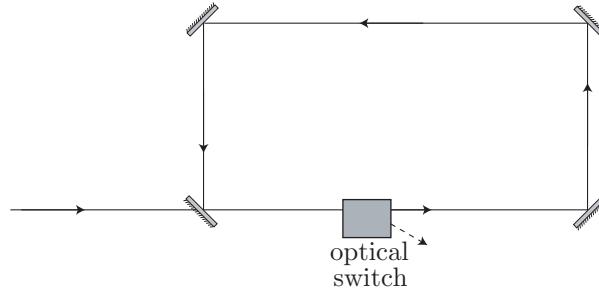
The idea behind Q -switching is to avoid the laser spiking by compressing all the relaxation oscillations into a single, big pulse. For example, for a flashlamp-pumped laser, the cavity is made lossy (i.e., the cavity Q is “switched low”) during the flashlamp pulse to allow the inversion to build up way beyond the high- Q threshold value. Near the end of the flashlamp pulse, the cavity is restored to its former high- Q condition, so that the laser finds itself way above threshold. In this case, the intensity is extracted in one short, enormous pulse, as shown in the simulation below. For a high-gain laser (such as Nd:YAG with an unstable resonator), the output pulse can be on the order of the cavity round-trip time (10 ns pulses are typical for a 1 m linear Nd:YAG laser).

Several techniques are available for switching the cavity Q . One common method is to use a **Pockels cell** (electro-optic crystal) in the cavity, which is essentially a nonlinear crystal that becomes birefringent when high voltage is applied across it. When the voltage is applied, the polarization is rotated going through the cell, and a polarizer in the cavity strongly attenuates the mode. (Actually it turns out that a polarizer isn’t necessary, the birefringence is sufficient to spoil the Q .) Another method is to use a rotating cavity mirror, so that the cavity Q is only high at the moment when the mirror reaches the right angle. Finally, an “automatic” Q -switch can be made from a **saturable absorber**. This is a medium that absorbs light, but due to the same saturation effect as in gain medium, high intensities pass through the medium nearly unaffected (this saturation effect is also called “bleaching”). The saturable absorber will spoil the Q until the gain reaches the point where an amplified spontaneous wave exceeds the saturation threshold, at which point the absorber bleaches and the cavity Q is switched high.



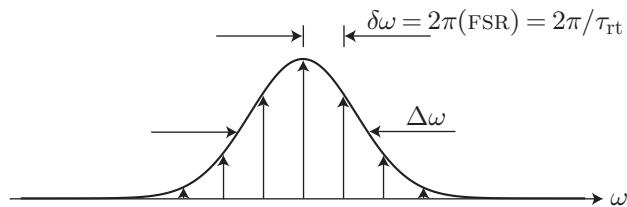
15.7.3 Cavity Dumper

Another simple source of pulse laser light is the **cavity dumper** (see Problem 7.3). Actually, this is more of a device to convert CW light into pulses, since it is not itself a laser. The idea is to take a ring cavity and inject CW laser light into one port. The intensity in the cavity builds up to $\sim(\mathcal{F}/2\pi)I_{\text{in}}$ in steady state, so a high cavity finesse is important in attaining large pulse intensities. Once the light has built up, the idea is to suddenly turn on an optical switch (such as an acousto-optic modulator) inside the cavity to couple out the high-intensity light. For efficient out-coupling, the pulse duration is equal to the round-trip time of the cavity.

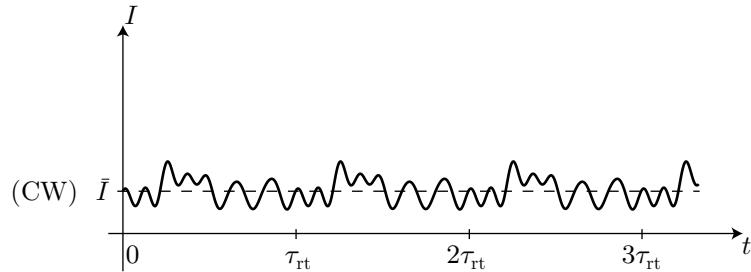


15.7.4 Mode Locking

Mode locking is another method for converting a CW laser into a pulsed laser. This is also the method that provides some of the shortest possible pulses. The idea is that a gain medium with inhomogeneous broadening can support simultaneous lasing on many modes, as we discussed in Section 15.3.2. This is because different modes interact with different atoms, and thus do not compete for pump energy. Suppose that the gain medium can support modes in a range of order $\Delta\omega$. The modes are spaced by the usual cavity spacing $\delta\omega = 2\pi/\tau_{\text{rt}}$.



Usually, the modes will have no particular phase relationship. The laser output is thus the interference of all the beams, but with random phases, giving a noisy output that fluctuates about the mean intensity \bar{I} .



But what if it is possible to give them all the same relative phase? Recall the Poisson sum rule (Problem 3.7),

$$\sum_{n=-\infty}^{\infty} \cos(2\pi nt) = \sum_{n=-\infty}^{\infty} \delta(t - n). \quad (15.105)$$

We can interpret the left-hand side as the sum of many waves of equal amplitude, frequency $\omega_n = 2\pi n$, and zero phase offset. The right-hand side is a periodic series of arbitrarily short pulses. In terms of the laser,

if we can *lock* the relative phases of many modes, we can expect the output to be a series of short, intense pulses.

Let's consider a slightly less simple model of mode locking. Recall from Section 5.8 that the interference of N equal-amplitude waves of the form $\exp[i(n-1)\phi]$ is

$$I = I_0 \frac{\sin^2(N\phi/2)}{\sin^2(\phi/2)}, \quad (15.106)$$

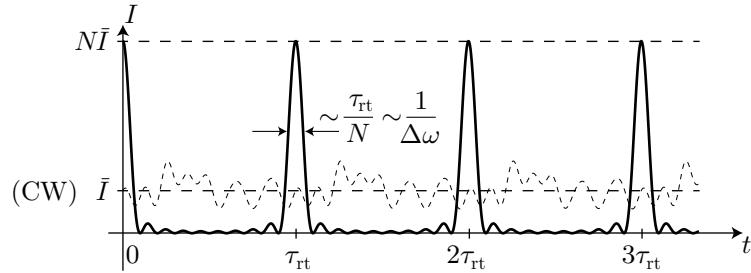
where I_0 is the intensity of a single wave. Interpreting $\phi \rightarrow (\delta\omega)t$, where $\delta\omega$ is much smaller than the central optical frequency ω_0 , we have a sum of the form

$$\left| \sum_{n=1}^N \sqrt{I_0} e^{i\omega_0 t} e^{i(n-1)(\delta\omega)t} \right|^2, \quad (15.107)$$

which leads to an intensity

$$I = I_0 \frac{\sin^2[N(\delta\omega)t/2]}{\sin^2[(\delta\omega)t/2]}. \quad (15.108)$$

This is a sequence of short pulses, with duration $\sim 1/\Delta\omega$, period τ_{rt} , and height $N\bar{I}$. The output for 10 equal-amplitude, phase-locked modes is shown here, compared to the random-phase version shown as a dashed line.



If there are many modes available, the pulses can be extremely intense and short. For example, a Ti:sapphire laser is popular for mode-locking because of its very wide emission range, spanning 600-1050 nm. The full width at half maximum of the emission range is more like $\Delta\lambda \sim 200$ nm (from 700-900 nm), which gives $\Delta\omega/2\pi \sim 9.5$ THz, so the pulse width is $\delta t \sim 1/\Delta\omega \sim 1.7$ fs. Real lasers have much longer pulse lengths due to a number of effects, but fs-scale pulses are now routine through this method.

There several ways to mode-lock a laser. One way is to “encourage” only the large intensities of the mode-locked operation and “discourage” the random-phase CW operation. For example, just as in the *Q*-switched laser, a saturable absorber can select the mode-locked behavior, since the pulses will bleach the absorber, while the CW light will be blocked. A similar effect is achieved by the **Kerr lens**, which takes advantage of the nonlinear optical behavior of the gain medium. At high intensities, the refractive index of the gain medium increases with the applied intensity. Thus, a high-intensity Gaussian beam effectively creates a lens for itself in the gain medium. An appropriately placed aperture (or effective aperture) can block the cavity light except when the Kerr effect focuses the mode through it, thus selecting again the high-intensity pulses. Lasers can also be mode-locked *actively*. For example, an electro-optic phase modulator in the cavity can put sidebands on each mode, separated by the modulation frequency (see Problem 13.1). If the modulation frequency matches the free spectral range of the laser cavity (or possible some multiple of it), the sidebands lead to coupling of the modes. The modes then act as coupled nonlinear oscillators, and tend to synchronize (you can also think of the sidebands *seeding* the other modes with light of the proper phase). A similar effect can be achieved by modulating the *amplitude* of the intracavity light, e.g., by periodically driving an acousto-optic modulator, where the modulation frequency matches the cavity free-spectral range.

15.8 Exercises

Problem 15.1

Consider a two-mirror cavity with a gain medium. The mirrors have intensity reflection coefficients R_1 and R_2 . The gain medium is thin and has a round-trip intensity gain coefficient G . (All this is in the text, the point is to go through it and fill in the details.)

- Write down the cavity intensity at time $t + \tau_{\text{rt}}$ in terms of the cavity intensity at time t , where τ_{rt} is the round-trip time of the cavity.
- Expand the above expression to first order in τ_{rt} , and hence obtain a differential equation for $I(t)$. What are the conditions under which this approximation is justified?
- Apply the result of part (b) to a cavity *without* a gain medium, to derive an expression for the cavity photon lifetime τ_p , defined as the time required for the cavity intensity to decay to $1/e$ times its initial value.

Problem 15.2

This is a prelude to problems below that require numerically solving differential equations (i.e., laser rate equations). Get thee to a computer running Mathematica, and run the commands in the following notebook:

http://atomoptics.uoregon.edu/~dstreck/teaching/mathematica/ode_example.nb

This notebook is also posted in pdf format here:

http://atomoptics.uoregon.edu/~dstreck/teaching/mathematica/ode_example.pdf

Run the supplied commands to solve two simple systems of ordinary differential equations. Give the analytic solutions to these equations, and comment on whether the numerical solutions are sensible. Make sure to understand the syntax, so you can apply the same commands to plotting the solutions to the laser rate equations in later problems.

Problem 15.3

- Obtain an analytic solution to the laser cavity intensity equation,

$$\frac{dI}{dt} = \frac{(G(I) R_1 R_2 - 1)}{\tau_{\text{rt}}} I, \quad (15.109)$$

where the gain G saturates according to

$$G(I) = 1 + \frac{g_0}{1 + \frac{I}{I_{\text{sat}}}}, \quad (15.110)$$

and where g_0 is a constant. Treat the case of a perfect cavity, $R_1 R_2 = 1$ (for this part only). You will not be able to obtain an explicit solution, $I(t)$, but you can get an implicit solution $t(I)$. (Hint: we did something very similar to this in the text.)

- Find the steady-state intensity for a laser modeled by the above equations.
- Plot the solution $I(t)$ to the above equations. For the purposes of the plot, take $g_0 = 2$, $I_0 = 1 \mu\text{W}/\text{cm}^2$, $I_{\text{sat}} = 1 \text{ W}/\text{cm}^2$, $R_1 R_2 = 0.99$, and plot the solution as a function of t/τ_{rt} . Solve the differential equation numerically to make the plot.

Problem 15.4

Starting with the gain relation for a medium of length ℓ_g ,

$$\log G + \frac{I_1}{I_{\text{sat}}} (G - 1) = \gamma_0(\omega) \ell_g, \quad (15.111)$$

show that in the limit of a thin gain medium, the gain coefficient takes the form that we considered in the simple laser model of Eq. (15.9):

$$G(I) = 1 + \frac{g_0}{1 + \frac{I}{I_{\text{sat}}}}. \quad (15.112)$$

What is g_0 in terms of the gain parameters?

Problem 15.5

Examine the validity of the adiabatic approximation in the three-level laser model as follows.

- (a) Write down the rate equations for the three-level atom, both with and without the adiabatic elimination of the third level. Then modify these in two ways. First, eliminate N_1 by letting N be the total number density of the atoms, and then use conservation of atom number. Write the equations in terms of the normalized populations $n_j := N_j/N$. Second, measure time in units of $1/A_{21}$. In other words, write the equations in terms of the scaled time parameter $t' := A_{21}t$.
- (b) Assume a pumping rate of $R_{13} = 2A_{21}$. Make plots of $n_2(t')$ for different values of A_{32} , corresponding to where the adiabatic approximation is bad, marginal, and good (say, $A_{32}/R_{13} = 1, 10$, and 100). Include the plot of the solution of the adiabatic rate equations for comparison. In all cases start with the atoms entirely in the ground state.

Problem 15.6

Suppose you build your own He-Ne laser. You use a symmetric cavity (geometrically speaking) by sealing two concave mirrors with the same radius of curvature $R = 100$ cm onto the ends of a glass tube of length L . You then fill the tube with the appropriate mixture of helium and neon, install proper electrodes, hook up the power supply, and so on. Assume that the laser operates on the red line at 632.8 nm, and that the refractive index of the helium/neon mixture is $n = 1$. Also assume that the output mirror has a reflectance of $R_{\text{out}} = 99\%$, while the other (back-end) mirror is perfectly reflecting.

- (a) Gas lasers have relatively small single-pass gain, and so the light must pass through the gain medium many times before exiting the cavity to sustain oscillation. For what range of length L might you expect a working laser?
- (b) Suppose you settle on a length $L = 75$ cm. Calculate the beam waist parameter w_0 , the spot size w at the mirrors, and the far-field divergence angle of your laser, assuming it lases.
- (c) Calculate the free spectral range of the resonator.
- (d) Calculate the finesse and Q factor of the resonator, and the order q of the lasing mode.
- (e) Calculate the frequency width of the resonator modes. This is a very crude estimate of the frequency width of the laser output, as it assumes that the gain medium plays negligible role in determining the frequency spectrum. This is only a good estimate if the laser operates close to threshold.
- (f) The He-Ne gain medium can support laser operation within a frequency range of about 1.3 GHz due to Doppler broadening at room temperature. Do you expect your laser to operate in a single longitudinal mode? Assume that the laser operates in a single *transverse* mode ($\text{TEM}_{0,0}$).
- (g) Suppose that the laser can oscillate in either the $\text{TEM}_{0,0}$ or the $\text{TEM}_{1,0}$ mode. For the same *longitudinal* mode, calculate the difference in the lasing frequencies of these two modes.
- (h) Suppose you change the temperature of the laser resonator by 1°C , say by cooling it off with air current from a fan. By how much does the frequency of the output light change? Assume that you constructed your laser using a glass tube made out of BK7 borosilicate crown glass, and that the laser is not too much hotter than room temperature. (Part of this exercise is to locate a reputable source for optical material properties, such as the thermal expansion coefficient of a popular optical glass.)
- (i) Calculate the minimum round-trip intensity gain G required for your laser to oscillate (lase).

- (j) Suppose that the output power of your laser is 5 mW. Calculate the optical power of the light circulating in the resonator (assuming that all the power is either reflected from or transmitted through the output mirror) and the peak intensity inside the resonator.
- (k) You notice that your laser is behaving strangely. Upon closer inspection, you find that you managed to trap a tiny bug inside the cavity while you were filling it. You observe that the bug is resting near the center of the output mirror, and as it moves around slightly the laser jumps between the Gaussian and donut modes. Estimate the diameter of the bug, using these assumptions and observations: (1) the bug is a perfectly absorbing disc, and is otherwise impervious to the laser radiation; (2) in addition to the transmission through the output mirror, the only sources of loss are absorption by the bug and loss around the edges of the 2 mm diameter mirrors; (3) the bug is much smaller than the beam diameter; (4) the TEM_{0,0} and donut modes compete so that only the mode with the least loss lasers at any given time; and (5) the gains of the two modes in the medium are roughly the same.

Problem 15.7

In this problem you will examine numerically the dynamics of a CW Nd:YAG laser. Assume that the laser is well modeled as a four-level laser producing output at a wavelength of 1.064 μm , and that the refractive index of YAG is $n = 1.83$ at the laser wavelength (be careful to note that 1.064 μm is the wavelength in *vacuum*!).

- (a) What does “CW Nd:YAG” mean?
- (b) Assume the following parameters for the gain medium:⁷ the lasing transition has a lifetime $\tau_{21} = 1.2$ ms and a linewidth $\Delta\omega/2\pi = 120$ GHz. Assume that you can model the lineshape function as a Lorentzian,

$$s(\omega) = \frac{\Delta\omega}{2\pi [(\omega_0 - \omega)^2 + (\Delta\omega/2)^2]}, \quad (15.113)$$

where ω_0 is the resonance frequency. Use this form to write down an expression for $s(\omega_0)$. What is the Einstein A coefficient for the transition? What is the corresponding resonant laser cross section $\sigma(\omega_0)$? The saturation intensity I_{sat} on resonance? (Check your answers: $I_{\text{sat}} = 262 \text{ W/cm}^2$.)

- (c) Assume a cavity with a perfectly reflecting back mirror and a 95% reflecting output mirror (5% transmission), with all other losses negligible. Also assume a 10 cm long gain medium. What is the threshold population inversion density $N_2 - N_1$ required for the laser to oscillate? (Assume $g_2 = g_1$ for this laser.)

- (d) Assuming an ideal 4-level laser, we can use the rate equation

$$\frac{dN_2}{dt} = R_{\text{pump}}N_0 - A_{21}N_2 \quad (15.114)$$

to model the level population densities, with $N_3 \approx 0$, $N_1 \approx 0$, and $N \approx N_2 + N_0$. (Right now we are still excluding the effect of laser light.) Find an expression for the steady-state inversion $(N_2 - N_1)/N$.

- (e) At what rate R_{pump} must you pump the medium to achieve an inversion density of 5 times the threshold value you found in (c)? Assume a dopant density $N = 10^{19} \text{ cm}^{-3}$. (Check your answers: $R_{\text{pump}} = 1.8 \text{ s}^{-1}$.)

- (f) Now including the laser light, the rate equation is

$$\frac{dN_2}{dt} = R_{\text{pump}}N_0 - A_{21}N_2 - \frac{2\sigma(\omega_0)I}{\hbar\omega_0}N_2, \quad (15.115)$$

where I is the laser intensity, and the extra factor of 2 in the last term accounts for the fact that the laser light counterpropagates through the medium. We will assume the cavity intensity equation

$$\frac{dI}{dt} = \frac{(R_1R_2 - 1)}{\tau_{\text{rt}}}I + \frac{2\ell_g\sigma(\omega_0)N_2}{\tau_{\text{rt}}}I + \frac{\eta_{\text{SE}}\ell_g\hbar\omega_0A_{21}N_2}{\tau_{\text{rt}}}, \quad (15.116)$$

⁷Bahaa E. A. Saleh and Malvin Carl Teich, *Fundamentals of Photonics* (Wiley, 1991) p. 478

where the second term now accounts for the gain (saturation is built into the rate equation for N_2), and the third term is a small contribution to the intensity due to spontaneous emission. Again, the second term here has a factor of 2 to account for the counterpropagating mode in the linear cavity. These equations are valid when the laser intensity changes slowly. Here we have defined η_{SE} as the fraction of spontaneously emitted photons that end up in the cavity mode (with proper direction, frequency, polarization, etc.; note that we are not properly treating effects due to amplified spontaneous emission). Assume a cavity length of 1 m (make sure to account for the laser rod in calculating the round-trip time τ_{rt} !). Solve these simultaneous equations and make plots of the laser intensity and the population inversion N_2/N . Assume also a small initial intensity (due to spontaneous emission) to get things started, and generate solutions over a few photon lifetimes.

