Basic Principles of Imaging Science II

Roger Easton

12 February 2005
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology
54 Lomb Memorial Drive
Rochester, NY 14623-5604
585-475-5969
585-475-5988 (FAX)
easton@cis.rit.edu
## Contents

**Preface** ix  
**1 Review for Optics** 1  
1.1 Optics: Introduction and Review 1  
**2 Review: Oscillations** 3  
2.1 Harmonic Oscillations 4  
2.2 Harmonic Oscillations – Energy Considerations 5  
2.2.1 Anharmonic Oscillations 6  
2.3 Representations of Harmonic Oscillations 7  
**3 Review: Complex Numbers** 9  
3.1 Graphical Representation of Complex Numbers 11  
3.2 Euler Relation – Complex Exponentials 12  
3.3 Arithmetic of Complex Exponentials 15  
3.3.1 De Moivre’s Theorem 15  
3.4 Description of Harmonic Oscillations via the Euler relation 16  
3.5 Oscillations as Projections of Circular Harmonic Motion 17  
3.6 Phasor Notation for Oscillations 18  
3.7 Superposition of Oscillations 19  
3.7.1 Digression: Nonlinear Optics and Second-Harmonic Generation: 19  
3.8 Superposition of Same-Frequency Oscillations 20  
3.8.1 Trigonometric Notation 20  
3.8.2 Phasor Representation 22  
3.8.3 Complex Notation 24  
3.9 Superposition of Many Same-Frequency Oscillators 24  
3.10 Superposition of Randomly Phased Oscillators 26  
3.11 Superposition of Nonrandomly Phased Oscillators 27  
3.12 Superposition of Oscillations with Different Frequencies 29  
3.12.1 Complex Notation 29  
3.13 Introduction to Fourier Analysis 32  
**4 Review: Traveling Waves** 35  
4.1 Introduction 35  
4.1.1 2-D Plot of 1-D Traveling Wave 37
4.2 Notation and Dimensions for Waves in a Medium 37
4.3 Velocity of Traveling Waves 38
4.4 Superposition of Traveling Waves 39
4.5 Standing Waves 40
4.6 Anharmonic Traveling Waves, Dispersion 41
4.7 Average Velocity and Modulation (Group) Velocity 41
  4.7.1 Example: Nondispersive Waves $(v_\phi)_1 = (v_\phi)_2$ 42
4.8 Dispersion Relation for Nondispersive Traveling Waves 43
4.9 Dispersive Traveling Waves 43
  4.9.1 Example: Dispersive Traveling Waves 44
  4.9.2 Propagation of Superposition of Waves in Nondispersive and Dispersive Media 46
  4.9.3 Energy and Information Transmission in Nondispersive and Dispersive Media 48

5 Review: Doppler Effect 51
  5.1 Transition from Acoustic Waves to Electromagnetic Waves 51
    5.1.1 Acoustic Doppler Effect, Source at Rest 51
    5.1.2 Acoustic Doppler Effect — Source in Motion 52
    5.1.3 Acoustic Doppler Effect — Both Source and Receiver Moving 53
  5.2 Doppler Effect for Light 54
    5.2.1 Difference between Light and Sound 54

6 Maxwell’s Equations for Electromagnetic Waves 57
  6.1 Vector Operations 57
    6.1.1 Scalar Product of Two Vectors 60
    6.1.2 Cross Product 62
    6.1.3 Triple Vector Product 63
  6.2 Vector Calculus 64
  6.3 Gradient 67
  6.4 Divergence 68
    6.4.1 Gauss’ Theorem for Divergence 70
  6.5 Curl 72
    6.5.1 Example of Function with Large Curl 73
  6.6 Laplacian 74
    6.6.1 Curl of Curl 75
  6.7 Electric and Magnetic Fields 76
    6.7.1 A Note on Units 76
    6.7.2 Magnetic Fields 77
  6.8 Maxwell’s Equations 79
    6.8.1 Gauss’ Law for Electric Fields 79
    6.8.2 Gauss’ Law for Magnetic Fields 80
    6.8.3 Faraday’s Law of Magnetic Induction 80
    6.8.4 Ampere’s Law 81
    6.8.5 Maxwell’s Equations 81
6.9 Wave Equation ......................................................... 82
  6.9.1 Electromagnetic Waves from Maxwell’s Equations ............. 85
  6.9.2 Poynting Vector ................................................... 88
  6.9.3 Redux on Phase Velocity of Electromagnetic Waves .......... 89
6.10 Consequences of Maxwell’s Equations ............................. 92
6.11 Dispersion Redux .................................................... 93
  6.11.1 Feynman’s Model for Refractive Index ....................... 95
6.12 Dual Nature of Light: Photons .................................... 107
  6.12.1 Momentum of Photons .......................................... 108
6.13 Optical Frequencies – Detector Response ......................... 109