For solving ODEs, numerical integrators don't like really big or really small numbers, nor units for that matter. Therefore you should switch to dimensionless variables. Solve this set of differential equations, which is equivalent to the set of unscaled equations above:

$$\begin{aligned}\frac{dn_2}{dt'} &= R'_{\text{pump}}(1 - n_2) - \left(\frac{\tau_p}{\tau_{21}}\right)n_2 - \left(\frac{\tau_p}{\tau_{21}}\right)2I'n_2 \\ \frac{dI'}{dt'} &= -I' + \frac{n_2}{n_{\text{th}}}I' + \eta_{\text{SE}}\frac{n_2}{n_{\text{th}}}.\end{aligned}\quad (15.117)$$

Here, $t' := t/\tau_p$, $I' := I/I_{\text{sat}}$, $n_2 := N_2/N$, $R'_{\text{pump}} := R_{\text{pump}}\tau_p$, $n_{0,2} = N_{0,2}/N$, $\tau_p := \tau_{\text{rt}}/(1 - R_1R_2)$ is the photon lifetime, and $n_{\text{th}} := (1 - R_1R_2)/(2\ell_g\sigma(\omega_0)N)$ is the threshold inversion density. (Use scaled variables in the plots as well.) Use initial conditions $n_2(0) = 0$ and $I(0) = 0$, and assume $\eta_{\text{SE}} = 10^{-12}$. Use the parameters you calculated or were given above to set up the parameters for these scaled equations. How does changing the ratio τ_p/τ_{21} affect the laser dynamics? Perform sanity checks on your results, e.g., by calculating expressions for steady-state values to compare with your data.

Problem 15.8

This is a follow-up to Problem 15.7 that models the CW Nd:YAG laser. Here you will examine numerically the dynamics of a *pulsed* Nd:YAG laser.

Use these assumptions, which are different from the last problem:

1. The laser uses a *ring* cavity, comprising four mirrors at each of the corners of the square beam path. Each leg of the beam path is 50 cm long, and the gain medium sits in one leg of the cavity. Appropriate measures have been taken to ensure the laser oscillates in only one direction.
2. The output coupler has an intensity reflection coefficient of $R_1 = 80\%$. The other three mirrors have reflectances of $R_2 = R_3 = R_4 = 99.5\%$. Any light not reflected is transmitted through the mirror.
3. The laser will be pumped at a rate R_{pump} that produces an inversion density N_2 of 10 times the threshold inversion N_{th} (in the absence of laser oscillation).

Everything else should be the same as last time:

1. The laser is well modeled as a four-level laser producing output at a wavelength of 1.064 μm . Ignore degeneracy of the laser transition levels.
2. The refractive index of YAG is $n = 1.83$ at the laser wavelength, so the laser wavelength inside the gain medium is 580 nm. The gain medium has length $\ell_g = 10$ cm.
3. The lasing transition has a lifetime $\tau_{21} = 1.2$ ms and a line width $\Delta\omega/2\pi = 120$ GHz (with a Lorentzian line-shape function $s(\omega)$). These parameters produce a resonant laser cross section $\sigma(\omega_0) = 5.9 \times 10^{-19}$ cm^2 and a saturation intensity $I_{\text{sat}} = 262$ W/ cm^2 .
4. The Nd dopant density is $N = 10^{19}$ cm^{-3} .

- (a) Recalculate the threshold inversion density N_2 required for this laser to oscillate. You must generalize the formula you used last week slightly; do not use the formula given below for N_{th} . What is the corresponding pump rate that produces an inversion of 10 times the threshold value (if the laser is not oscillating)? (Check your answer: $R_{\text{pump}} = 35 \text{ s}^{-1}$.)
- (b) The output mirror we chose is better for creating short pulses, since it quickly couples energy out the resonator. What reflectance of the output coupler would maximize the output intensity for CW operation? What is the corresponding output intensity?
- (c) We will model the laser with the pair of equations

$$\begin{aligned}\frac{dN_2}{dt} &= R_{\text{pump}}N_0 - A_{21}N_2 - \frac{\sigma(\omega_0)I}{\hbar\omega_0}N_2, \\ \frac{dI}{dt} &= \frac{(P_s - 1)}{\tau}I + \frac{\ell_g\sigma(\omega_0)N_2}{\tau}I + \frac{\eta_{\text{SE}}\ell_g\hbar\omega_0A_{21}N_2}{\tau}.\end{aligned}\quad (15.118)$$

where $P_s := R_1R_2R_3R_4$ is the round-trip survival probability, τ_{rt} is the cavity round-trip time, I is the intensity circulating in the cavity, and η_{SE} is the fraction of spontaneously emitted photons that end up in the cavity mode (with proper direction, frequency, polarization, etc.; note that we are not properly treating effects due to amplified spontaneous emission). These are the same equations as before, but with spontaneous emission into the cavity mode explicitly accounted for. These equations are really only valid for a high-finesse resonator, so they will be marginally accurate but good enough for our purposes.

Using the scaled and otherwise defined variables $t' := t/\tau_p$, $I' := I/I_{\text{sat}}$, $N_{\text{th}} := (1 - P_s)/\ell_g\sigma(\omega_0)$, $n_{\text{th}} := N_{\text{th}}/N$, $n_2 := N_2/N$, $R'_{\text{pump}} := R_{\text{pump}}\tau_p$, $\tau_p := \tau_{\text{rt}}/(1 - P_s)$, $I_{\text{sat}} := \hbar\omega_0/\sigma(\omega_0)\tau_{21}$, and $A_{21} = \tau_{21}^{-1}$, show that the above rate equations are equivalent to the following equations:

$$\begin{aligned}\frac{dn_2}{dt'} &= R'_{\text{pump}}(1 - n_2) - \left(\frac{\tau_p}{\tau_{21}}\right)n_2 - \left(\frac{\tau_p}{\tau_{21}}\right)I'n_2 \\ \frac{dI'}{dt'} &= -I' + \frac{n_2}{n_{\text{th}}}I' + \eta_{\text{SE}}\frac{n_2}{n_{\text{th}}}.\end{aligned}\quad (15.119)$$

(d) Assume that the gain medium is pumped by a flashlamp modeled by a square pulse of duration $\tau_{\text{pulse}} = 1 \text{ ms}$ (i.e., define a pulse function $f_{\text{pulse}}(t')$ which is unity between $t = 0$ and τ_{pulse} , which can be written in terms of the Heaviside step function as $f_{\text{pulse}}(t') = U_H(t') - U_H(t' - \tau_{\text{pulse}}/\tau_p)$, and then replace $R'_{\text{pump}} \rightarrow R'_{\text{pump}}f_{\text{pulse}}(t')$ in the above equations). Plot the solution to the scaled rate equations for these conditions, with the initial conditions $n_2(0) = I'(0) = 0$ (assume $\eta_{\text{SE}} = 10^{-12}$).

(e) Now plot the laser output for a Q -switched pulse. We will model Q -switching the cavity by allowing n_{th} to be effectively time dependent. Let $f_Q(t')$ have the value 12 during the flashlamp pulse (to model a Q -spoiled cavity) and then have the value unity immediately afterwards; then make the replacement $n_{\text{th}} \rightarrow n_{\text{th}}f_Q(t')$ in the above equations (in addition to the modification you made in (d)). Compare your results to what you saw for the “long-pulse mode” of (d). What is the duration of the Q -switched pulse (in ns)?

Problem 15.9

Recall the simple laser model

$$\frac{dI}{dt} = \frac{(GR_1R_2 - 1)}{\tau}I,\quad (15.120)$$

where the gain G saturates according to

$$G = 1 + \frac{g}{1 + \frac{I}{I_{\text{sat}}}}.\quad (15.121)$$

Under the conditions where this equation is valid, show that this model is equivalent to the model in Problem 15.7

$$\frac{dI}{dt} = \frac{(R_1R_2 - 1)}{\tau}I + \frac{2\ell_g\sigma(\nu_0)N_2}{\tau}I,\quad (15.122)$$

which is specific to the four-level laser (note that we have dropped the spontaneous emission term). Do this in the following steps:

- (a) Argue that both G and $R_1 R_2$ must be close to 1. Then use this to argue that the first model can be rewritten (approximately) as

$$\frac{dI}{dt} = \frac{(R_1 R_2 - 1)}{\tau} I + \frac{(G - 1)}{\tau} I. \quad (15.123)$$

- (b) Derive an expression for the *round-trip* gain G in the regime where G is close to 1, and hence recover the equation for the second model above.

- (c) Complete the argument by using the results from the four-level laser to write an expression for g . Be careful to make sure your results are valid for a round trip, not just a single pass through the gain medium.

Problem 15.10

For a ring laser modeled by the gain equations (as in Problem 15.8)

$$\begin{aligned} \frac{dn_2}{dt'} &= R'_{\text{pump}}(1 - n_2) - \left(\frac{\tau_p}{\tau_{21}}\right) n_2 - \left(\frac{\tau_p}{\tau_{21}}\right) I' n_2 \\ \frac{dI'}{dt'} &= -I' + \frac{n_2}{n_{\text{th}}} I' + \eta_{\text{SE}} \frac{n_2}{n_{\text{th}}}, \end{aligned} \quad (15.124)$$

where P_s is the round-trip survival probability, τ_{rt} is the cavity round-trip time, I is the intensity circulating in the cavity, $\tau_p := \tau_{\text{rt}}/(1 - P_s)$, $t' := t/\tau_p$, $I' := I/I_{\text{sat}}$, $N_{\text{th}} := (1 - P_s)/\ell_g \sigma(\omega_0)$, $n_{\text{th}} := N_{\text{th}}/N$, $n_2 := N_2/N$, $R'_{\text{pump}} := R_{\text{pump}} \tau_p$, $I_{\text{sat}} := \hbar \omega_0 / (\sigma(\omega_0) \tau_{21})$, and $A_{21} = \tau_{21}^{-1}$, you will compute the frequency and damping rates of the *relaxation oscillations* for small disturbances about the equilibrium (these are the same oscillations as in the laser spikes when they have nearly damped away). Show that the intensity oscillation frequency and damping rate are given by ω_γ and $\gamma/2$, respectively, where

$$\begin{aligned} \omega_\gamma &= \sqrt{\omega^2 - (\gamma/2)^2} \\ \omega &= \sqrt{\frac{r - 1}{\tau_p \tau_{21}}} \\ \gamma &= \frac{r}{\tau_{21}(1 - n_{\text{th}})} \\ r &:= \frac{R'_{\text{pump}}}{R'_{\text{pump,th}}}, \end{aligned} \quad (15.125)$$

and $R'_{\text{pump,th}}$ is the scaled threshold pumping rate (the value of R'_{pump} above which the laser oscillates). You can do this via the following outline.

- (a) Neglect spontaneous emission and show that in steady state, $n_2 = n_{\text{th}}$ and $I'_{\text{ss}} = r - 1$. What is $R'_{\text{pump,th}}$?
- (b) Ignore spontaneous emission and linearize the above pair of rate equations about the equilibrium solution by letting $n_2 = n_{\text{th}} + \delta n_2$ and $I' = I'_{\text{ss}} + \delta I'$, putting these into the rate equations, and discarding terms beyond first order in δn_2 and $\delta I'$. You should now have a pair of rate equations for the small disturbances δn_2 and $\delta I'$.
- (c) Recall that the damped harmonic oscillator is governed by the equation

$$m\ddot{x} + m\gamma\dot{x} + m\omega^2 x = 0. \quad (15.126)$$

Recall that in the underdamped regime, the solution with $p(0) = 0$ is

$$x(t) = x(0)e^{-(\gamma/2)t} \cos(\omega_\gamma t), \quad (15.127)$$

where ω_γ is defined above.

Show that the harmonic oscillator equation can be rewritten as the pair of first-order equations

$$\begin{aligned}\dot{x} &= \frac{p}{m} \\ \dot{p} &= -m\omega^2 x - \gamma p,\end{aligned}\tag{15.128}$$

where $p = m\dot{x}$. Now draw a formal analogy between the laser gain equations and the damped harmonic oscillator, and write down expressions for the oscillation frequency and damping rate. Remember that the above rate equations measure time in units of τ_p , so you must undo this to get the above answers.

(d) To get a sense of scale, put in some numbers to see a typical oscillation frequency. Assume a low-loss ring cavity ($n_{\text{th}} \ll 1$, $\tau_p \sim 10\tau_{\text{rt}}$) just above threshold ($r = 1.1$) with a ~ 1 m cavity path length ($\tau_{\text{rt}} \approx 3$ ns), and a Nd:YAG gain medium ($\tau_{21} = 1.2$ ms). Roughly how many oscillations do you expect to see before they damp away?

Problem 15.11

(This is a simple review question to help you tie together some of the concepts.) Consider a laser whose gain medium is described by a frequency-dependent gain coefficient $\gamma(\omega)$. Your answers to the questions below should be *qualitative* and need not involve calculations or equations.

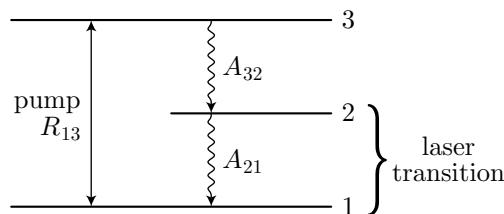
- (a) What is the condition for the laser to oscillate?
- (b) Suppose that many cavity modes satisfy the condition in (a), but in steady state, the laser operates only in a single mode. What is the underlying effect that causes single-mode operation? Explain how this effect leads to single-mode operation.
- (c) A different gain medium in the same cavity causes the laser to oscillate in many modes simultaneously. Explain how the medium is different from the medium in part (b).
- (d) What is mode-locking? Give an example of how to achieve mode locking in a multimode laser.

Problem 15.12

Due to temperature fluctuations, the length and therefore the frequency of a helium-neon (He-Ne) laser drifts slightly. Suppose you observe that a particular He-Ne laser oscillates in only one (longitudinal) mode at some temperatures, but two modes at others. Estimate the length of the laser cavity, assuming an operating wavelength $\lambda = 632.8$ nm, a gain linewidth $\Delta\nu = 1.5$ GHz, and $\text{TEM}_{0,0}$ transverse-mode operation.

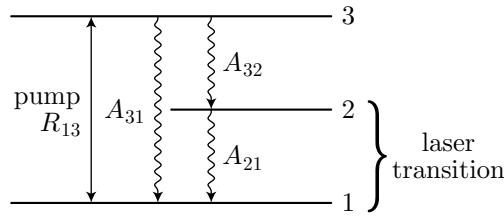
Problem 15.13

Consider the three-level laser scheme shown here. (The first three parts of this problem are a review of the chapter material.)



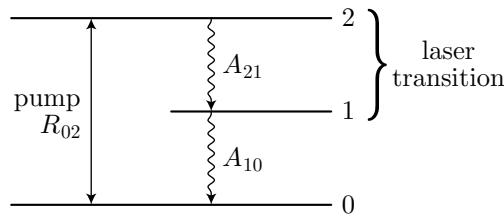
- (a) Write down the rate equations for this laser.
- (b) What are the conditions on R_{13} , A_{32} , and A_{21} that ensure the best laser operation?
- (c) In the regime you described in part (b), carry out the adiabatic elimination of the appropriate state to obtain a set of rate equations for an effective two-level atom.
- (d) Suppose we add an extra decay path as shown below, described by the (possibly large) coefficient A_{31} . Perform a similar analysis (with the same adiabatic elimination), and write down an expression

for the small-signal gain coefficient γ_0 in terms of the coefficient for the gain medium in parts (a)-(c), assuming everything else is the same. That is, what is the effect of A_{31} on the gain coefficient?



Problem 15.14

Consider the *inverted* three-level laser scheme shown here.



- Write down the rate equations for this laser.
- What are the conditions on R_{02} , A_{21} , and A_{10} that ensure the best laser operation?
- In the regime you described in part (b), carry out the adiabatic elimination of the appropriate state to obtain a set of rate equations for an effective two-level atom.
- The *usual* three-level laser that we covered in Section 15.4.1 has the lower two levels defining the laser transition. Argue *qualitatively* that in the best possible cases (optimal pumping and decay rates), for the same pumping rate, laser cross section, and atomic number density, the small-signal gain coefficient for the usual scheme is *twice* the small-signal gain coefficient for the inverted scheme shown here. Recall that the gain coefficient is given by $\gamma(\omega) = \sigma(\omega)[N_e - N_g]$, where “e” and “g” refer to the excited and ground levels of the laser transition, respectively. *Hint:* you don’t need to do any calculations, just reason out what the steady-state populations would be in the optimal cases.
- Give a *qualitative* argument for why the saturation intensity for the inverted three-level scheme will be *twice* that of the saturation intensity for the usual three-level scheme. Assume the laser cross sections are the same in both cases.

Chapter 16

Dispersion and Wave Propagation

Now that we have discussed the generation of laser light and the propagation of laser light through a gain medium, we will take a deeper look at propagation in media. In particular, we will look at the effects of **dispersion**—dependence of dielectric material properties on the optical frequency—on light propagation.

16.1 Causality and the Kramers–Kronig Relations

Recall from Section 4.3 that the polarization density of a dielectric is the material response to the electric field,

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E}, \quad (16.1)$$

where χ is the susceptibility. For dispersive media, the susceptibility is frequency-dependent ($\chi = \chi(\omega)$),

$$\tilde{\mathbf{P}}(\omega) = \epsilon_0 \chi(\omega) \tilde{\mathbf{E}}(\omega), \quad (16.2)$$

where $\tilde{\mathbf{E}}(\omega)$ and $\tilde{\mathbf{P}}(\omega)$ is the magnitude of the component $\exp(-i\omega t)$. We can also think of $\chi(\omega)$ as the (complex-valued) *frequency-space transfer function* for the medium. It turns out that just by knowing that $\chi(\omega)$ is a response function corresponding to a physical system, and thus describing a *causal* response, we can put substantial constraints on its form.

By the convolution theorem, the frequency-space multiplication in Eq. (16.2) is equivalent to a convolution in the time domain of the electric field with the inverse fourier transform $g_\chi(t)$ of $\chi(\omega)$:

$$\mathbf{P}(t) = \epsilon_0 \int_{-\infty}^{\infty} g_\chi(t') \mathbf{E}(t - t') dt'. \quad (16.3)$$

Since $\mathbf{P}(t)$ is the material response to $\mathbf{E}(t)$, causality requires that $\mathbf{P}(t)$ can't depend on $\mathbf{E}(t')$ for future times $t' > t$. Thus, from the convolution, we can see that causality implies $g_\chi(t) = 0$ for $t < 0$. Equivalently, we can write causality as the mathematical statement

$$g_\chi(t) = g_\chi(t) U_H(t), \quad (16.4)$$

where $U_H(t)$ is the Heaviside step function. Recalling from Section 12.5.2.3 that

$$\mathcal{F}[U_H(t)] \equiv \tilde{U}_H(\omega) = \pi \delta(\omega) + \frac{i}{\omega}, \quad (16.5)$$

we can compute the Fourier transform of Eq. (16.4), using the convolution theorem again, but this time in

the opposite direction than usual:

$$\begin{aligned}
\chi(\omega) &= \mathcal{F}[g_\chi(t)] = \mathcal{F}[g_\chi(t)U_H(t)] \\
&= \mathcal{F}[\mathcal{F}^{-1}[\chi(\omega)]\mathcal{F}^{-1}[\tilde{U}_H(\omega)]] \\
&= \frac{1}{2\pi}\chi(\omega) * \tilde{U}_H(\omega) \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \chi(\omega') \left[\pi\delta(\omega - \omega') + \frac{i}{\omega - \omega'} \right] d\omega' \\
&= \frac{\chi(\omega)}{2} + \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{\chi(\omega')}{\omega - \omega'} d\omega' \\
&= \frac{\chi(\omega)}{2} - \frac{i}{2}\mathcal{H}[\chi(\omega)]. \tag{16.6}
\end{aligned}$$

Here again, \mathcal{H} denotes the Hilbert transform, defined by

$$\mathcal{H}[f(x)] := \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(x')}{x' - x} dx'. \tag{16.7}$$

The factor of $1/2\pi$ in applying the convolution theorem in the above derivation comes again from applying the convolution theorem in the reverse sense: the usual convolution theorem reads $\mathcal{F}[f(t) * g(t)] = \mathcal{F}[f(t)]\mathcal{F}[g(t)]$, but in the opposite direction we have $\mathcal{F}^{-1}[\tilde{f}(\omega) * \tilde{g}(\omega)] = 2\pi\mathcal{F}^{-1}[\tilde{f}(\omega)]\mathcal{F}^{-1}[\tilde{g}(\omega)]$, due to the factor of 2π that appears in the inverse Fourier transform equation (3.21).

Thus, from Eq. (16.6) we have that

$$\chi(\omega) = -i\mathcal{H}[\chi(\omega)]. \tag{16.8}$$

Writing out the real and imaginary parts of this equation,

$$\begin{aligned}
\text{Re}[\chi(\omega)] &= \mathcal{H}\{\text{Im}[\chi(\omega)]\} \\
\text{Im}[\chi(\omega)] &= -\mathcal{H}\{\text{Re}[\chi(\omega)]\}. \tag{16.9}
\end{aligned}$$

Writing out the Hilbert transform explicitly, we arrive at the standard form of the **Kramers–Kronig relations**:

$$\begin{aligned}
\text{Re}[\chi(\omega)] &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Im}[\chi(\omega')]}{\omega' - \omega} d\omega' \\
\text{Im}[\chi(\omega)] &= -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Re}[\chi(\omega')]}{\omega' - \omega} d\omega'. \tag{16.10}
\end{aligned}$$

Notice that this is a strong constraint on the form of $\chi(\omega)$: the real part of χ determines the imaginary part via the Hilbert transform and vice versa.

The Kramers–Kronig relations are often written in terms of the complex permittivity (dielectric constant) $\tilde{\epsilon}(\omega) = \epsilon_0[1 + \chi(\omega)]$, where we recall from Section 4.3 that the electric displacement is given in terms of the permittivity by $\tilde{\mathbf{D}}(\omega) = \tilde{\epsilon}(\omega)\tilde{\mathbf{E}}(\omega)$. The Kramers–Kronig relations for the permittivity then read

$$\begin{aligned}
\text{Re}[\tilde{\epsilon}(\omega)] &= \epsilon_0 + \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Im}[\tilde{\epsilon}(\omega')]}{\omega' - \omega} d\omega' \\
\text{Im}[\tilde{\epsilon}(\omega)] &= -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Re}[\tilde{\epsilon}(\omega')] - \epsilon_0}{\omega' - \omega} d\omega'. \tag{16.11}
\end{aligned}$$

Again, the real part of $\tilde{\epsilon}$ determines the imaginary part via the Hilbert transform and vice versa.

16.1.0.1 DC Component

But one should then ask, why do the Kramers–Kronig relations for the susceptibility [Eqs. (16.10)] and the permittivity [Eqs. (16.11)] have different forms, where ϵ_0 needs to be subtracted from $\text{Re}[\tilde{\epsilon}(\omega)]$? After all, from $\tilde{\mathbf{D}}(\omega) = \tilde{\epsilon}(\omega)\tilde{\mathbf{E}}(\omega)$, $\tilde{\epsilon}$ looks like an input-output response function just like χ . The difference lies in the dc component of $\epsilon(\omega)$, which is absent in χ . This is because at very large frequencies, the material electrons fail to respond to the electric field, and thus

$$\lim_{\omega \rightarrow \pm\infty} \chi(\omega) = 0; \quad \lim_{\omega \rightarrow \pm\infty} \tilde{\epsilon}(\omega) = \epsilon_0. \quad (16.12)$$

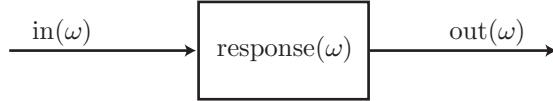
If we repeat the derivation of Eqs. (16.6), we find that if χ had an asymptotic dc component of $\chi_\infty := \chi(\omega \rightarrow \infty)$, the response function $g_\chi(t)$ picks up the additional term $\chi_\infty \delta(t)$. It works out if $U_H(t)$ passes “half” of $\delta(t)$, so that

$$\chi(\omega) - \chi_\infty = -i\mathcal{H}[\chi(\omega) - \chi_\infty]. \quad (16.13)$$

Thus, *any causal response function with its asymptotic value subtracted away* satisfies the Kramers–Kronig relations. Thus, Eqs. (16.11) follow from the fact that $[\tilde{\epsilon}(\omega) - \epsilon_0]$ is the appropriate response function for the Kramers–Kronig relations.

16.1.1 Refractive Index

Notice that the Kramers–Kronig relations are not specific to electromagnetism. Rather, they apply to any frequency-space transfer function that corresponds to a physical, causal response.



For example, they could apply to the response function of an electronic amplifier or feedback loop. For our purposes, it most relevant to note that the Kramers–Kronig relations also apply to the refractive index. Recall the propagation of a plane wave in a medium with a complex refractive index \tilde{n} , which carries over into the dispersive case:

$$E^{(+)}(z) = E_0^{(+)} e^{i\tilde{n}k_0 z} \longrightarrow \tilde{E}^{(+)}(\omega; z) = \tilde{E}_0^{(+)}(\omega) e^{i\tilde{n}(\omega)k_0 z}. \quad (16.14)$$

The corresponding infinitesimal evolution is

$$\tilde{E}^{(+)}(\omega; z + dz) = \tilde{E}^{(+)}(\omega; z)[1 + i\tilde{n}(\omega)k_0 dz], \quad (16.15)$$

which is equivalent to the differential equation

$$\frac{d\tilde{E}^{(+)}(\omega; z)}{dz} = ik_0\tilde{n}(\omega)\tilde{E}^{(+)}(\omega; z). \quad (16.16)$$

This is another input-output relation, with input $\tilde{E}^{(+)}$, output $d\tilde{E}^{(+)}/dz$, and transfer function proportional to $\tilde{n}(\omega)$. Thus, the refractive index $(\tilde{n}(\omega) - 1)$, after subtracting the asymptotic dc part as in the discussion of Section 16.1.0.1, since \tilde{n} tends to unity at high frequencies) obeys the Kramers–Kronig relations

$$\tilde{n}(\omega) - 1 = -i\mathcal{H}[\tilde{n}(\omega) - 1], \quad (16.17)$$

or in inefficient form,

$$\begin{aligned} \text{Re}[\tilde{n}(\omega)] &= 1 + \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Im}[\tilde{n}(\omega')]}{\omega' - \omega} d\omega' \\ \text{Im}[\tilde{n}(\omega)] &= -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Re}[\tilde{n}(\omega')] - 1}{\omega' - \omega} d\omega'. \end{aligned} \quad (16.18)$$

Recall that the *complex* refractive index corresponds to both phase delay *and* absorption (or gain). This is evident from Eq. (16.16), where we can write the solution as

$$\tilde{E}^{(+)}(\omega; z) = \tilde{E}_0^{(+)}(\omega) e^{i\text{Re}[\tilde{n}(\omega)]k_0 z} e^{-\text{Im}[\tilde{n}(\omega)]k_0 z}. \quad (16.19)$$

The first exponential factor is a propagating plane wave with wave number $\text{Re}[\tilde{n}(\omega)]k_0$. The second exponential factor represents exponential decay (or gain if $\text{Im}[\tilde{n}(\omega)] < 0$).

Since $\text{Re}[\tilde{n}(\omega)]$ controls the phase shift of propagating light and $\text{Im}[\tilde{n}(\omega)]$ controls the absorption (gain), we see why the Kramers–Kronig relations are so significant. Since the Hilbert transformation is “something like” a derivative (see Section 12.5.2.3), derivatives in the real part of $\tilde{n}(\omega)$ induce large imaginary parts, and vice versa. Thus, for example, variation with frequency of the (real) refractive index implies regions of absorption. Similarly, variation in the absorption coefficient implies regions of strong (real) refractive index.

16.1.1.1 Example: Lorentzian Absorption

One particularly important absorption line shape is the Lorentzian [see Eq. (15.37)], which we can write as

$$\text{Im}[\tilde{n}(\omega)] = \frac{a(\omega)}{2k_0} = \frac{(a_0/2k_0)(\Delta\omega/2)^2}{(\omega_0 - \omega)^2 + (\Delta\omega/2)^2}. \quad (16.20)$$

The convention here is that $a(\omega)$ is the *intensity* absorption coefficient, so that $I(z) = I_0 \exp[-a(\omega)z]$, and $a_0 = a(\omega_0)$. Note that $a_0 > 0$ corresponds to absorption, while $a_0 < 0$ corresponds to gain.

Now we need to evaluate the integral

$$\text{Re}[\tilde{n}(\omega)] - 1 = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Im}[\tilde{n}(\omega')]}{\omega' - \omega} d\omega' \quad (16.21)$$

to find the phase behavior. Let’s first do this with a simple, approximate method. Assume that $a(\omega)$ is sharply peaked around ω_0 ($\Delta\omega \ll \omega_0$). Then we can approximate the Lorentzian by a δ -function, by considering the normalized form of the line shape:

$$\frac{\Delta\omega}{2\pi[(\omega_0 - \omega)^2 + (\Delta\omega/2)^2]} \approx \delta(\omega - \omega_0). \quad (16.22)$$

In this case,

$$\text{Im}[\tilde{n}(\omega)] \approx \frac{a_0}{2k_0} \frac{\pi\Delta\omega}{2} \delta(\omega - \omega_0), \quad (16.23)$$

so that

$$\text{Re}[\tilde{n}(\omega)] - 1 \approx \frac{a_0}{4k_0} \Delta\omega \int_{-\infty}^{\infty} \frac{\delta(\omega' - \omega_0)}{\omega' - \omega} d\omega' = \frac{a_0}{4k_0} \frac{\Delta\omega}{(\omega_0 - \omega)}. \quad (16.24)$$

Notice that the refractive index scales as $(\omega_0 - \omega)$ for large $|\omega_0 - \omega|$, while the absorption coefficient scales as $(\omega_0 - \omega)^2$. This turns out to be an important result in atom optics.

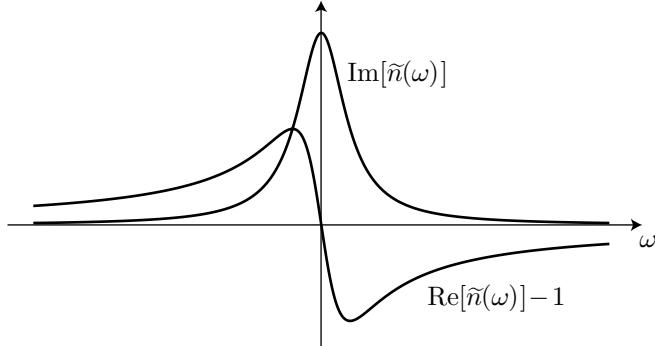
It isn’t too difficult to compute the Hilbert transform without approximations. Let’s compute the Hilbert transform of the scaled Lorentzian $1/(1+x^2)$. We can do this using the convolution theorem, the fact that the Fourier transform of $1/(1+x^2)$ is $\pi \exp(-|k|)$ (see Problem 11.7), and the fact that the Fourier transform of $1/x$ is $i \text{sgn}(k)$ (see Problem 12.16):

$$\begin{aligned} \mathcal{H}\left[\frac{1}{1+x^2}\right] &= -\mathcal{F}^{-1}[i\pi \text{sgn}(k) \exp(-|k|)] \\ &= \frac{-i}{2\pi} \int_0^\infty \pi \exp(-k) e^{ikx} dk + \text{c.c.} \\ &= \frac{-i}{2} \int_0^\infty e^{-(1-ix)k} dk + \text{c.c.} \\ &= \frac{i}{2(1-ix)} + \text{c.c.} \\ &= -\frac{x}{1+x^2}. \end{aligned} \quad (16.25)$$

Then with $x = 2(\omega - \omega_0)/\Delta\omega$, we can write

$$\text{Re}[\tilde{n}(\omega)] = 1 + \frac{2(\omega_0 - \omega)}{\Delta\omega} \text{Im}[\tilde{n}(\omega)]. \quad (16.26)$$

We will derive this formula again when we examine models of atomic media.



16.2 Pulse Propagation and Group Velocity

16.2.1 Phase Velocity

Now we will consider the velocity of light propagation in a dispersive medium. We will start out by considering the monochromatic case, and then we will generalize to the case of pulses with a spectrum of finite width.

Recall the plane-wave solution to the electromagnetic wave equation:

$$E^{(+)}(x, t) = E_0^{(+)} e^{i(kx - \omega t)}. \quad (16.27)$$

The phase of the wave is thus $\phi = kx - \omega t$. We can define the *phase velocity* as the velocity of the points of constant phase. To derive an expression for the phase velocity, consider the infinitesimal transformation $x \rightarrow x + dx$ and $t \rightarrow t + dt$, while ϕ is held constant. Then the phase is the same before and after the transformation,

$$\phi = k(x + dx) - \omega(t + dt) = kx - \omega t, \quad (16.28)$$

so that $k dx = \omega dt$. This defines the phase velocity v_{ph} as

$$v_{\text{ph}} := \frac{dx}{dt} = \frac{\omega}{k}. \quad (16.29)$$

Recall that for a plane wave to be a solution of the wave equation,

$$v_{\text{ph}} = \frac{\omega}{k} = c = \frac{c_0}{n}, \quad (16.30)$$

where n is the refractive index, and $c_0 = 1/\sqrt{\epsilon_0 \mu_0}$ is what we defined to be the speed of light in vacuum in Section 4.1. In general, the refractive index depends on the optical wavelength and thus on the wave number, $n = n(k)$, for a dispersive dielectric medium. Thus, the phase velocity $v_{\text{ph}} = v_{\text{ph}}(k)$ is likewise wavelength-dependent. By the Kramers–Kronig relations, we know that the dispersive refractive index must also be complex, so when we write n here, we are referring only to the real part of \tilde{n} .

16.2.2 Group Velocity

But what happens with pulses, which contain a continuum of frequency components? In a dispersive medium, each component sees a different phase velocity:

$$E^{(+)}(x, t) = \frac{1}{2\pi} \int_0^\infty \tilde{E}^{(+)}(k) e^{i(kx - \omega t)} dk. \quad (16.31)$$

Here we account for dispersion by regarding $\omega = \omega(k)$ and use k as the independent variable, which we can do via $\omega(k) = k v_{\text{ph}}(k) = c_0 k / \tilde{n}(k)$. This works for small deviations of the refractive index from unity, such that $k(\omega)$ is an invertible function, but the ideas here are generally applicable.

For $\omega = \omega_0$, where ω_0 is a constant, all the components have the same phase velocity, so all the time dependence in Eq. (16.31) is of the form

$$\left(x - \frac{\omega_0}{k} t \right) = (x - v_{\text{ph}} t). \quad (16.32)$$

Thus, $E^{(+)}(x, t) = E^{(+)}(x - v_{\text{ph}} t)$, so the pulse propagates at the phase velocity $v_{\text{ph}} = \omega_0/k$.

Now suppose we have a dispersive medium, but let's assume that the dispersion relation is linear. More generally, we can expand an arbitrary dispersion relation to first order about a center wave number k_c . In either case,

$$\omega(k) = \omega_0 + \frac{d\omega}{dk} \Big|_{k_c} (k - k_c) = \omega_0 + v_g(k - k_c), \quad (16.33)$$

where $\omega_0 = \omega(k_c)$, and we have defined the **group velocity**

$$v_g := \frac{d\omega}{dk} \Big|_{k_c}, \quad (16.34)$$

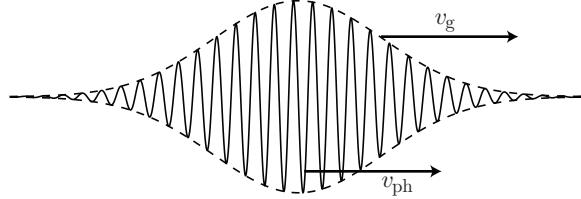
whose significance will become apparent presently. Now we can write the pulse function (16.31) as

$$\begin{aligned} E^{(+)}(x, t) &= \frac{1}{2\pi} \int_0^\infty \tilde{E}^{(+)}(k) e^{i[kx - \omega(k)t]} dk \\ &= \frac{1}{2\pi} \int_0^\infty \tilde{E}^{(+)}(k) e^{i(kx - \omega_0 t)} e^{-iv_g(k - k_c)t} dk \\ &= \frac{1}{2\pi} e^{ik_0 v_g t} e^{-i\omega_0 t} \int_0^\infty \tilde{E}^{(+)}(k) e^{ik(x - v_g t)} dk \end{aligned} \quad (16.35)$$

Now the phases outside the integral are uninteresting overall phase factors, and all the spacetime dependence is of the form

$$E^{(+)}(x, t) = E^{(+)}(x - v_g t). \quad (16.36)$$

Thus, the pulse propagates at the group velocity $v_g = d\omega/dk$. The expansion of Eq. (16.33) is valid if the pulse spectrum is narrow enough for the higher-order terms to be neglected. Note that there are multiple relevant propagation velocities here: while the phase fronts of the individual waves still propagate at the appropriate phase velocities, the *overall envelope* of the pulse propagates at the group velocity.



Recall that the usual refractive index, which controls the phase velocity, can be written

$$n = \frac{c_0}{v_{\text{ph}}}, \quad (16.37)$$

so we can define an analogous **group index** for the group velocity by

$$n_g := \frac{c_0}{v_g} = c_0 \frac{dk}{d\omega} = n + \omega \frac{dn}{d\omega}, \quad (16.38)$$

since $k = n\omega/c_0$. Also, the vacuum wavelength is

$$\lambda_0 = \frac{2\pi}{k_0} = \frac{2\pi c_0}{\omega}, \quad (16.39)$$

where $k_0 = k/n$ is the vacuum wave number. Then

$$\frac{d\lambda_0}{d\omega} = -\frac{2\pi c_0}{\omega^2} = -\frac{\lambda_0}{\omega}, \quad (16.40)$$

and so we can write the group index (16.38) as

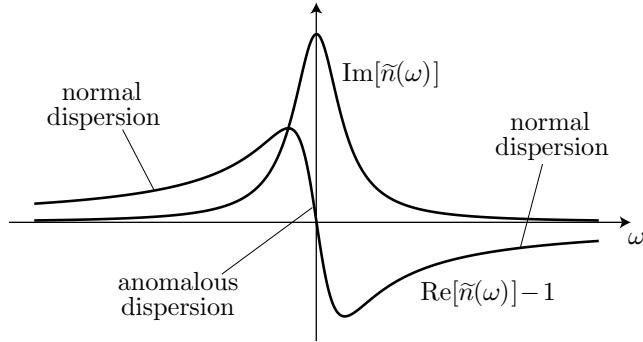
$$n_g = n + \omega \frac{dn}{d\omega} = n + \omega \frac{dn}{d\lambda_0} \frac{d\lambda_0}{d\omega} = n - \lambda_0 \frac{dn}{d\lambda_0} \approx n - \lambda \frac{dn}{d\lambda}, \quad (16.41)$$

where the last equality follows from assuming that $n(\lambda)$ is a slowly varying function. This form for the group index is convenient, since the normal refractive index n is commonly specified in terms of wavelength.

There are, broadly speaking three cases to consider with respect to the group index.

1. **No dispersion.** ($dn/d\lambda_0 = 0$) This is the nondispersive case that we already discussed, when the group and refractive indices are equivalent: $n_g = n$.
2. **Normal dispersion.** ($dn/d\lambda_0 < 0$) Here the pulse envelope travels **more slowly** than the phase fronts; this happens when the index *decreases* with wavelength, and implies that $n_g > n$. This case is called “normal” because this is the observed behavior when white light is dispersed by a glass prism: blue wavelengths see a larger refractive index than red wavelengths, and are thus deflected more by the prism.
3. **Anomalous dispersion.** ($dn/d\lambda_0 > 0$) Here the pulse envelope travels **more quickly** than the phase fronts; this happens when the index *increases* with wavelength, and implies that $n_g < n$.

For the Lorentzian line-shape function (Section 16.1.1.1), normal dispersion occurs everywhere except near the line center, where the slope of the index curve changes sign. Here, there is a region of width $\Delta\omega$ (usually very narrow) of anomalous dispersion.



In the anomalous case, it is even possible to have $v_g > c_0$. This “superluminal” propagation seems to violate (relativistic) causality, since it seems possible to send a signal pulse faster than the speed of light. However, a more careful analysis of the pulse shape shows that the signal can't be sent at a higher velocity than c_0 .