7 Propagation of Light Waves ......................................... 111
  7.1 Wavefronts ......................................................... 111
    7.1.1 Plane Waves ................................................... 111
    7.1.2 Cylindrical Waves .............................................. 112
    7.1.3 Spherical Waves ................................................. 113
  7.2 Huygens’ Principle ................................................ 114

8 Interaction of Light and Matter ..................................... 115
  8.1 Electromagnetic Waves at an Interface ............................ 115
    8.1.1 Snell’s Law for Reflection and Refraction ................. 117
    8.1.2 Boundary Conditions for Electric and Magnetic Fields ..... 119
    8.1.3 Transverse Electric Waves, s Polarization .................. 122
    8.1.4 Transverse Magnetic Waves (p polarization) ............... 126
    8.1.5 Comparison of Coefficients for TE and TM Waves .......... 127
    8.1.6 Angular Dependence of Reflection and Transmittance at “Rare-
to-Dense” Interface .................................................. 129
    8.1.7 Reflection and Transmittance at “Dense-to-Rare” Interface, Crit-
tical Angle ........................................................... 132
    8.1.8 Practical Applications for Fresnel’s Equations ............. 133
  8.2 Index of Refraction of Glass ....................................... 133
    8.2.1 Optical Path Length ............................................ 136
  8.3 Polarization .......................................................... 136
    8.3.1 Plane Polarization = Linear Polarization .................... 136
    8.3.2 Circular Polarization ........................................... 137
    8.3.3 Nomenclature for Circular Polarization ...................... 138
    8.3.4 Elliptical Polarization, Reflections .......................... 139
    8.3.5 Change of Handedness on Reflection .......................... 139
  8.4 Description of Polarization States ................................ 140
    8.4.1 Jones Vector ..................................................... 140
  8.5 Generation of Polarized Light .................................... 141
    8.5.1 Selective Emission: ............................................ 141
    8.5.2 Selective Transmission or Absorption ........................ 141
    8.5.3 Generating Polarized Light by Reflection – Brewster’s Angle . 142
    8.5.4 Polarization by Scattering ..................................... 143
CONTENTS

8.6 Birefringence – Double Refraction ........................................... 144
  8.6.1 Examples: ............................................................... 144
  8.6.2 Phase Delays in Birefringent Materials – Wave Plates ............ 145
  8.6.3 Circular Polarizer: .................................................... 146
8.7 Critical Angle – Total Internal Reflection .................................. 147

9 Optical Imaging ............................................................................. 149
  9.1 Transition from Wave Optics to Ray Optics ................................ 149
    9.1.1 Notational Conventions .............................................. 149
    9.1.2 Fermat’s Principle .................................................... 151

10 Image Formation in the Ray Model .............................................. 155
  10.1 Refraction at a Spherical Surface ........................................ 155
    10.1.1 Paraxial Approximation ............................................ 158
    10.1.2 Nature of Objects and Images: .................................... 159
  10.2 Imaging With Lenses ........................................................ 161
    10.2.1 Examples: ............................................................... 162
  10.3 Magnifications ..................................................................... 163
    10.3.1 Transverse Magnification: ........................................... 163
    10.3.2 Longitudinal Magnification: ........................................ 164
  10.4 Spherical Mirrors .................................................................. 166
  10.5 Systems of Thin Lenses ....................................................... 169
    10.5.1 Back Focal Distance .................................................. 171
    10.5.2 Front Focal Distance .................................................. 172
    10.5.3 Thin Lenses in Contact ............................................... 172
    10.5.4 “Effective Focal Length” of a System with Two Lenses .......... 173
    10.5.5 Positive Lenses Separated by $t < f_1 + f_2$ ....................... 174
    10.5.6 Cardinal Points ....................................................... 183
    10.5.7 Positive Thin Lenses Separated by $t = f_1 + f_2$ ................. 184
    10.5.8 Positive Thin Lenses Separated by $t = f_1$ or $t = f_2$ .......... 185
    10.5.9 Positive Thin Lenses Separated by $t > f_1 + f_2$ ............... 186
    10.5.10 Systems of Two Positive Thin Lenses with Different Focal Lengths 187
    10.5.11 Newtonian Form of Imaging Equation ............................. 188
    10.5.12 Another System of Two Thin Lenses: the Telephoto Lens ...... 189
  10.6 Stops and Pupils ................................................................... 193
    10.6.1 Stop and Pupils of Galilean and Keplerian Telescopes .......... 194
    10.6.2 System $f$-Number .................................................... 196
  10.7 Ray Tracing .......................................................................... 197
    10.7.1 Marginal and Chief Rays ............................................. 197
  10.8 Paraxial Ray Tracing Equations ............................................. 198
    10.8.1 Paraxial Refraction .................................................... 199
    10.8.2 Paraxial Transfer ....................................................... 200
    10.8.3 Linearity of the Refraction and Transfer Equations .............. 201
    10.8.4 Paraxial Ray Tracing .................................................. 201
    10.8.5 Matrix Formulation of Paraxial Ray Tracing ...................... 204
CONTENTS