16.2.3 Pulse Spreading

Higher order terms in the dispersion relation lead to pulse spreading (second order) or distortion (third and higher orders). For example, up to second order,

$$\omega(k) = \omega_0 + v_g(k - k_c) + \omega_2(k - k_c)^2, \quad (16.42)$$

where

$$\omega_2 := \frac{1}{2} \left. \frac{d^2\omega}{dk^2} \right|_{k_c}. \quad (16.43)$$

Then the pulse function (16.31) becomes

$$E^{(+)}(x, t) = \frac{1}{2\pi} e^{ik_0 v_g t} e^{-i\omega_0 t} \int_0^\infty \tilde{E}(k)^{(+)} e^{ik(x - v_g t)} e^{i\omega_2(k - k_c)^2 t} dk \quad (16.44)$$

For a Gaussian pulse,

$$E^{(+)}(x, t=0) = E_0^{(+)} e^{-x^2/4(\Delta x)^2} e^{ik_0 x}, \quad (16.45)$$

so that the Fourier transform is approximately

$$\tilde{E}(k) \approx E_0^{(+)} 2\sqrt{\pi} (\Delta x) e^{-(\Delta x)^2(k-k_c)^2}. \quad (16.46)$$

Thus, the propagation to time t leads to a larger effective width

$$(\Delta x)^2 \longrightarrow (\Delta x)^2 + i\omega_2 t, \quad (16.47)$$

which disperses the pulse in the same way as the complex q parameter of the Gaussian beam of Section 6.5 (p. 97). To see this, we can define a q parameter for this problem by

$$q := \tau - iq_0, \quad (16.48)$$

where $q_0 = 2(\Delta x)^2$ and $\tau = 2\omega_2 t$, so that the propagation replacement becomes

$$(\Delta x)^2 \longrightarrow (\Delta x)^2 + i\omega_2 t = \frac{iq}{2}. \quad (16.49)$$

Now we can take the inverse Fourier transform, thus evaluating the integral of Eq. (16.44). The Fourier transform after propagation through time t is

$$\tilde{E}^{(+)}(k, t) = E_0^{(+)} 2\sqrt{\pi} (\Delta x) e^{iq(k-k_c)^2} e^{-ikv_g t}, \quad (16.50)$$

and so up to an overall phase, the inverse transform is

$$E^{(+)}(x, t) = E_0^{(+)} \frac{\Delta x}{q} \exp\left[\frac{ik(x-v_g t)^2}{2q}\right] e^{ik_c x}. \quad (16.51)$$

This has the same form as the Gaussian beam in terms of the q parameter [Eq. (6.41)],

$$E^{(+)}(r) = E_0^{(+)} \frac{q_0}{q(z)} \exp\left[\frac{ikr^2}{2q(z)}\right] \exp(ikz), \quad (16.52)$$

where $q = z - iz_0 = z - i\pi w_0^2/\lambda$. Since the Gaussian beam satisfies the expansion law [Eq. (6.13)],

$$w(z) = w_0 \sqrt{1 + \left(\frac{z}{z_0}\right)^2}, \quad (16.53)$$

the analogue in pulse propagation in a dispersive medium (by comparing the two forms of the q parameter) is that the *pulse width* Δx increases with a similar functional form with the time parameter τ .

The other aspect of the complex q parameter is that it controls the radius of curvature of the Gaussian beam [Eq. (6.14)],

$$R(z) = z \left[1 + \left(\frac{z_0}{z}\right)^2\right]. \quad (16.54)$$

The Gaussian beam acquires curvature because it is a superposition of plane waves with slightly different directions in their wave vector \mathbf{k} , and as the beam propagates, these begin to separate. For propagation in the $+z$ -direction, the plane waves with wave vectors with positive x -components will move towards one edge of the beam profile, while the ones with negative x -components move towards the other. The analogue in the pulse propagation is that the Gaussian pulse is a superposition of plane waves of different wavelength, and the red/blue wavelength components begin to move to the front/back of the pulse (or the other way around, depending on the sign of ω_2). Here, ω_2 is related to the **group velocity dispersion**, given by $dv_g/d\omega$, which controls this dispersion effect.

16.3 Slow and Fast Light

A recent “hot topic” in the quantum optics community has been the slow and fast propagation of light in atomic media.¹ The general question is, how is it possible to make v_g very small or very large? Answering this question is fundamentally interesting, and can also be technologically interesting: for example, a “slow” medium can act as a compact delay line or even memory storage for optical information.

Let’s consider a medium with Lorentzian absorption (appropriate for stationary gas-phase atoms). From Section 16.1.1.1, the absorptive part of the refractive index is thus given by

$$\text{Im}[\tilde{n}(\omega)] = \frac{a(\omega)}{2k_0} = \frac{(a_0/2k_0)(\Delta\omega/2)^2}{(\omega_0 - \omega)^2 + (\Delta\omega/2)^2}, \quad (16.55)$$

while the phase index is

$$n = \text{Re}[\tilde{n}(\omega)] = 1 + \frac{2(\omega_0 - \omega)}{\Delta\omega} \text{Im}[\tilde{n}(\omega)]. \quad (16.56)$$

The maximum value of $|n - 1|$ occurs for $(\omega - \omega_0) = \pm\Delta\omega/2$, giving a maximum deviation of n from unity of

$$\delta n_{\max} = \frac{a_0}{4k_0}. \quad (16.57)$$

To evaluate the group index [Eq. (16.38)], we need to consider the derivative $dn/d\omega$. The maximum value of the derivative occurs at $(\omega - \omega_0) = \pm\sqrt{3}\Delta\omega/2$, giving

$$\left. \frac{dn}{d\omega} \right|_{\max} = \frac{a_0}{8k_0\Delta\omega} = \frac{\delta n_{\max}}{2\Delta\omega}. \quad (16.58)$$

Thus, we can write the maximum group index as

$$n_{g,\max} = n + \omega \left. \frac{dn}{d\omega} \right|_{\max} = 1 + \left[1 + \frac{\omega}{2\Delta\omega} \right] \delta n_{\max}. \quad (16.59)$$

Let’s put in some typical numbers to see the magnitude of the effect. For alkali vapors, the refractive index is not too large ($n \sim 1.1$), the resonant frequency is in the optical ($\omega_0/2\pi \sim 400$ THz), but the line width is quite narrow ($\Delta\omega/2\pi \sim 6$ MHz). Putting in these numbers gives

$$n_{g,\max} \sim 3 \times 10^6, \quad (16.60)$$

an absolutely enormous index (so that pulses should propagate at ~ 100 m/s. Again, this is valid for *stationary* atoms. For a vapor at room temperature, the line width is more like $\Delta\omega/2\pi \sim 1$ GHz due to Doppler broadening (in fact the line-shape function should be Gaussian rather than Lorentzian, but we’ll ignore this here). The corresponding group index is smaller, but still quite large:

$$n_{g,\max} \sim 2 \times 10^4. \quad (16.61)$$

However, it is practically impossible to observe these effects in ordinary vapors, because a large group index is accompanied by strong absorption:

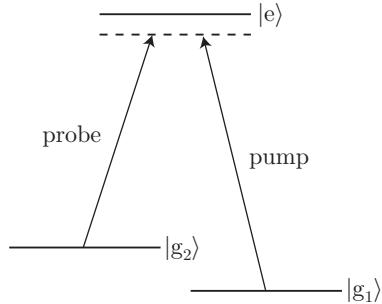
$$a \sim (\delta n)4k_0 = \frac{4\omega(\delta n)}{c_0} \sim 3 \times 10^4 \text{ cm}^{-1}. \quad (16.62)$$

Thus, we need to be more crafty when setting up a medium with an extreme group index.

¹For a good review, see Robert W. Boyd and Daniel J. Gauthier, “‘Slow’ and ‘Fast’ Light,” in *Progress in Optics*, vol. 43, E. Wolf, ed. (Elsevier, Amsterdam, 2002), p. 497. We roughly follow some of their arguments here.

16.3.1 Quantum Coherence: Slow Light

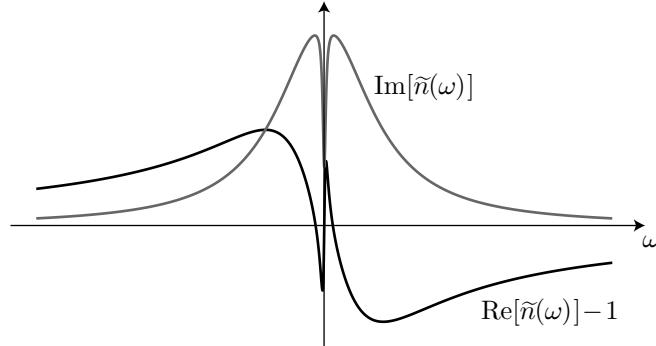
The solution to the problem of achieving a high group index in a *transparent* medium is to use a quantum interference effect called **electromagnetically induced transparency** (EIT). Roughly speaking, this effect comes about in a medium of three-level atoms in the “ Λ ” configuration, where two fields (the pump and probe) couple the two ground states to the excited state.



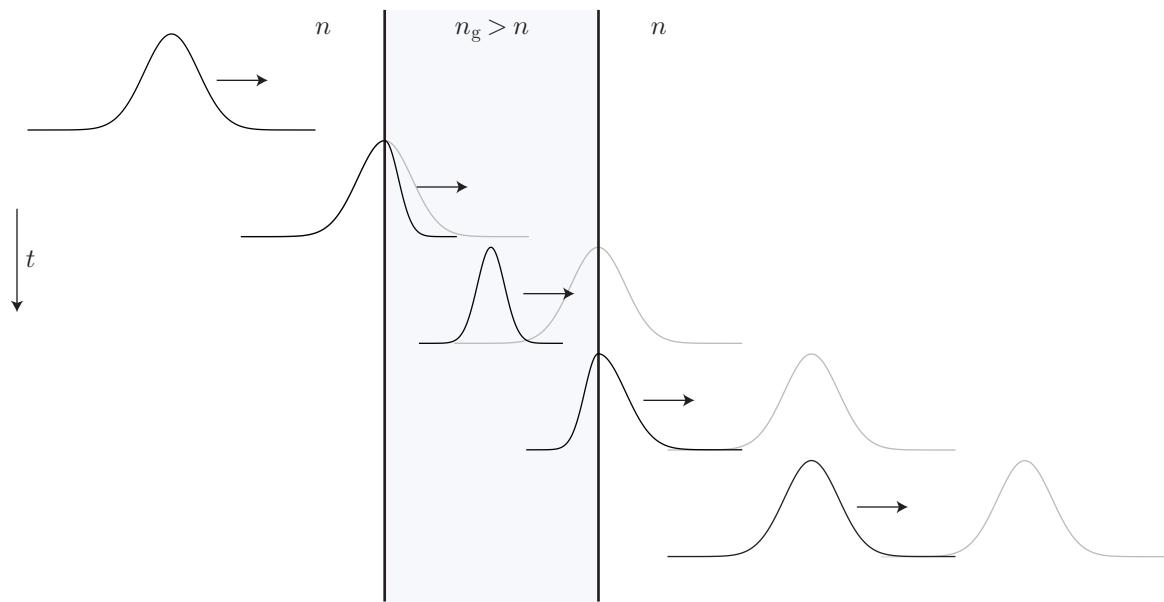
Due to quantum interference, the pump saturates the probe transition, but only if the two detunings match precisely:

$$\omega_{\text{probe}} - \omega_{\text{eg}_2} = \omega_{\text{pump}} - \omega_{\text{eg}_1}. \quad (16.63)$$

This creates a narrow “dip” in the usual Lorentzian absorption spectrum for the probe transition. For small intensities, the width of the dip can be small, on the order of the decay rate between the ground states (which can be as low as Hz to kHz).



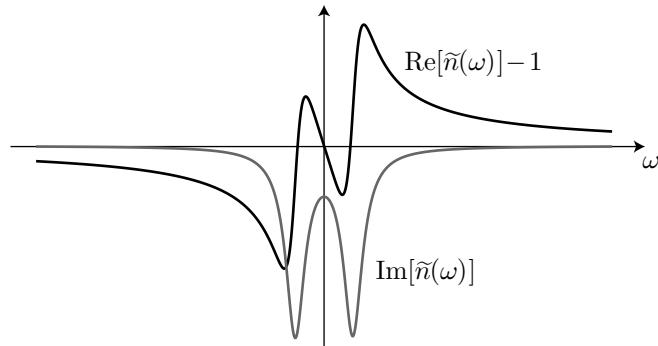
The physical significance of the dip is that the medium is transparent to the probe field, when it is properly tuned. The dip also induces a steep slope in the refractive index in the same frequency range. Since $d\eta/d\omega$ is large, the group index is correspondingly large, even though the absorption is small. A pulse propagating through the medium is thus delayed compared to pulse propagation without the medium. Again, this is a quantum-mechanical effect, so we can't describe this via the rate-equation approach that we used to model lasers. Rather, we need a fully quantum-mechanical description of the atom, which we will not get into here.



A number of slow-light experiments have achieved impressively small group velocities, down to 8 m/s for the slowest achieved velocities.²

16.3.2 Fast Light

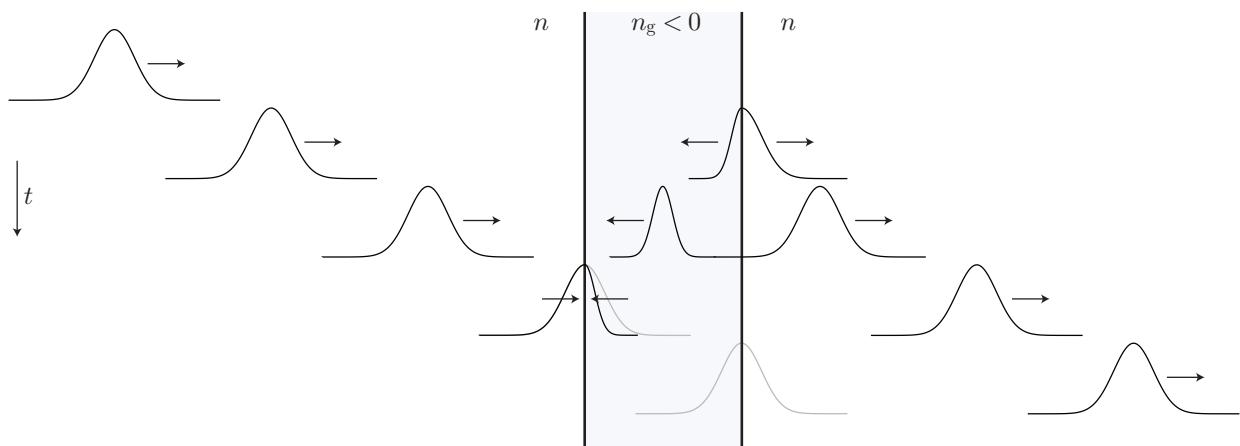
The alternative to slow light is “fast” light. One approach that was realized experimentally by Wang *et al.*³ is to use *two* pump beams to induce two *gain* peaks in the refractive-index profile. These appear as negative peaks in the absorption curve.



This arrangement flips the slope of the phase index, so that the sign of $dn/d\omega$ becomes negative. The magnitude of the derivative can be large enough that the group index can become negative, as we can see from Eq. (16.38). In the experiment of Wang *et al.*, the group index was $n_g = -310$, so that the group velocity was $v_g = -c_0/310$. Thus, a pulse entering the medium must propagate *backwards* through it. Given the continuity conditions at the boundaries, this happens if a “pulse pair” is created at the far boundary, where the backwards-propagating pulse meets the incoming pulse and annihilates it. The outgoing pulse is thus advanced with respect to the “normal” propagation of the incoming pulse, and the propagation appears to be “faster than c_0 .”

²D. Budker, D. F. Kimball, S. M. Rochester, and V. V. Yashchuk, “Nonlinear Magneto-optics and Reduced Group Velocity of Light in Atomic Vapor with Slow Ground State Relaxation,” *Physical Review Letters* **83**, 1767 (1999).

³L. J. Wang, A. Kuzmich and A. Dogariu, “Gain-assisted superluminal light propagation,” *Nature* **406**, 277 (2000).



Again, there can be no violation of causality, and in fact this behavior is completely consistent with the Kramers–Kronig relations that enforce causality.

16.4 Exercises

Problem 16.1

Consider an optical medium, where you observe that at a particular wavelength (say, 500 nm), the gas has a refractive index $n = 1.1$, but *exactly* zero absorption. From this information, what can you conclude about the existence of absorption of the gas at *other* wavelengths? *Explain.*

Problem 16.2

Show that the Kramers–Kronig relations for the complex permittivity

$$\begin{aligned}\operatorname{Re}[\tilde{\epsilon}(\omega)] &= \epsilon_0 + \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\operatorname{Im}[\tilde{\epsilon}(\omega')]}{\omega' - \omega} d\omega' \\ \operatorname{Im}[\tilde{\epsilon}(\omega)] &= -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\operatorname{Re}[\tilde{\epsilon}(\omega')] - \epsilon_0}{\omega' - \omega} d\omega'\end{aligned}\quad (16.64)$$

can be written in the equivalent form

$$\begin{aligned}\operatorname{Re}[\tilde{\epsilon}(\omega)] &= \epsilon_0 + \frac{2}{\pi} \int_0^{\infty} \frac{\omega' \operatorname{Im}[\tilde{\epsilon}(\omega')]}{\omega'^2 - \omega^2} d\omega' \\ \operatorname{Im}[\tilde{\epsilon}(\omega)] &= -\frac{2\omega}{\pi} \int_0^{\infty} \frac{\operatorname{Re}[\tilde{\epsilon}(\omega')] - \epsilon_0}{\omega'^2 - \omega^2} d\omega'.\end{aligned}\quad (16.65)$$

To do this, you will need to argue that the time-domain polarization response $g_X(t)$ applies to *real* fields and polarizations, and thus $\tilde{\epsilon}(-\omega) = \tilde{\epsilon}^*(\omega)$.

Problem 16.3

Show that for a rarefied medium ($|\chi| \ll 1$), the refractive index $n \equiv \operatorname{Re}[\tilde{n}]$ and intensity absorption coefficient a can be written in terms of the susceptibility as

$$\begin{aligned}n &\approx 1 + \frac{\operatorname{Re}[\chi]}{2} \\ a &\approx k_0 \operatorname{Im}[\chi].\end{aligned}\quad (16.66)$$

Problem 16.4

Show that the pulse spread due to propagation in a fiber composed of a dispersive material can be written⁴

$$\Delta t_{\text{pulse}} \approx \frac{\lambda_0}{c_0} \left| \frac{d^2 n}{d\lambda_0^2} \right| (\Delta\lambda_0) d,\quad (16.67)$$

where $\Delta\lambda_0$ is the wavelength range of the pulse, and d is the propagation distance. Use the following outline.

(a) Show that the group velocity may be written

$$v_g \approx \frac{c_0}{n} \left[1 + \frac{\lambda_0}{n} \frac{dn}{d\lambda_0} \right].\quad (16.68)$$

(b) Then use the fact that the pulse propagation time is $t_{\text{pulse}} = d/v_g$ and $\Delta v_g \approx (dv_g/d\lambda_0)\Delta\lambda_0$ to obtain the above expression, in the regime where $(dn/d\lambda_0)^2 \ll |d^2 n/d\lambda_0^2|$.

⁴See, for example, Peter W. Milonni and Joseph H. Eberly, *Lasers* (Wiley, 1988).

- (c) Compute the pulse spread per unit length per unit of wavelength spread for a single-mode, fused-silica fiber where the material makes the dominant contribution to dispersion. Calculate the dispersion at the popular diode-laser wavelengths of 850 nm, 1310 nm, and 1550 nm. Give your answers in units of ps/(nm · km).

To get the refractive-index derivatives, we can use the Sellmeier formula for fused silica⁵

$$n^2(\lambda_0) - 1 = \frac{0.68374049400 \lambda_0^2}{\lambda_0^2 - 0.00460352869} + \frac{0.42032361300 \lambda_0^2}{\lambda_0^2 - 0.01339688560} + \frac{0.58502748000 \lambda_0^2}{\lambda_0^2 - 64.49327320000}, \quad (16.69)$$

where λ_0 is in μm . This gives $d^2n/d\lambda_0^2 = 0.0296 \mu\text{m}^{-2}$ at 850 nm, $d^2n/d\lambda_0^2 = -0.00112 \mu\text{m}^{-2}$ at 1310 nm, and $d^2n/d\lambda_0^2 = -0.00472 \mu\text{m}^{-2}$ at 1550 nm.

Problem 16.5

Consider a bichromatic optical wave of frequencies ω_1 and ω_2 propagating through a dispersive medium.

- (a) Show that the optical phase of the wave propagates at velocity ω/k , while the envelope of the interference pattern propagates at velocity $\Delta\omega/\Delta k$, where $\omega := (\omega_1 + \omega_2)/2$, $\Delta\omega := \omega_2 - \omega_1$, $k := (k_1 + k_2)/2$, $\Delta k := k_2 - k_1$, $\omega_1 = c_0 k_1 / n_1$, $\omega_2 = c_0 k_2 / n_2$, and n_1 and n_2 are the refractive indices at frequencies ω_1 and ω_2 , respectively. Assume that $|\Delta\omega| \ll \omega$ and that the two frequency components have equal amplitudes.

- (b) What are the conditions for the envelope to propagate backwards?

⁵“HPFS Fused Silica KrF Grade,” Corning Incorporated data sheet, 2003.

Chapter 17

Classical Light–Atom Interactions

We will now model the interaction between light and atoms, using a classical model of the atom. This will allow us to treat a variety of phenomena from the refractive index of atomic vapors to the conductivity of metals to laser cooling and trapping of atoms.

We will model the atom as a classical harmonic oscillator, an electron bound to the nucleus by a harmonic force (linear spring):

$$m\ddot{\mathbf{x}} + m\omega_0^2 \mathbf{x} = 0. \quad (17.1)$$

Here, \mathbf{x} represents the *average* position of the electron, since quantum-mechanically, the electron is not localized, and ω_0 is the resonant frequency of the harmonic potential. The above equation is also in center-of-mass coordinates, so that we can ignore the motion of the nucleus. Thus, m is the *reduced mass* of the electron, given by

$$m = \frac{m_e m_n}{m_e + m_n}, \quad (17.2)$$

where m_e is the electron mass, and m_n is the nuclear mass. Generally $m_e \ll m_n$, so

$$m \approx m_e \left(1 - \frac{m_e}{m_n}\right), \quad (17.3)$$

and generally, it is a good approximation to use $m \approx m_e$.

Why use a classical calculation, when an atom is a manifestly quantum-mechanical object? It turns out that the classical calculation gets many phenomena correct, and these results can be justified by a quantum calculation. Essentially, the classical calculation is good for small excitations, when the harmonic potential, the lowest-order approximation to an arbitrary potential, is an accurate model. In particular, the classical model does not predict any saturation effects, and as we will see, it requires a bit of patching to make it quantitatively correct, even in the limit of small intensity.

17.1 Polarizability

We will now consider the interaction of the atom with a monochromatic field of the form

$$\mathbf{E}^{(+)}(t) = \hat{\varepsilon} E_0^{(+)} e^{-i\omega t}, \quad (17.4)$$

where $\hat{\varepsilon}$ is the unit polarization vector. In writing down this expression, we are making the **dipole approximation**: we are assuming that the size of the atom is much smaller than the optical wavelength, so that the electron only sees the field at the nuclear position. Thus, we need not consider the spatial dependence or propagation direction of the field. The force on the electron due to the field is

$$\mathbf{F}^{(+)} = -e\mathbf{E}^{(+)}, \quad (17.5)$$

where e is the **fundamental charge**, the magnitude of the electron charge (so that the electron charge is $-e$).

Then the equation of motion for the electron becomes

$$m\ddot{\mathbf{x}} + m\omega_0^2 \mathbf{x} = -\hat{\epsilon}eE_0^{(+)}e^{-i\omega t}. \quad (17.6)$$

We need only worry about the electron motion in the direction of the electric field; we will ignore any motion except that induced by the field, as we will justify when considering the damped version of the harmonic oscillator.

We now see if we can assume that the solution has the same time dependence as the field:

$$\mathbf{x}^{(+)}(t) = \hat{\epsilon}x_0^{(+)}e^{-i\omega t}. \quad (17.7)$$

With this solution, Eq. (17.6) becomes

$$-m\omega^2 x_0^{(+)} + m\omega_0^2 x_0^{(+)} = -eE_0^{(+)}e^{-i\omega t}, \quad (17.8)$$

which we can solve for $x_0^{(+)}$ to obtain the solution

$$x_0^{(+)} = \frac{eE_0^{(+)} / m}{\omega^2 - \omega_0^2}. \quad (17.9)$$

Again, we are breaking the electron displacement into its positive and negative components $x(t) = x^{(+)}(t) + x^{(-)}(t)$.

The dipole moment of the atom is

$$\mathbf{d}^{(+)} = -e\mathbf{x}^{(+)}, \quad (17.10)$$

where $\mathbf{x} = \hat{\epsilon}\mathbf{x}$. Since the dipole moment is induced by the field (the electron displacement is zero in equilibrium), we can define the **polarizability** α to describe how easily the field induces the dipole moment by

$$\mathbf{d}^{(+)} = \alpha(\omega)\mathbf{E}^{(+)}. \quad (17.11)$$

From Eqs. (17.9) and (17.10), we can write the polarizability as

$$\alpha(\omega) = \frac{e^2/m}{\omega_0^2 - \omega^2}. \quad (17.12)$$

The polarizability completely characterizes the response of the atom to the applied field. Of course, this is the frequency-space response function, which we have obtained via an implicit Fourier transform of the applied field.

17.1.1 Connection to Dielectric Media

Recall that the polarization density \mathbf{P} (from Section 4.3) is the dipole moment per unit volume. Thus, for an atomic vapor of number density N ,

$$\mathbf{P}^{(+)} = N\mathbf{d}^{(+)} = N\alpha(\omega)\mathbf{E}^{(+)} = \hat{\epsilon} \frac{Ne^2/m}{\omega^2 - \omega_0^2} E_0^{(+)} e^{-i\omega t}. \quad (17.13)$$

This expression is valid for a *rarefied* medium, where the interactions between the atoms are negligible. In dense media, correlations between dipoles cause deviations from these results. We can thus write the susceptibility for the vapor as

$$\chi(\omega) = \frac{Ne^2/m\epsilon_0}{\omega^2 - \omega_0^2}, \quad (17.14)$$

in view of the defining relation $\mathbf{P} = \epsilon_0\chi\mathbf{E}$.

17.1.2 Conducting Media: Plasma Model

In conducting media, the electrons are not tightly bound to the nucleus. We can treat this by letting $\omega_0 \rightarrow 0$, modeling the conductor as a plasma. Then the susceptibility becomes

$$\chi(\omega) = -\frac{Ne^2/m\epsilon_0}{\omega^2} = -\left(\frac{\omega_p}{\omega}\right)^2, \quad (17.15)$$

where

$$\omega_p := \sqrt{\frac{Ne^2}{m\epsilon_0}} \quad (17.16)$$

is the **plasma frequency**. The complex refractive index (see Section 9.5) is given in terms of the susceptibility by

$$\tilde{n}(\omega) = \sqrt{1 + \chi(\omega)} = \sqrt{1 - \left(\frac{\omega_p}{\omega}\right)^2}. \quad (17.17)$$

The behavior is quite different in the limiting cases of low and high frequency.

1. **Low frequency** ($\omega < \omega_p$). In this case, \tilde{n} is purely imaginary, so the reflection coefficient (Section 9.5) is of the form

$$r = \frac{n_1 - in_2}{n_1 + in_2}, \quad (17.18)$$

where n_1 and n_2 are real, so that $|r| = 1$. Thus, we find perfect reflection from a metal surface for frequencies below the plasma frequency.

2. **High frequency** ($\omega \gg \omega_p$). Here $\tilde{n} \rightarrow 1$, so that the metal becomes transparent at high frequencies. For typical metals, the plasma frequency occurs in the ultraviolet, so that metals are transparent to ultraviolet and shorter wavelengths.

Again, this model is valid for a rarefied plasma, so it is only qualitatively correct for metals, where density-induced correlations are more important.

17.2 Damping: Lorentz Model

A better model of the atom is a *damped* harmonic oscillator. This improved model is known as the **Lorentz model** of the atom, and the equation of motion is

$$m\ddot{\mathbf{x}} + m\gamma\dot{\mathbf{x}} + m\omega_0^2\mathbf{x} = -\hat{\varepsilon}eE_0^{(+)}e^{-i\omega t}. \quad (17.19)$$

The damping (“friction”) term models radiation reaction due to the charge acceleration (the classical analogue of spontaneous emission) and collisions with other atoms. A quantum-mechanical calculation shows that for an isolated atom, the damping rate is the same as the Einstein A coefficient (spontaneous emission rate): $\gamma = A_{21}$.

Again, we assume a solution of the form $\mathbf{x}^{(+)}(t) = \hat{\varepsilon}x_0^{(+)}e^{-i\omega t}$. Following the method above, the solution is

$$x_0^{(+)} = \frac{eE_0^{(+)} / m}{\omega^2 - \omega_0^2 + i\gamma\omega}. \quad (17.20)$$

Now the displacement is complex, reflecting a *phase lag* of the displacement behind the field, with phase angle

$$\delta = \tan^{-1} \left(\frac{\gamma\omega}{\omega_0^2 - \omega^2} \right). \quad (17.21)$$

The phase lag approaches zero for $\omega \ll \omega_0$ and π for $\omega \gg \omega_0$ ($\delta = \pi/2$ exactly on resonance). Then for this case, the polarizability becomes

$$\alpha(\omega) = \frac{e^2 / m}{\omega_0^2 - \omega^2 - i\gamma\omega}. \quad (17.22)$$

The susceptibility likewise becomes

$$\chi(\omega) = \frac{Ne^2/m\epsilon_0}{\omega_0^2 - \omega^2 - i\gamma\omega}. \quad (17.23)$$

It is worth reiterating here that α and χ are complex quantities defined for the positive-rotating fields via $\mathbf{d}^{(+)} = \alpha(\omega)\mathbf{E}^{(+)}$ and $\mathbf{P}^{(+)} = \epsilon_0\chi\mathbf{E}^{(+)}$, and therefore must be treated appropriately.

If χ is small (as for a dilute vapor),

$$\tilde{n}(\omega) = \sqrt{1 + \chi(\omega)} \approx 1 + \frac{\chi(\omega)}{2} = 1 + \frac{Ne^2}{2m\epsilon_0} \frac{(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2} + i \frac{Ne^2}{2m\epsilon_0} \frac{\gamma\omega}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}. \quad (17.24)$$

Then we can read off the phase index as the real part,

$$n(\omega) \approx 1 + \frac{Ne^2}{2m\epsilon_0} \frac{(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}, \quad (17.25)$$

while the absorption coefficient becomes

$$a(\omega) = 2k_0 \text{Im}[\tilde{n}(\omega)] \approx \frac{Ne^2\omega^2}{m\epsilon_0 c_0} \frac{\gamma}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}. \quad (17.26)$$

The region where significant absorption occurs for small detunings of the field from the atomic resonance, $|\omega - \omega_0| \ll \omega_0$. Then

$$\omega_0^2 - \omega^2 = (\omega_0 - \omega)(\omega_0 + \omega) \approx 2\omega(\omega_0 - \omega). \quad (17.27)$$

This is effectively equivalent to the *rotating-wave approximation* in quantum optics. With this approximation, the phase index and absorption become

$$\begin{aligned} n(\omega) &\approx 1 + \frac{Ne^2}{2m\epsilon_0} \frac{(\omega_0 - \omega)/2\omega}{(\omega_0 - \omega)^2 + (\gamma/2)^2} \\ a(\omega) &\approx \frac{Ne^2}{m\epsilon_0 c_0 \gamma} \frac{(\gamma/2)^2}{(\omega_0 - \omega)^2 + (\gamma/2)^2}. \end{aligned} \quad (17.28)$$

Hence, we recover the Lorentzian absorption profile from Sections 15.2.5 and 16.1.1.1 (hence the name) with full width at half maximum $\Delta\omega = \gamma$ and resonant absorption coefficient $a_0 = Ne^2/m\epsilon_0 c_0 \gamma$. Also, we see that in the same regime,

$$n - 1 = \frac{2(\omega_0 - \omega)}{\gamma} \left[\frac{Ne^2}{2m\epsilon_0 \gamma \omega} \frac{(\gamma/2)^2}{(\omega_0 - \omega)^2 + (\gamma/2)^2} \right] = \frac{2(\omega_0 - \omega)}{\gamma} \text{Im}[\tilde{n}(\omega)], \quad (17.29)$$

as required by the Kramers–Kronig relations [Eq. (16.26)].

In general, we can have atoms with multiple electrons that we need to sum over. Then the polarizability and susceptibility become

$$\begin{aligned} \alpha(\omega) &= \sum_j \frac{e^2}{m} \frac{f_{0j}}{(\omega_{0j}^2 - \omega^2 - i\gamma_j \omega)} \\ \chi(\omega) &= \sum_j \frac{Ne^2}{m\epsilon_0} \frac{f_{0j}}{(\omega_{0j}^2 - \omega^2 - i\gamma_j \omega)}. \end{aligned} \quad (17.30)$$

Here, f_{0j} is the *absorption oscillator strength*, which acts as a weighting factor for each electron. The quantum-mechanical (and correct) interpretation of these expressions is that each term in the sum represents a transition from the ground level 0 to excited level j . The oscillator strength can only be obtained from a quantum calculation, and is necessary to make the classical calculation quantitatively correct. Because of the quantitative importance of this factor, we will explore it in more detail.

17.2.1 Oscillator Strength

Since the absorption coefficient scales as the susceptibility and thus the oscillator strength (for a dilute gas), the oscillator strength also scales with the cross section [see Eq. (15.36) for the laser cross section that we defined earlier]. On resonance, the cross section for absorption is defined by

$$a(\omega_0) = \sigma(\omega_0)N = \sigma_0 N. \quad (17.31)$$

Thus, using Eq. (17.26), we can write the classical absorption cross section as

$$\sigma_{\text{classical}}(\omega_0) = \frac{e^2 \omega^2}{m\epsilon_0 c_0} \frac{\gamma}{(\omega_0^2 - \omega^2)^2 + \gamma^2 \omega^2} \Big|_{\omega=\omega_0} = \frac{e^2}{m\epsilon_0 c_0 \gamma}. \quad (17.32)$$

This cross section is not quantitatively correct, as the correct cross section from Eq. (15.39) for the transition to level j is

$$\sigma_{0j} = \sigma_j(\omega_0) = \frac{\lambda_{0j}^2}{2\pi} = \frac{\pi c_0^2}{\omega_{0j}^2}. \quad (17.33)$$

Note that this cross section assumes an orientational average (and thus a factor of $1/3$) that is generally appropriate for our purposes. We can then define the absorption oscillator strength to be the “fudge factor” to fix the classical cross section:

$$f_{0j} := \frac{\sigma_{0j}}{\sigma_{0,\text{classical}}} = \frac{2\pi\epsilon_0 m c_0^3 \gamma_j}{e^2 \omega_{0j}^2}. \quad (17.34)$$

We can also write the cross section as

$$\sigma_{0j} = f_{0j} \frac{e^2}{m\epsilon_0 c_0 \gamma_j}, \quad (17.35)$$

which will be useful later. More commonly, the absorption oscillator strength is defined to include the degeneracy of the level structure,¹

$$f_{0j} = \frac{2\pi\epsilon_0 m c_0^3 \gamma_j}{e^2 \omega_{0j}^2} \frac{g_j}{g_0}, \quad (17.36)$$

with a separate expression defined for the *emission* oscillator strength f_{0j} (which just flips the degeneracy ratio).

Also, in the limit of large frequency, the susceptibility of Eqs. (17.30) becomes

$$\chi(\omega) \longrightarrow -\frac{Ne^2}{m\epsilon_0 \omega^2} \sum_j f_{0j}. \quad (17.37)$$

In this limit, the induced electron displacements are small, and thus the damping and harmonic-potential forces are not important. We thus expect to recover the behavior of the free-electron plasma of Eq. (17.15) in the high-frequency limit (i.e., the conductor without damping):

$$\chi(\omega) = -\left(\frac{\omega_p}{\omega}\right)^2 = -\frac{Ne^2}{m\epsilon_0 \omega^2}. \quad (17.38)$$

Comparing these two expressions, we find the **Thomas–Reiche–Kuhn sum rule** for the oscillator strength:

$$\sum_j f_{0j} = 1. \quad (17.39)$$

Since $f_{0j} > 0$, the sum rule tells us that $f_{0j} < 1$. The interpretation is that the classical cross section represents the *maximum possible* cross section, which turns out to be distributed over all the possible transitions from the ground level. Note that transitions to *unbound* (ionized) states are also included in this sum, making it difficult to verify this with atomic transition data.²

¹Alan Corney, *Atomic and Laser Spectroscopy* (Oxford, 1987).

²See Peter W. Milonni and Joseph H. Eberly, *Lasers* (Wiley, 1988), p. 239.

17.2.2 Conductor with Damping: Drude Model

We can again use the Lorentz model to derive an improved model of conductors or plasmas. This model, the **Drude model**, accounts for damping, and accounts for electron collisions in the metal or plasma. We again take the Lorentz-model results and let $\omega_0 \rightarrow 0$ so that the electrons are not bound. In this limit, Eqs. (17.30) give

$$\chi(\omega) = \frac{-Ne^2/m\epsilon_0}{\omega^2 + i\gamma\omega} = \frac{-\omega_p^2}{\omega^2 + i\gamma\omega}. \quad (17.40)$$

The complex refractive index becomes

$$\tilde{n}(\omega) = \sqrt{1 + \chi(\omega)} = \sqrt{1 - \frac{\omega_p^2}{\omega^2 + i\gamma\omega}}. \quad (17.41)$$

In the dc limit ($\omega \rightarrow 0$),

$$\tilde{n}(\omega) \rightarrow \sqrt{1 + i\frac{\omega_p^2}{\gamma\omega}} \approx \sqrt{i\frac{\omega_p^2}{\gamma\omega}}. \quad (17.42)$$

Thus, the refractive index is not purely imaginary at any nonzero frequency, even though the imaginary part is much larger than the real part. This leads to a reflectance that is slightly less than unity, which is a distinct improvement over the plasma-model result of perfect reflectance. In the high-frequency limit, the damping part is not important, and the Drude model reduces to the undamped plasma model. The imaginary (damping) part of the conductivity here becomes significant at infrared and higher frequencies, gradually rolling off the material reflectance below the plasma frequency.

Another useful quantity for a conductor is the induced current density, which we can write as

$$\mathbf{J} = -Ne\dot{x}. \quad (17.43)$$

Since $\dot{x} = -i\omega x$, we can use the solution for $x_0^{(+)}$ from Eq. (17.20) to write

$$\mathbf{J} = \frac{Ne^2/m}{\gamma - i\omega} \mathbf{E} \quad (17.44)$$

in the unbound limit. Recalling that the conductivity σ is defined by $\mathbf{J} = \sigma\mathbf{E}$, we can write the Drude-model conductivity as

$$\sigma = \frac{Ne^2/m}{\gamma - i\omega} = \frac{\sigma_0}{1 - i\omega/\gamma}, \quad (17.45)$$

where

$$\sigma_0 := \frac{Ne^2}{m\gamma} \quad (17.46)$$

is the dc conductivity. We thus have a model for the frequency dependence of the metal conductivity that we ignored in Section 9.5. It is worth reiterating that this is only a simple model and valid for rarefied plasmas.

17.3 Atom Optics: Mechanical Effects of Light on Atoms

Now we will have a brief look at the field of **atom optics**, or optics with matter (de Broglie) waves. We will only be looking here at how to trap and cool atoms with laser light using the classical Lorentz model of the atom, so in a sense we will be doing “geometrical atom optics.”

Broadly speaking, there are two types of mechanical forces that light can have on atoms. The first, the *dipole force*, is related to the potential energy of the induced dipole in the electric field, and is thus related to the real part of $\alpha(\omega)$ [see Eq. (17.53)]. The second is *radiation pressure* due to absorption and rescattering of the incident light, which is thus related to the imaginary part of $\alpha(\omega)$ [see Eq. (17.89)].

17.3.1 Dipole Force

The dipole moment of the atom is induced by the external field, so we can write the potential energy of the induced dipole as

$$V_{\text{dipole}} = -\frac{\mathbf{d} \cdot \mathbf{E}}{2} = -\frac{d E}{2}. \quad (17.47)$$

The extra factor of 1/2 compared to the usual dipole energy is because the dipole is induced, and thus

$$V_{\text{dipole}} = - \int_0^E (d) dE = - \int_0^E \alpha E dE = -\frac{1}{2} \alpha E^2. \quad (17.48)$$

Since we found the solution for the positive-frequency component of the field, we should write out the potential in terms of the same components:

$$V_{\text{dipole}} = -\frac{1}{2} (\mathbf{d}^{(+)} + \mathbf{d}^{(-)}) \cdot (\mathbf{E}^{(+)} + \mathbf{E}^{(-)}). \quad (17.49)$$

Noting that

$$\mathbf{d}^{(\pm)} \sim e^{\mp i\omega t}, \quad \mathbf{E}^{(\pm)} \sim e^{\mp i\omega t}, \quad (17.50)$$

we can see that the terms of the form

$$\mathbf{d}^{(\pm)} \cdot \mathbf{E}^{(\pm)} \sim e^{\mp i2\omega t} \quad (17.51)$$

rotate at twice the optical frequency, which is too fast for the atoms to respond mechanically. So we will drop these terms in the time average (the same average that leads to the intensity). The terms of the form

$$\mathbf{d}^{(\pm)} \cdot \mathbf{E}^{(\mp)} \sim 1 \quad (17.52)$$

are dc, so we can keep these. Thus,

$$\begin{aligned} V_{\text{dipole}} &= -\frac{1}{2} \mathbf{d}^{(+)} \cdot \mathbf{E}^{(-)} - \frac{1}{2} \mathbf{d}^{(-)} \cdot \mathbf{E}^{(+)} \\ &= -\frac{1}{2} [\alpha(\omega) \mathbf{E}^{(+)}] \cdot \mathbf{E}^{(-)} - \frac{1}{2} [\alpha(\omega) \mathbf{E}^{(-)}] \cdot \mathbf{E}^{(+)} \\ &= -\text{Re}[\alpha(\omega)] |E^{(+)}|^2 \\ &= -\frac{\eta_0}{2} \text{Re}[\alpha(\omega)] I(\mathbf{r}). \end{aligned} \quad (17.53)$$

Here, η_0 is the vacuum wave impedance from Eq. (4.63), and we are regarding the electric-field envelope $E^{(+)}(\mathbf{r})$ to be a slowly varying function of position. Putting in the explicit form for the polarizability, we can write the dipole potential as

$$V_{\text{dipole}} = \frac{-e^2}{2m\epsilon_0 c_0} \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2 \omega^2} I(\mathbf{r}). \quad (17.54)$$

Thus, the atom sees a spatial potential proportional to $I(\mathbf{r})$ and to $(n - 1)$. This potential-shift effect (also known as the **ac Stark shift**) is the atomic counterpart of the phase shift (due to $n - 1$) of a beam propagating through a vapor. Both effects follow from the coupling of the field to the atom.

The corresponding force is given by the potential gradient

$$\mathbf{F}_{\text{dipole}} = -\nabla V_{\text{dipole}} \propto \nabla I(\mathbf{r}). \quad (17.55)$$

Thus, the dipole force responds to intensity *gradients*. If the dipole is viewed as two slightly separated, opposite charges, there is only a net force if the two charges see a different electric field, which is only possible in the ideal dipole limit if the field has a gradient.

The sign of the dipole potential is set solely by the detuning of the field from the atomic resonance. Defining the detuning $\Delta := \omega - \omega_0$, we can write the dipole potential as

$$V_{\text{dipole}} = \frac{e^2}{2m\epsilon_0 c_0} \frac{(\omega_0 + \omega)\Delta}{[(\omega_0 + \omega)\Delta]^2 + \gamma^2\omega^2} I(\mathbf{r}). \quad (17.56)$$

Everything in this expression is positive except for the factor of Δ in the numerator. Thus, for positive Δ ($\omega > \omega_0$, or *blue detuning*), $V_{\text{dipole}} > 0$, while for negative Δ ($\omega < \omega_0$, or *red detuning*), $V_{\text{dipole}} < 0$. That is, a bright spot in space (e.g., due to a tightly focused Gaussian beam) will repel an atom for blue detunings, forming a potential barrier, while for red detunings, the spot attracts atoms and forms a potential well.

The sign dependence of V_{dipole} makes sense in terms of the phase lag (17.21). Recall that for small frequencies ($\Delta < 0$), the phase lag of the dipole behind the field is smaller than $\pi/2$, while for large frequencies ($\Delta > 0$), the phase lag is between $\pi/2$ and π . Since $V_{\text{dipole}} \propto -\mathbf{d} \cdot \mathbf{E}$, the phase lag is important because then \mathbf{d} and \mathbf{E} are mostly aligned or mostly opposed for $\Delta < 0$ and $\Delta > 0$, respectively. Thus, $V_{\text{dipole}} \gtrless 0$ for $\Delta \gtrless 0$.

17.3.1.1 Dipole Potential: Standard Form

By writing the dipole potential in a more standard form, we can see that it matches the result of a quantum calculation, at least in the limit of low intensity. To do this, we first need to patch the classical result of Eq. (17.54) as before by including the oscillator strength and summing over all transitions:

$$V_{\text{dipole}} = - \sum_j \frac{e^2 f_{0j}}{2m\epsilon_0 c_0} \frac{\omega_{0j}^2 - \omega^2}{(\omega_{0j}^2 - \omega^2)^2 + \gamma_j^2 \omega^2} I(\mathbf{r}). \quad (17.57)$$

Now to put this in more standard form, we need to define the saturation intensity for the atom. We defined the saturation intensity before in Eq. (15.75) for a four-level laser, but we need to redo this for the two-level atom. Recall that a nondegenerate, two-level atom has the steady state solution when driven on resonance from Eq. (15.23)

$$\frac{N_2}{N_1} = \frac{B_{12}\rho(\omega_0)}{A_{21} + B_{21}\rho(\omega_0)}. \quad (17.58)$$

Using $N_1 + N_2 = N$ and $B_{12} = B_{21}$, we can solve this to find

$$\frac{N_2}{N} = \frac{B_{21}\rho(\omega)}{A_{21} + 2B_{21}\rho(\omega)} = \frac{\sigma_0 I / \hbar\omega_0}{A_{21} + 2\sigma_0 I / \hbar\omega_0} = \frac{I / 2I_{\text{sat}}}{1 + I / I_{\text{sat}}}, \quad (17.59)$$

where σ_0 is the resonant cross section and we have defined the saturation intensity as

$$I_{\text{sat}} := \frac{\hbar\omega_0\gamma}{2\sigma_0}, \quad (17.60)$$

and we have identified $\gamma = A_{21}$. Note that this definition differs by a factor of 2 from the four-level laser definition (15.75), since here the ground state is not naturally depleted. For the maximum possible resonant cross section of $\sigma_0 = 3\lambda^2/2\pi$ (where there is no average over the dipole orientation), the saturation intensity is $I_{\text{sat}} = 1.10 \text{ mW/cm}^2$ for ^{133}Cs on the D₂ transition (852 nm), while for ^{87}Rb on the same transition (780 nm), the saturation intensity is $I_{\text{sat}} = 1.67 \text{ mW/cm}^2$. We can also write the saturation intensity in terms of the oscillator strength by using Eq. (17.35), with the result

$$I_{\text{sat},j} = \frac{\hbar\omega_{0j}m\epsilon_0 c_0 \gamma_j^2}{2e^2 f_{0j}}. \quad (17.61)$$

Even though the above numerical values are often quoted for the saturation intensity, this is actually a context-dependent quantity (as we can see from the difference of its value in the two- and four-level cases). A safe but cumbersome approach is to use the quantum-mechanical formalism for angular momentum to directly calculate the cross section and thus saturation intensity.³

³Daniel A. Steck, “Cesium D Line Data,” 2003. Available online at <http://steck.us/alkalidata>.

Using Eq. (17.61), we can write the dipole potential (17.57) as

$$V_{\text{dipole}} = - \sum_j \frac{\hbar\omega_{0j}\gamma_j^2}{4} \frac{\omega_{0j}^2 - \omega^2}{(\omega_{0j}^2 - \omega^2)^2 + \gamma_j^2\omega^2} \frac{I(\mathbf{r})}{I_{\text{sat},j}}. \quad (17.62)$$

This is the general expression for any frequency, so long as the intensity is small. To simplify this, we can look at the functional form far away from all resonances ($|\omega_{0j} - \omega| \gg \gamma_j$ for all j) so that

$$\begin{aligned} V_{\text{dipole}} &= \sum_j \frac{\hbar\omega_{0j}\gamma_j^2}{4} \frac{1}{(\omega^2 - \omega_{0j}^2)} \frac{I(\mathbf{r})}{I_{\text{sat},j}} \\ &= \sum_j \frac{\hbar\gamma_j^2}{8} \left(\frac{1}{\omega - \omega_{0j}} - \frac{1}{\omega + \omega_{0j}} \right) \frac{I(\mathbf{r})}{I_{\text{sat},j}}. \end{aligned} \quad (17.63)$$

The first term in the parentheses is the inverse of the detuning, and represents the Stark shift due to the atomic resonances. The second term can be interpreted as the weak, additional Stark shift due to resonances at the corresponding *negative frequencies*. This secondary shift is always negative (like a red detuning), and corresponds to the **Bloch–Siegert shift**. Note that this expression also recovers the *dc* Stark shift (or equivalently, the dc polarizability up to some universal factor) when $\omega = 0$, when both terms contribute equal, negative energy shifts.

If one resonance is dominant (that is, the laser is tuned far away from resonance, but much closer to one than all the others), then we can make the rotating-wave approximation and neglect the second term in the parentheses of Eq. (17.63) to obtain

$$V_{\text{dipole}} = \frac{\hbar\gamma_j^2}{8\Delta} \frac{I(\mathbf{r})}{I_{\text{sat}}}, \quad (17.64)$$

where again $\Delta = \omega - \omega_0$ is the detuning from resonance. Note that for a far-detuned, linearly polarized laser creating this potential, it turns out that $\sigma_0 = \lambda_0^2/2\pi$ is the appropriate resonant cross section, so the above saturation intensity values should be multiplied by 3 before being used in this formula.

Typically, a focused, red-detuned, Gaussian laser beam is used to make a **dipole trap** or *far-off resonance trap* (FORT)⁴ for atoms via the dipole force.⁵ Below is an example image of about 10^5 ^{87}Rb atoms confined in a dipole trap formed by a 10 W, 1090 nm Gaussian laser beam (far below the 780 and 794 nm main resonances) focused to a $31\ \mu\text{m}$ beam waist ($1/e^2$ radius), implying a Rayleigh length (depth of focus along the beam direction) of 2.8 mm.



The dipole trap clearly runs from left to right with a slight downward angle; the dimensions of the image are 270×29 CCD pixels (6.59×0.71 mm). This is an **absorption image**, where the shadow cast by the atoms in a brief, resonant laser probe is imaged and recorded on a CCD camera. (The image greyscale is inverted so the atoms appear bright rather than dark.)

But now the important question to address is under what conditions the trap is *stable*, since as the atom scatters photons, it heats up until it boils out of the trap. So we will need to take a closer look at the *radiation* of the Lorentz atom.

17.3.2 Radiation Pressure

Now we will examine the forces due to absorption and reemission of the incident light. We will begin by examining the radiation process in detail.

⁴Steven L. Rolston, Christoph Gerz, Kristian Helmerson, P. S. Jessen, Paul D. Lett, William D. Phillips, R. J. Spreeuw, and C. I. Westbrook, “Trapping atoms with optical potentials,” *Proceedings of SPIE* **1726**, 205 (1992) (doi: 10.1117/12.130392).

⁵The first observation of atoms trapped in a dipole-force potential was Steven Chu, J. E. Bjorkholm, A. Ashkin, and A. Cable, “Experimental Observation of Optically Trapped Atoms,” *Physical Review Letters* **57**, 314 (1986) (doi: 10.1103/PhysRevLett.57.314).

17.3.2.1 Dipole Radiation

The electric and magnetic fields for an oscillating dipole are⁶

$$\begin{aligned}\mathbf{E}^{(+)}(\mathbf{r}, t) &= \frac{1}{4\pi\epsilon_0} [3(\hat{\varepsilon} \cdot \hat{r})\hat{r} - \hat{\varepsilon}] \left[\frac{d^{(+)}(t_r)}{r^3} + \frac{\dot{d}^{(+)}(t_r)}{c_0 r^2} \right] + \frac{1}{4\pi\epsilon_0} [(\hat{\varepsilon} \cdot \hat{r})\hat{r} - \hat{\varepsilon}] \frac{\ddot{d}^{(+)}(t_r)}{c_0^2 r} \\ \mathbf{H}^{(+)}(\mathbf{r}, t) &= \frac{c_0}{4\pi} (\hat{\varepsilon} \times \hat{r}) \left[\frac{\dot{d}^{(+)}(t_r)}{c_0 r^2} + \frac{\ddot{d}^{(+)}(t_r)}{c_0^2 r} \right],\end{aligned}\quad (17.65)$$

where $t_r = t - r/c_0$ is the retarded time, and $\hat{\varepsilon}$ is the polarization unit vector of the applied field (and thus the dipole orientation vector). Only the $1/r$ terms actually transport energy to infinity (i.e., they correspond to radiation), so we can drop the rest to obtain

$$\begin{aligned}\mathbf{E}^{(+)}(\mathbf{r}, t) &\approx \frac{1}{4\pi\epsilon_0 c_0^2} [(\hat{\varepsilon} \cdot \hat{r})\hat{r} - \hat{\varepsilon}] \frac{\ddot{d}^{(+)}(t_r)}{r} \\ \mathbf{H}^{(+)}(\mathbf{r}, t) &\approx \frac{1}{4\pi c_0} (\hat{\varepsilon} \times \hat{r}) \frac{\ddot{d}^{(+)}(t_r)}{r}.\end{aligned}\quad (17.66)$$

The energy transport is governed by the Poynting vector, which we can write as

$$\begin{aligned}\langle \mathbf{S} \rangle &= \mathbf{E}^{(+)} \times \mathbf{H}^{(-)} + \text{c.c.} \\ &= \frac{1}{16\pi^2\epsilon_0 c_0^3} \frac{|\ddot{d}^{(+)}|^2}{r^2} [(\hat{\varepsilon} \cdot \hat{r})\hat{r} - \hat{\varepsilon}] \times (\hat{\varepsilon}^* \times \hat{r}) + \text{c.c.} \\ &= \frac{\hat{r}}{16\pi^2\epsilon_0 c_0^3} \frac{|\ddot{d}^{(+)}|^2}{r^2} \left(1 - |\hat{r} \cdot \hat{\varepsilon}|^2 \right) + \text{c.c.},\end{aligned}\quad (17.67)$$

where we have used

$$[(\hat{\varepsilon} \cdot \hat{r})\hat{r} - \hat{\varepsilon}] \times (\hat{\varepsilon}^* \times \hat{r}) = \left(1 - |\hat{r} \cdot \hat{\varepsilon}|^2 \right) \hat{r} \quad (17.68)$$

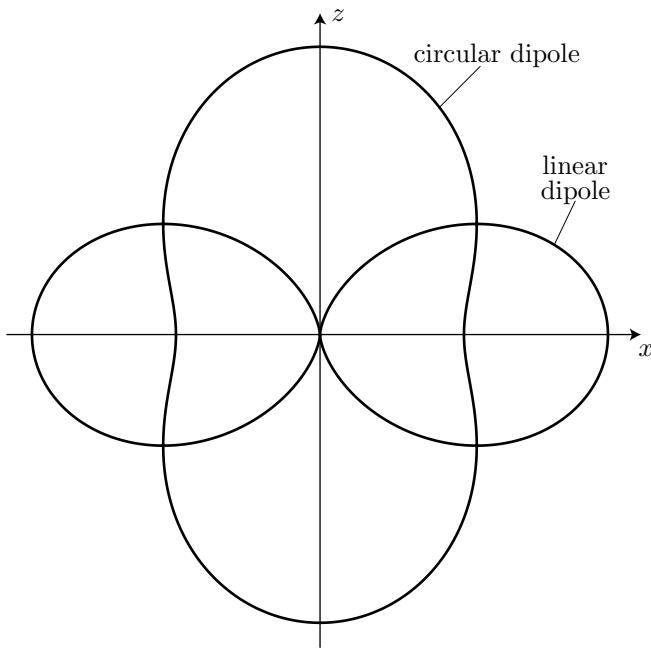
for the angular dependence.

There are two main possibilities for the polarization vector: the incident light can be linearly or circularly polarized.

1. **Linear polarization** ($\hat{\varepsilon} = \hat{z}$): $1 - |\hat{r} \cdot \hat{\varepsilon}|^2 = \sin^2 \theta$. This is the usual “doughnut-shaped” radiation pattern for an oscillating dipole.
2. **Circular polarization** ($\hat{\varepsilon} = \hat{\varepsilon}_\pm := \mp(\hat{x} \pm i\hat{y})/\sqrt{2}$): $1 - |\hat{r} \cdot \hat{\varepsilon}|^2 = (1 + \cos^2 \theta)/2$. This is a “peanut-shaped” radiation pattern for a rotating dipole.

Here, θ is the angle from the z -axis, while ϕ is the angle around the azimuth. Note that any arbitrary polarization can be represented as a superposition of these three basis vectors. The (intensity/power) radiation patterns for the linear and circular dipole cases are shown here.

⁶See John David Jackson, *Classical Electrodynamics*, 3rd ed. (Wiley, 1999), p. 411 or Peter W. Milonni and Joseph H. Eberly, *Lasers* (Wiley, 1988), p. 44.



The three-dimensional distributions are generated by sweeping these patterns around the z -axis.

The corresponding electric fields for the dipole radiation are polarized. From Eq. (17.66), we can see that the polarization vector is proportional to $(\hat{\varepsilon} \cdot \hat{r})\hat{r} - \hat{\varepsilon}$. For linear polarization ($\hat{\varepsilon} = \hat{z}$), this factor turns out to be $\sin \theta \hat{\theta}$, while for circular polarization ($\hat{\varepsilon} = \hat{\varepsilon}_\pm = \mp(\hat{x} \pm i\hat{y})/\sqrt{2}$), the polarization vector is proportional to $(\cos \theta \hat{\theta} \mp i\hat{\phi})e^{\mp i\phi}/\sqrt{2}$.

Now let's define the angular-distribution function via

$$f_{\hat{\varepsilon}}(\theta, \phi) := \frac{3}{8\pi} \left(1 - |\hat{r} \cdot \hat{\varepsilon}|^2 \right). \quad (17.69)$$

For linear and circular polarization, this takes the form

$$\begin{aligned} f_{\hat{z}}(\theta, \phi) &= \frac{3}{8\pi} \sin^2(\theta) \\ f_{\pm}(\theta, \phi) &= \frac{3}{16\pi} [1 + \cos^2(\theta)]. \end{aligned} \quad (17.70)$$

This function has the nice property that it is normalized, and thus represents a probability distribution for photon emission in quantum mechanics:

$$\int f_{\hat{\varepsilon}}(\theta, \phi) d\Omega = 1. \quad (17.71)$$

Here, $d\Omega = \sin \theta d\theta d\phi$ is the usual solid-angle element.

Now we can write the Poynting vector in terms of the angular-distribution function as

$$\langle \mathbf{S} \rangle = \frac{\hat{r}}{3\pi\epsilon_0 c_0^3} \frac{|\ddot{d}^{(+)}|^2}{r^2} f_{\hat{\varepsilon}}(\theta, \phi). \quad (17.72)$$

The power radiated per unit solid angle is then

$$\frac{dP_{\text{rad}}}{d\Omega} = r^2 \langle \mathbf{S} \rangle \cdot \hat{r} = \frac{|\ddot{d}^{(+)}|^2}{3\pi\epsilon_0 c_0^3} f_{\hat{\varepsilon}}(\theta, \phi), \quad (17.73)$$

and the total radiated power is

$$P_{\text{rad}} = \int d\Omega \frac{dP_{\text{rad}}}{d\Omega} = \frac{|\ddot{x}^{(+)}|^2}{3\pi\epsilon_0 c_0^3} = \frac{e^2 |\ddot{x}^{(+)}|^2}{3\pi\epsilon_0 c_0^3}. \quad (17.74)$$

Of course, the incident intensity is contained implicitly in the electron acceleration \ddot{x} .

17.3.2.2 Damping Coefficient

Now we can connect the radiated power to the damping term in the Lorentz model,⁷ Eq. (17.6). Note that the radiated power in Eq. (17.74) is the *time-averaged* power, since we used the complex representation. In terms of the real displacement, we can make the replacement

$$|\ddot{x}^{(+)}|^2 \longrightarrow \frac{\langle \ddot{x}^2 \rangle}{2}, \quad (17.75)$$

where the angle brackets denote the time average. Then the average work done by radiation reaction must balance the energy emitted into the field:

$$\begin{aligned} \int_{x_0}^x \mathbf{F}_{\text{rr}} \cdot d\mathbf{x}' &= \int_{t_0}^t F_{\text{rr}} \dot{x}(t') dt' = -\frac{e^2}{6\pi\epsilon_0 c_0^3} \int_{t_0}^t (\ddot{x})^2 dt' \\ &= -\frac{e^2}{6\pi\epsilon_0 c_0^3} \left[\dot{x}\ddot{x} \Big|_{t_0}^t - \int_{t_0}^t \dot{x}\ddot{x} dt' \right]. \end{aligned} \quad (17.76)$$

Here \mathbf{F}_{rr} refers to the radiation-reaction force. If we pick $t - t_0$ to be an integer multiple of the optical period, the boundary term vanishes (it is also negligible for large $t - t_0$). Then the radiation-reaction force is

$$\mathbf{F}_{\text{rr}} = \frac{e^2}{6\pi\epsilon_0 c_0^3} \ddot{\mathbf{x}}. \quad (17.77)$$

Going back to the complex representation and noting that the displacement is a harmonic function,

$$\mathbf{F}_{\text{rr}}^{(+)} = \frac{e^2}{6\pi\epsilon_0 c_0^3} \ddot{\mathbf{x}}^{(+)} \approx -\frac{e^2 \omega_0^2}{6\pi\epsilon_0 c_0^3} \dot{\mathbf{x}}^{(+)}. \quad (17.78)$$

If we define

$$\gamma = \frac{e^2 \omega_0^2}{6\pi m \epsilon_0 c_0^3}, \quad (17.79)$$

then we recover the damping term in the harmonic oscillator:

$$\mathbf{F}_{\text{rr}}^{(+)} = -m\gamma \dot{z}^{(+)}. \quad (17.80)$$

This is the classical result for the spontaneous emission rate, which isn't quite correct. Again, we can patch this with the substitution $e^2/m \longrightarrow (e^2/m)f_{0j}$, with the result

$$\gamma_j = \frac{e^2 \omega_{0j}^2 f_{0j}}{6\pi m \epsilon_0 c_0^3}. \quad (17.81)$$

This is consistent with Eq. (17.34) if we take $\sigma_{0j} = 3\lambda_0^2/2\pi$ (i.e., no orientational average for the dipole). Again, there are some subtleties here regarding the cross sections and orientational averages that are better handled by angular-momentum algebra.

⁷This argument follows Alan Corney, *Atomic and Laser Spectroscopy* (Oxford, 1987), p. 230.

17.3.2.3 Photon Scattering Rate

Now we can compute the rate of photon scattering as a way to get to the rate of momentum transfer. We can write the total radiated power from Eq. (17.74) in terms of the polarizability as

$$P_{\text{rad}} = \frac{\omega^4 |\alpha(\omega)|^2}{6\pi\epsilon_0^2 c_0^4} I, \quad (17.82)$$

where we used $d^{(+)} = \alpha(\omega)E_0^{(+)}e^{-i\omega t}$.

As a brief aside, though, we can write down the scattering cross section from the total radiated power, given the defining relation $P_{\text{rad}} = \sigma I$:

$$\sigma_{\text{Rayleigh}} = \frac{\omega^4 |\alpha(\omega)|^2}{6\pi\epsilon_0^2 c_0^4}. \quad (17.83)$$

The overall scaling is as ω^4 (neglecting the small modification due to the polarizability). This is the usual explanation for why the sky is blue and sunsets are red: blue wavelengths are preferentially scattered by the atmosphere, while red wavelengths are preferentially transmitted.

We can continue by writing out explicitly the polarizability in Eq. (17.82), using Eq. (17.30):

$$P_{\text{rad}} = \frac{e^2 \omega^4}{6\pi m^2 \epsilon_0^2 c_0^2} \left| \sum_j \frac{f_{0j}}{\omega_{0j}^2 - \omega^2 - i\gamma_j \omega} \right|^2 I. \quad (17.84)$$

Using Eq. (17.81) to eliminate the oscillator strengths,

$$\begin{aligned} P_{\text{rad}} &= 6\pi c_0^2 \left| \sum_j \frac{\omega^2}{\omega_{0j}^2} \frac{\gamma_j}{\omega_{0j}^2 - \omega^2 - i\gamma_j \omega} \right|^2 I \\ &= \frac{\hbar}{2} \left| \sum_j \frac{\omega^2}{\sqrt{\omega_{0j}}} \frac{\gamma_j^{3/2}}{\omega_{0j}^2 - \omega^2 - i\gamma_j \omega} \sqrt{\frac{I}{I_{\text{sat},j}}} \right|^2, \end{aligned} \quad (17.85)$$

where we used $\sigma_{0j} = 3\lambda_0^2/2\pi$ to write the saturation intensity as

$$I_{\text{sat},j} = \frac{\hbar\omega_{0j}\gamma_j}{2\sigma_{0j}} = \frac{\hbar\omega_{0j}^3\gamma_j}{4\pi c_0^2}. \quad (17.86)$$

The photon scattering rate R_{sc} is the radiated power divided by the photon energy $\hbar\omega$:

$$R_{\text{sc}} = \frac{P_{\text{rad}}}{\hbar\omega} = \left| \sum_j \frac{\omega^{3/2}}{\sqrt{2\omega_{0j}}} \frac{\gamma_j^{3/2}}{\omega_{0j}^2 - \omega^2 - i\gamma_j \omega} \sqrt{\frac{I}{I_{\text{sat},j}}} \right|^2. \quad (17.87)$$

Again, this expression simplifies greatly for certain detunings. Near one resonance, we can ignore the contribution of the others:

$$\begin{aligned} R_{\text{sc}} &\approx \frac{\omega^3}{2\omega_0} \frac{\gamma^3}{|\omega_0^2 - \omega^2 - i\gamma\omega|^2} \frac{I}{I_{\text{sat}}} \\ &= \frac{\omega^3}{2\omega_0} \frac{\gamma^3}{(\omega_0^2 - \omega^2)^2 + \gamma^2 \omega^2} \frac{I}{I_{\text{sat}}} \end{aligned} \quad (17.88)$$

Using Eq. (17.30) restricted to a single resonance, we find

$$R_{\text{sc}} = \frac{\eta_0}{\hbar} \frac{\omega^2}{\omega_0^2} \text{Im}[\alpha] I(\mathbf{r}), \quad (17.89)$$

which shows the connection of the scattering rate (and hence the radiation pressure force below) to the absorptive part of the polarizability.

Far away from the dominant resonance ($|\Delta| \gg \gamma$), but still close enough for the resonance to still dominate, we find that

$$R_{\text{sc}} \approx \frac{\gamma^3}{8\Delta^2} \frac{I}{I_{\text{sat}}} = \frac{\gamma}{\hbar\Delta} V_{\text{dipole}}, \quad (17.90)$$

where in the last formula we have used Eq. (17.64) for the dipole potential. This result is of prime importance in the design of an optical dipole trap. The photon scattering rate represents *heating* of the atoms, as we will discuss, because of the random nature of photon emission. But the scattering rate and dipole potential scale as

$$R_{\text{sc}} \propto \frac{I}{\Delta^2}; \quad V_{\text{dipole}} \propto \frac{I}{\Delta}, \quad (17.91)$$

so that for a given desired potential depth, the scattering (heating) rate can be made very small by making the detuning Δ large, and compensate by increasing the intensity. Thus, dipole traps with small heating rates and hence long lifetimes (up to minutes for dipole traps created by CO₂ laser light) can be created in this way. Note that for a linearly polarized, far-detuned dipole trap where these scalings are valid, the same saturation intensities are to be used to calculate the dipole force and scattering rate, as discussed above.

17.3.2.4 Scattering Force

Each photon carries a momentum of $\hbar k_0$. Thus, the photon scattering rate implies a rate of momentum transfer and thus a *radiation pressure force* of

$$\mathbf{F}_{\text{rad}} = \hbar k_0 R_{\text{sc}} \hat{k}, \quad (17.92)$$

where \hat{k} is the unit vector along the direction of the wave vector (propagation direction) \mathbf{k} . Note that even though we have invoked the concept of the photon here, which will be convenient when discussing the heating rate, everything here is really classical, since the classical field momentum is related to the absorbed beam power: $F = dp/dt = P_{\text{abs}}/c_0 = \sigma I/c_0$.

To get a sense of scale of the momenta involved we can compute the **recoil velocity** v_r , defined as the velocity corresponding to one photon recoil momentum $\hbar k_0$:

$$v_r = \frac{\hbar k_0}{m}. \quad (17.93)$$

For ¹³³Cs at 852 nm, $v_r = 3.5$ mm/s, and for ⁸⁷Rb at 780 nm, $v_r = 5.9$ mm/s, so the recoil velocity is orders of magnitude smaller than typical room-temperature velocities.

Close to a single resonance, the scattering rate from Eq. (17.88) is

$$R_{\text{sc}} = \frac{(\gamma/2)^3}{\Delta^2 + (\gamma/2)^2} \frac{I}{I_{\text{sat}}}, \quad (17.94)$$

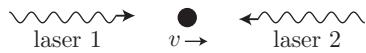
so that the radiation pressure force becomes

$$\mathbf{F}_{\text{rad}} = \frac{\hbar k_0 (\gamma/2)^3}{\Delta^2 + (\gamma/2)^2} \frac{I}{I_{\text{sat}}} \hat{k}. \quad (17.95)$$

Again, depending on the polarization and exactly how close the detuning is (i.e., whether or not the hyperfine structure is resolved), the appropriate value of the saturation intensity might be very different, so some caution is necessary in applying these formulae.

17.3.3 Laser Cooling: Optical Molasses

Now let's explore how we can use the radiation-pressure force to cool atoms. The simplest setup we can consider is an atom moving with velocity \mathbf{v} , exposed to identical but counterpropagating laser fields along the velocity direction.



The radiation-pressure force on the atom due to the two fields from Eq. (17.95) is

$$F_{\text{rad}} = \hbar k_0 (\gamma/2)^3 \left(\frac{1}{\Delta_1^2 + (\gamma/2)^2} - \frac{1}{\Delta_2^2 + (\gamma/2)^2} \right) \frac{I}{I_{\text{sat}}}, \quad (17.96)$$

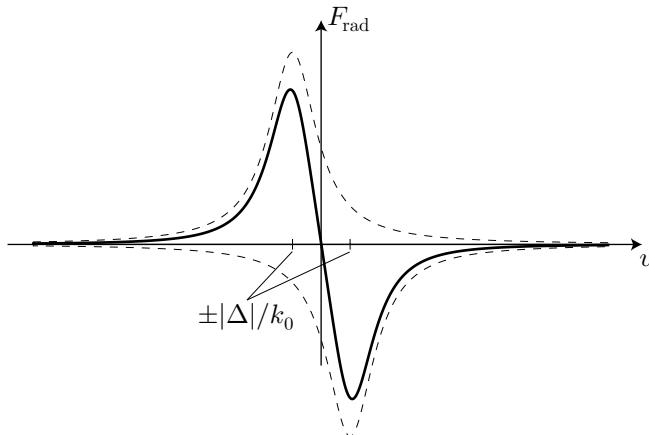
where $\Delta_{1,2}$ are the effective detunings of the two lasers. The detunings of the two lasers are the same in the laboratory frame, but the idea behind Doppler cooling is to tune the lasers *below* the atomic resonance, so that the beam that opposes the atomic velocity is Doppler-shifted into resonance, thus tending to stop the atom. With the pictured setup, the frequency of laser 1 is Doppler shifted (red shifted) by $-k_0 v$, while the frequency of laser 2 is Doppler shifted (blue shifted) by $+k_0 v$. Since the detunings are given by $\Delta_{1,2} = \omega_{1,2} - \omega_0$, we can write

$$\begin{aligned} \Delta_1 &= \Delta - k_0 v \\ \Delta_2 &= \Delta + k_0 v, \end{aligned} \quad (17.97)$$

where $\Delta = \omega - \omega_0$ is the detuning in the laboratory frame. Then the force is

$$F_{\text{rad}} = \hbar k_0 (\gamma/2)^3 \left(\frac{1}{(\Delta - k_0 v)^2 + (\gamma/2)^2} - \frac{1}{(\Delta + k_0 v)^2 + (\gamma/2)^2} \right) \frac{I}{I_{\text{sat}}}. \quad (17.98)$$

Regarded as a function of velocity, this expression is the difference of two Lorentzians, each displaced by $|\Delta|/k_0$ from zero velocity. This force is plotted for $\Delta = -\gamma/2$, and the two offset Lorentzians are shown as dashed lines.



For small velocity [$v \ll \max(|\Delta|, \gamma)/k_0$], we can expand to lowest order in v to obtain the viscous damping (“friction”) force:

$$F_{\text{rad}} = \frac{\hbar k_0^2 \gamma^3}{2} \frac{\Delta}{[\Delta^2 + (\gamma/2)^2]^2} \frac{I}{I_{\text{sat}}} v. \quad (17.99)$$

Because this force is damping for $\Delta < 0$, typically leading to heavily overdamped motion for trapped atoms, this light configuration is called **optical molasses**.⁸ The maximum damping rate occurs for $\Delta = -\gamma/2\sqrt{3}$, although it turns out that the optimal detuning is actually something else for reasons we will soon discuss.

The velocity capture range is the range in velocity for which the force is appreciable. Thus, the capture range is on the order of

$$\pm \frac{|\Delta|}{k} \sim \pm \frac{\gamma}{2k} = \pm \frac{\gamma \lambda}{4\pi}, \quad (17.100)$$

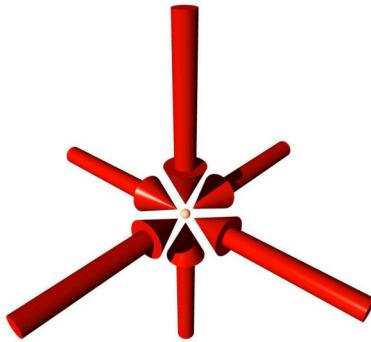
assuming $\Delta \sim -\gamma/2$. For both ^{133}Cs ($\gamma = 32.8 \times 10^6 \text{ s}^{-1}$, $\lambda_0 = 852 \text{ nm}$) and ^{87}Rb ($\gamma = 38.1 \times 10^6 \text{ s}^{-1}$, $\lambda_0 = 780 \text{ nm}$), the capture range is about $\pm 2 \text{ m/s}$. Thus, only fairly slowly moving atoms can be cooled

⁸Steven Chu, J. E. Bjorkholm, A. Ashkin, and A. Cable, “Experimental Observation of Optically Trapped Atoms,” *Physical Review Letters* **57**, 314 (1986).

at all with this method. Traditionally to load atomic traps, atoms were slowed by other methods from hot atomic beams to below the capture velocity and then trapped. However, it is possible to load a trap from room-temperature vapor with this method by capturing only the small fraction of the atoms with small enough velocity.

17.3.3.1 Doppler Cooling Limit

For laser cooling in three dimensions, it is sufficient to simply combine three of the above one-dimensional setups, one along each axis.⁹



Then we can write the force vector for small velocities as,

$$\mathbf{F}_{\text{rad}} = \frac{\hbar k_0^2 \gamma^3}{2} \frac{\Delta}{[\Delta^2 + (\gamma/2)^2]^2} \frac{I}{I_{\text{sat}}} \mathbf{v}. \quad (17.101)$$

where I is still the intensity of a *single* beam.

Our treatment so far makes it appear as though the atomic velocity may be damped completely away. However, we have only considered the *average* cooling force. There are also *fluctuations* of the cooling force that lead to a temperature limit. We will now derive this temperature limit, the **Doppler limit**, for the cooling mechanism presented here.

Let's look at the variance of the velocity distribution:

$$\frac{d}{dt} \langle v^2 \rangle = 2 \left\langle \mathbf{v} \cdot \frac{d\mathbf{v}}{dt} \right\rangle = \frac{2}{m_a} \left\langle \mathbf{v} \cdot \frac{d\mathbf{p}}{dt} \right\rangle = \frac{2}{m_a} \langle \mathbf{v} \cdot \mathbf{F}_{\text{rad}} \rangle. \quad (17.102)$$

Here, m_a is the atomic mass, and the angle brackets denote an ensemble average. With the small-velocity expression (17.99) for the average cooling force, this equation of motion becomes

$$\frac{d}{dt} \langle v^2 \rangle = \frac{\hbar k_0^2 \gamma^3}{m_a} \frac{\Delta}{[\Delta^2 + (\gamma/2)^2]^2} \frac{I}{I_{\text{sat}}} \langle v^2 \rangle. \quad (17.103)$$

Again, according to this differential equation, the velocity damps to zero for $\Delta < 0$.

Now we will include the force fluctuations heuristically, since the fluctuations are quantum-mechanical in origin (although there is a more general connection between damping and fluctuations known as the *fluctuation-dissipation theorem*). In the course of scattering a photon from one of the laser beams, there is a photon absorption and a photon emission. Each absorption leads to a momentum “kick” of magnitude $\hbar k_0$, and the direction is random but along one of the six beams. The emission is also in a random direction (not in a dipole-radiation pattern if we assume all polarizations to be equally present), leading to a second kick of magnitude $\hbar k_0$ in a random direction. Thus, a scattering event is effectively equivalent to two steps in a random walk in velocity space, where the step size is $\hbar k_0/m_a$. These scattering events happen at the scattering rate

$$R_{\text{sc}} = \frac{(\gamma/2)^3}{\Delta^2 + (\gamma/2)^2} \frac{6I}{I_{\text{sat}}}, \quad (17.104)$$

⁹Graphics by Windell Oskay.

since there are six beams present. Recall from Section 11.5.1 that for a random walk, each step increases $\langle v^2 \rangle$ by a fixed amount, given by the variance after one step starting from the origin. The three-dimensional probability distribution for a single scattering event is confined to a shell of radius $\hbar k_0/m_a$ in velocity space for either the absorption or emission event. The probability distribution is also inversion symmetric in either case. Thus if the atom is initially at rest, then after one step in the random walk, we can write

$$\langle v_{\text{initial}}^2 \rangle = 0 \quad \longrightarrow \quad \langle v_{\text{final}}^2 \rangle = \left(\frac{\hbar k_0}{m_a} \right)^2, \quad (17.105)$$

so that $\langle v^2 \rangle$ increases at the rate

$$2R_{\text{sc}} \left(\frac{\hbar k_0}{m_a} \right)^2. \quad (17.106)$$

Including the heating rate in Eq. (17.103), we find

$$\frac{d}{dt} \langle v^2 \rangle = \frac{\hbar k_0^2 \gamma^3}{m_a} \frac{\Delta}{[\Delta^2 + (\gamma/2)^2]^2} \frac{I}{I_{\text{sat}}} \langle v^2 \rangle + \frac{3\gamma^3}{2} \frac{1}{\Delta^2 + (\gamma/2)^2} \frac{I}{I_{\text{sat}}} \left(\frac{\hbar k_0}{m_a} \right)^2. \quad (17.107)$$

In steady state, we can set the right-hand side to zero, with the result

$$\langle v^2 \rangle = \frac{3\hbar\gamma}{4m_a} \frac{1 + (2\Delta/\gamma)^2}{(-2\Delta/\gamma)}. \quad (17.108)$$

This is an expression for the equilibrium kinetic energy, which we can convert to a temperature via

$$\frac{1}{2} m_a \langle v^2 \rangle = \frac{3}{2} k_B T, \quad (17.109)$$

where k_B is the Boltzmann constant. This gives

$$k_B T = \frac{\hbar\gamma}{4} \frac{1 + (2\Delta/\gamma)^2}{(-2\Delta/\gamma)}. \quad (17.110)$$

The temperature is minimized for the detuning $\Delta = -\gamma/2$, giving the **Doppler temperature** T_D :

$$k_B T_D = \frac{\hbar\gamma}{2}. \quad (17.111)$$

This temperature is the best expected for Doppler cooling. For ^{133}Cs at 852 nm, $T_D = 125 \mu\text{K}$, and for ^{87}Rb at 780 nm, $T_D = 146 \mu\text{K}$. These temperatures are extremely low. We can compare these temperatures to the **recoil temperature** T_r , which is the temperature corresponding to atoms with an average momentum of one photon recoil $\hbar k_0$ (i.e., a one-dimensional rms momentum of one photon recoil):

$$k_B T_r = \frac{(\hbar k_0)^2}{m_a}. \quad (17.112)$$

For ^{133}Cs , $T_r = 198 \text{ nK}$, and for ^{87}Rb , $T_r = 362 \text{ nK}$, so the Doppler limit is $T_D = 631 T_r$ for ^{133}Cs and $T_D = 403 T_r$ for ^{87}Rb . Since the (one-dimensional) rms velocity is

$$v_{\text{rms}} = \sqrt{\frac{T_D}{T_r}} \left(\frac{\hbar k_0}{m_a} \right), \quad (17.113)$$

which is 8.8 cm/s for ^{133}Cs and 12 cm/s for ^{87}Rb . These velocities are about three orders of magnitude slower than room-temperature rms velocities.

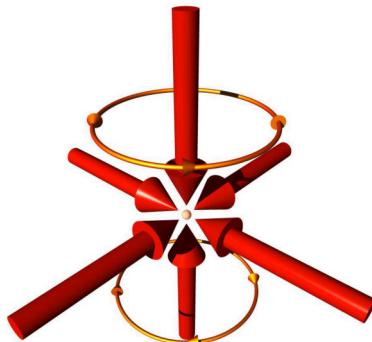
It turns out that for alkali vapors, typical laser-cooled samples exhibit temperatures well below the Doppler limit. Such “sub-Doppler” cooling is due to the degenerate level structure of alkali atoms.¹⁰ For example, ^{133}Cs can be laser cooled with the same general setup described above to about $2.5 \mu\text{K}$.¹¹

¹⁰Sub-Doppler temperature were observed in some of the first laser cooling experiments. The first reported observation is P. D. Lett, W. D. Phillips, S. L. Rolston, C. E. Tanner, R. N. Watts, and C. I. Westbrook, “Optical molasses,” *Journal of the Optical Society of America B* **6**, 2084 (1989). A classic treatment of sub-Doppler cooling mechanisms is J. Dalibard and C. Cohen-Tannoudji, “Laser cooling below the Doppler limit by polarization gradients: simple theoretical models,” *Journal of the Optical Society of America B* **6**, 2023 (1989).

¹¹C. Salomon, J. Dalibard, W. D. Phillips, A. Clairon, and S. Guellati, “Laser Cooling of Cesium Atoms below 3 μK ,” *Europhysics Letters* **12**, 683 (1990).

17.3.3.2 Magneto-Optical Trap

Optical molasses tends to stop atoms, making them “stuck,” but it does not confine atoms to a particular place. A slight modification to the three-dimensional optical molasses is to impose the magnetic field due to two opposed current loops in the “anti-Helmholtz” configuration. This arrangement is called the **magneto-optical trap** (MOT).¹²



The magnetic field vanishes at the center point of the trap, thus defining a point for atoms to accumulate. We will not go into the operation of the trap in detail, but essentially the idea is very similar to laser cooling. The additional complication is that the laser beams must all be correctly (circularly) polarized to address magnetic substates in the degenerate excited level. The magnetic field gives a position-dependent “Zeeman” shift of the transition frequency, such that if the atom is away from the center of the trap, the appropriate beam comes into resonance and pushes it towards the trap center.¹³

¹²Graphics by Windell Oskay.

¹³Jean Dalibard proposed the idea for the magneto-optical trap, and the MOT was first demonstrated by E. L. Raab, M. Prentiss, Alex Cable, Steven Chu, and D. E. Pritchard, “Trapping of Neutral Sodium Atoms with Radiation Pressure,” *Physical Review Letters* **59**, 2631 (1987).

17.4 Exercises

Problem 17.1

What are the two types of optical forces on atoms? How are they related to the atomic polarizability?

Index

ABCD law, 97–102
cascading rule, 99
deeper meaning, 100
ABCD matrix, 24, 98
factorization of general, 99–100
Q factor, 118
Q switch, 150
Q-switching, 305–306
f-number, 102

Abbé v-constant, 47
aberration theory, 24
absorption, 289, 294
absorption image, 339
absorption oscillator strength, 334–335
acousto-optic diffraction, 253–269
Bragg regime, 259–268
Klein–Cook parameter, 265
Raman–Nath regime, 254–259, 266, 268–269
rise time, 268–269
 TeO_2 modulator, 264–266

acousto-optic modulator
as spectrum analyzer, 268–269

action functional, 46

adiabatic approximation, 310

air-glass interface
cool movie, 157

Airy disk, 246

amplified spontaneous emission, 284

anamorphic prism pair, 43

antireflection (AR) coating, 197
single layer, 178–179, 198
two layer, 179–181, 198

approximations
silly, 249

associated Laguerre polynomials, 105

atom optics, 336–349
dipole force, 336–339
Doppler temperature, 346–347
magneto-optical trap (MOT), 347–348
optical molasses, 344–347
radiation pressure, 339–347

attenuation index, 162

autocorrelation function, 272

Babinet’s principle, 247–248
ball lens, 48
beam radius, 92
beam splitter, 82, 84
beam waist parameter, 92
Bessel function, 246
Bessel functions, 255
generating function for, 255
big mess, 93
birefringence, 144–147
BK7 borosilicate crown glass, 47
bleaching, 306
Boltzmann statistics, 291
boundary conditions, 151–152
Bragg reflector, 181
Brewster’s angle, 155
bug
in laser, 311

calculus of variations, 20

cat
convolution of, 201–202

cat’s eye, 48

Cauchy principal value, 238, 251–252

Cauchy probability distribution, 213

cavity dumper, 129, 306–307

CeF_3 , 180, 181

central dark-ground imaging, 237, 248–249

central limit theorem, 205–207

characteristic function, 206

characteristic polynomial, 17

coherence, 271–281
length, 278–280
time, 278–280

coherence time, 281

complex dielectric constant, 162

complex notation, 70–73

complex refractive index, 162

conductivity, 161
Drude model, 336

conductors
Drude model, 335–336
Fresnel relations for, 165–166
plasma model, 332–333

- wave propagation in, 161–166
- confocal parameter, 96
- confocal resonator, 33
- convolution, 199–214
- definition, 200
 - of box functions, 201–202
 - of two Gaussians, 203–204
 - with delta function, 200
- convolution theorem, 202–203, 246
- corner-cube reflector, 43
- cross section
- absorption, 335
 - laser, 292
 - natural, on resonance, 292, 293
- Dawson’s integral, 249–250
- degree of coherence
- first-order temporal, 274
- delta function, 61–63
- derivative of, 251
- determinant, 17
- dielectric constant, 70
- complex, 162
- dielectric medium
- simple, 69–70
- diffraction
- acousto-optic, 253–269
 - Babinet’s principle, 247–248
 - Fraunhofer, 221–228, 246–248
 - Fresnel, 227–231, 248
- diffraction grating, 225–227, 246–247
- diffraction limit, 221
- diffusion coefficient, 207
- diffusion process, 207
- diode laser, 167
- diopter, 46
- dipole force, 336–339
- directional filter, 232
- dispersion, 47, 317–330
- anomalous, 323
 - fast light, 324–325, 327–328
 - material, for optical fiber, 329–330
 - normal, 323
 - slow light, 324–327
- Doppler broadening, 292, 296, 310
- Doppler temperature, 346–347
- Drude model, 335–336
- eigenvalues, 17
- eigenvectors, 17
- Einstein A and B coefficients, 290–291
- Einstein rate equations, 290–293
- EIT, 326
- electric displacement, 69, 318
- electric flux density, 69
- electric permittivity, 70
- electro-optic effect, 144
- electromagnetic field
- energy density, 78
- electromagnetically induced transparency, 326
- energy density, 78
- energy spectral density, 272
- error analysis, 204–205
- Euler–Lagrange equation, 46
- evanescent field, 157
- extinction coefficient, 162
- extraordinary wave, 146
- F2 flint glass, 47
- Fabry–Perot cavity, 113–134, 171, 210–211
- Q factor, 118
 - confocal, 128, 130–131
 - convolution and, 210–211
 - finesse, 115
 - free spectral range, 114
 - Gaussian modes, 124–127
 - glass plate etalon, 173–174
 - Hermite–Gaussian modes, 127–128
 - line width, 116
 - photon lifetime, 117
 - physical mode condition, 125
 - resonance condition, 113
 - resonators in ray optics, 29–34
 - roughness limit to finesse, 131–132
 - survival probability, 117
- Fabry–Perot etalon, 113
- Fabry–Perot interferometer, 113
- far-off resonance trap (FORT), 339
- Faraday effect, 147, 148
- Verdet constant, 148
- Fermat’s Principle, 19–23
- Fermat’s principle, 45
- fiend, diabolical, 268
- film reflection, 171–178
- matrix formalism, 174–178
 - reflection-summation model, 171–174
- finesse, 115, 174
- fish
- underwater, 43
- fluctuation–dissipation theorem, 346
- Fourier optics, 217–245
- central point, 218
 - decomposition into plane waves, 218–219
 - diffraction limit, 221
 - holography, 242–245
 - nonparaxial regime, 221

- paraxial regime, 220–221
recipe, 219–220
spatial filters, 231–242
Fourier series, 55–58
 accuracy of, 65
 of rectified sine wave, 57–58
Fourier transform, 58–61, 71
 computed by lens, 60
 multiple dimensions, 217
 of Gaussian, 59–60
 spatial convention, 202
 thin lens as analog computer, 224–225
Fraunhofer diffraction, 221–228, 246–248
 Airy disk, 246
 circular aperture, 246
 diffraction grating, 225–227, 246–247
 double slit, 225, 246
 rectangular aperture, 246
 thin lens, 224–225
 validity conditions, 222–224
Fraunhofer lines, 47
free spectral range, 114
Fresnel diffraction, 227–231, 248
 knife edge, 248
 single slit, 228–231
Fresnel relations, 151–167
Fresnel rhomb, 159
frustrated total internal reflection, 158
FTIR spectroscopy, 274–275
function, 15
functions
 orthogonal, 63
gain coefficient, 293–296
gain saturation, 287–288, 296
Gaussian beam, 233–235, 246, 248
 knife-edge measurement (10-90 rule), 106
Gaussian beams, 92–102
 $ABCD$ law, 97–102
 focusing by thin lens, 100–101
 minimum spot size for focusing lens, 101–102
 parameter specification of, 96
 vector, 96–97
geometrical optics, 19–34
Gibbs phenomenon, 64
Gouy phase, 94–95, 98
grating, diffraction, 225–227, 246–247
gravimeter, interferometric, 85
Green function, 207–210, 213–214
group index, 322–323
group velocity, 321–323
group velocity dispersion, 324
half-wave plate, 140, 141
Hamilton's principle, 46
Hamiltonian mechanics, 45
harmonic oscillator, 331–336
 damped, 207–210, 213, 333–336
Heaviside step function, 65, 237–238
 Fourier transform, 237–238
Helmholtz equation, 73
Hermite polynomials, 103, 111
 generating function, 111
Hermite–Gauss function, 214–215
Hermite–Gaussian beam, 102–105
high reflection (HR) coating, 181
high-pass filter, 232
Hilbert transform, 239, 249–250, 252, 318
holography, 242–245, 252
 off-axis, 244–245
homogenous medium, 20
impedance, 75–76
 of vacuum, 76
impulse-response function, 207–210
index of refraction, 70
induction heater, 163
intensity, 68
interface, air-conductor
 cool movie, 166
interface, air-glass
 cool movie, 154
interference
 between partially coherent sources, 280
 coherence, 271–281
 constructive, 82
 destructive, 82
 multiple waves, 87–88
 two beams, 81–82
 two tilted plane waves, 86–87
 visibility, 277–278
interferometer
 Mach-Zehnder, 82–84, 89
 Michelson, 84–86, 89, 274–275
internal reflection, 156–159
 evanescent field, 157
 skin depth, 157
Jones matrix, 139–144
 cascaded system, 142
 linear polarizer, 140
 normal modes, 144
 polarization rotator, 142
 realistic polarizer, 149
 rotation of, 142–144
 wave retarder, 140–142

- Jones vector, 138–139
 normalization, 139
 orthogonality, 139
- Kerr lens, 308
 Klein–Cook parameter, 265
 Kramers–Kronig relations, 317–321, 329
- Lagrangian, 46
 Laguerre–Gaussian beam, 104–105
 laser
 Q switch, 150
 diode, 167
 He–Ne, 304
 Nd:YAG, 304, 311–313
 Ti:sapphire, 308
 laser spiking, 304–305
 lasers, 283–316
 Q-switching, 305–306
 cavity dumper, 306–307
 continuous-wave, 302–304
 four-level, 298–299
 gain, 286, 301–302
 gain coefficient, 293–296, 299–302
 gain media, 283–286
 gain saturation, 287–288, 296, 300–301
 gain, small-signal, 287
 key components of, 283
 laser spiking, 304–305
 light–atom interactions, 288–293
 mode locking, 307–308
 multi-mode operation, 296
 optimum output, 303
 optimum output coupler, 288
 properties of, 283
 pulsed, 304–308
 pump saturation, 301
 pumping schemes, 296–299
 pumps, 283
 resonator, 286
 single-mode operation, 295–296
 steady-state behavior, 288
 three-level, 297–298, 310
 threshold behavior, 286–288, 295
- Law of Malus, 149
 Legendre transformation, 46
 Lensmaker’s formula, 29
 light–atom interactions, 288–293
 LIGO, 85
 line shape, 291
 natural, 292
 linear algebra, 15–17
 linear transformation, 16, 18, 24
- Liouville’s theorem, 46
 Lord Rayleigh, 108
 Lorentz model, 333–336
 damping coefficient, 342
 Lorentzian
 absorption, 320–321, 325, 334
 line shape, 292
 loss probability, 303
 low-pass filter, 232
- Mach-Zehnder interferometer, 82–84, 89
 magnetic flux density, 69
 magnetization density, 69
 magneto-optical trap (MOT), 347–348
 Malus, Law of, 149
 matrix
 characteristic polynomial, 17–18
 definition, 15
 determinant, 17–18
 eigenvalues, 17–18
 eigenvectors, 17
 identity, 16
 inverse, 16
 linear transformation as, 18
 product, 16
 sum, 16
 trace, 18
 transpose, 16
 Maxwell equations, 77
 in dielectric media, 68–72
 in vacuum, 67–68
 melting coin
 cool movie, 164
 MgF₂, 179
 Michelson interferometer, 84–86, 89, 274–275
 Michelson–Morley experiment, 85
 mirror
 elliptical, 22
 parabolic, 23, 44
 plane, 20
 spherical, 22
 mode, 103
 mode locking, 307–308
 monochromatic waves, 70–72
 multiple refraction, 145
- natural line width, 293
- optical activity, 147–148
 optical axis, 23
 optical isolator, 149
 optical molasses, 344–347
 optical path length, 19

- optical power (of lens), 46
optical spectrum analyzer, 113, 122–124
 confocal, 128
ordinary wave, 146
orthogonal functions, 63
oscillator strength, 334–335
- paraxial approximation, 23, 24
paraxial wave equation, 91–92
Parseval's Theorem, 281
Parseval's theorem, 279
permeability, 67
permittivity, 67, 318
permutation symbol, 17
phase velocity, 321
photon lifetime, 309
photon scattering rate, 342–344
photonic crystal, 181
pinhole camera, 108
planar resonator, 33
Planck blackbody distribution, 291
plane wave, 77
plane waves, 73–76
plasma frequency, 333
plasma model, 332–333
Pockels cell, 150
Poisson equation
 Green function, 214
Poisson sum rule, 65, 307
polarizability, 332
polarization, 75, 77, 135–150
 birefringence, 144–147
 circular, 137–138
 cool movie, 142
 electro-optic effect, 144
 ellipse, 135–136
 elliptical, 138
 Faraday effect, 147, 148
 Jones matrix, 139–144
 Jones vector, 138–139
 linear, 136–137
 optical activity, 147–148
 S and P, 153
polarization (of dielectric), 70
polarization density, 69, 317, 332
polarizer (linear), 140
polarizing angle, 155
power spectral density, 273
 one- vs. two-sided, 275–277
power-equivalent width, 278, 281
Poynting vector, 68, 78
pulse spread
 in optical fiber, 329–330
- quantum efficiency, 304
quantum harmonic oscillator, 103
quantum Zeno effect, 149
quarter-wave plate, 140
quarter-wave stack, 181
- radiation pressure, 339–347
radiation reaction, 342
radiation, dipole, 339–342
 angular distribution, 341–342
random walk, 206–207, 346
rate equations
 Einstein's, 290–293
ray matrix, 24
ray optics, 19–34
ray tracing, 34–42
ray-transfer matrix, 24
Rayleigh length, 92
recoil temperature, 347
recoil velocity, 344
reflectance, 155
refractive index
 complex, 162, 319–321
 fused silica, 330
refractive interface, 21–22
resonator
 confocal, 34
 confocal-planar, 33
 planar, 33
 spherical, 33
resonators, 29–34
 equivalent waveguide, 29
 periodic rays, 31–32
 stability condition, 30–34
Rodrigues formula, 111
rotation matrix, 143
- sample mean, 207
saturable absorber, 306, 308
saturation intensity, 287
 for four-level laser, 300
 for two-level atom, 338
scattering rate, photon, 342–344
schlieren imaging, 237–240, 249
Schrödinger equation, 92
Sellmeier formula
 for fused silica, 330
set, 15
 cartesian product, 15
signum function, 238
skin depth, 157, 163, 168
skin effect, 163
small-signal gain coefficient, 287

- Snell's Law, 21–22, 27
 spatial filter, 231–242
 4-*f* imaging system, 231, 233, 237, 240
 4-*f* setup, 250–252
 central dark-ground imaging, 237, 248–249
 cleaning Gaussian beam, 233–235, 248
 directional, 232
 high-pass, 232
 low-pass, 232
 phase-object visualization, 235–242
 schlieren imaging, 237–240, 249
 Zernike phase-contrast imaging, 236–237
 speed of light, 19, 68
 spherical wave, 106
 spontaneous emission, 289, 294
 standard deviation, 204, 205, 207
 of mean, 207
 step function, 237–238
 Fourier transform, 237–238
 stimulated emission, 289, 293–294
 Stokes relations, 83–84
 applied to beam splitter, 84
 survival probability, 303
 susceptibility, 317, 332

 telescope, 47
 TEM (transverse electromagnetic) wave, 97
 TEM_{*l,m*} mode, 103
 thin-film reflection, 171–178
 matrix formalism, 174–178
 reflection-summation model, 171–174
 thin-lens law, 45
 Thomas–Reiche–Kuhn sum rule, 335
 total internal reflection, 36, 156–159, 246
 frustrated, 158
 transfer matrix, 24
 transmittance, 155

 uncertainty principle, 94, 96
 uncertainty relation, 278, 280, 281

 variational principle, 20
 vector, 16
 vector space, 15–16
 Verdet constant, 148
 visibility, 277–278
 Voigt profile, 292

 wave equation, 68, 77, 161–162, 168
 paraxial, 91–92
 wave fronts, 73
 wave meter, 85
 wave plate, 140
 multiple-order, 145
 zero-order, 145
 Wiener–Khinchin theorem, 272–277
 optical, 273–277
 Wigner transform, 100

 Young double slit, 89, 225

 Zernike phase-contrast imaging, 236–237
 negative, 237
 positive, 236
 ZrO₂, 180, 181

ISBN 000-00-00000-00-0

A standard linear barcode representing the ISBN number 000-00-00000-00-0.

0 000000 000000