10.8.6 Examples of System Matrices: 213

11 Waves and Imaging 217
11.1 Interference of Waves 217
11.1.1 Superposition of Two Plane Waves of the Same Frequency 222
11.1.2 Superposition of Two Plane Waves with Different Frequencies 228
11.1.3 Fringe Visibility – Coherence 230
11.1.4 Coherence Time and Coherence Length 236
11.1.5 Effect of Polarization of Electric Field on Fringe Visibility 238
11.2 Interferometers 238
11.2.1 Division-of-Amplitude Interferometers 239
11.2.2 Applications of the Michelson Interferometer 243
11.2.3 Other Types of Division-of-Amplitude Interferometers 244
11.3 Diffraction 247
11.3.1 Diffraction Integrals 249
11.3.2 Fresnel Diffraction 251
11.3.3 Fresnel Diffraction Integral as a Convolution 253
11.3.4 Diffraction Integral Valid Far from Source 257

12 Basic Principles of Digital Image Processing 265
12.1 Digital Processing 267

13 Review of Sampling 269
13.1 Digitization 269
13.2 Sampling 270
13.2.1 Ideal Sampling 270
13.3 Aliasing – Whittaker-Shannon Sampling Theorem 273
13.4 Realistic Sampling – Averaging by the Detector 274

14 Review of Quantization 281
14.1 Tone-Transfer Curve 281
14.2 Quantization 282
14.3 Quantization Error ("Noise") 284
14.3.1 Example: Variance of a Sinusoid 288
14.3.2 Example: Variance of a Square Wave 288
14.3.3 Variance of "Noise" from a Gaussian Distribution 289
14.3.4 Approximations to $SNR$ 290
14.3.5 $SNR$ of Quantization 290
14.4 Quantizers with Memory – Error Diffusion 293
14.5 Image Display Systems – Digital - to - Analog Conversion 294
14.6 Image Interpolation 295
14.6.1 Ideal Interpolation 296
14.6.2 Modulation Transfer Function of Sampling 297
14.6.3 MTF of Realistic Sampling (Finite Detectors) 298
14.7 Effect of Phase Reversal on Image Quality 299
14.8 Summary of Effects of Sampling and Quantization ........... 300
14.9 Spatial Resolution ........................................ 301

15 Point Operations ............................................ 303
15.1 Point Operations on Single Images .......................... 304
15.1.1 Image Histograms ....................................... 305
15.1.2 Histograms of Typical Images ............................ 306
15.1.3 Cumulative Histogram ................................. 307
15.1.4 Histogram Modification for Image Enhancement ........... 308
15.1.5 Jones Plots ........................................... 310
15.2 Histogram Equalization (“Flattening”) ....................... 311
15.2.1 Example of Histogram Equalization – 1-D “Image” ....... 314
15.2.2 Example of Histogram Equalization – 2-D “Image” ....... 316
15.2.3 Nonlinear Nature of Histogram Equalization ............ 316
15.3 Histogram Specification .................................... 317
15.4 Application of Histograms to Tone-Transfer Correction ..... 318
15.5 Application of Histograms to Image Segmentation ......... 319
15.6 Point Operations on Multiple Images ........................ 321
15.7 Digression: Introduction to Vision and Color ............... 322
15.7.1 The Eye .............................................. 322
15.7.2 Convergence ........................................... 326
15.7.3 Eye Motions ........................................... 330
15.7.4 Eye Lens .............................................. 330
15.7.5 Spatial resolution ...................................... 332
15.8 Color Vision ................................................ 332
15.8.1 Color Matching, Metamers ............................... 333
15.8.2 XYZ Color Space ....................................... 334
15.8.3 CIE Chromaticity Diagram ............................. 336
15.8.4 Color Reproduction ..................................... 338
15.8.5 Color Spaces .......................................... 339
15.8.6 Conversion from RGB to HSL .......................... 344
15.8.7 Phenomena of Color Vision .............................. 347
15.9 Pixel Processing of Multiple Images ......................... 347
15.10 Multispectral Histograms for Feature Extraction ......... 349
15.10.1 2-D Histograms of Simple Color Image ................. 351
15.10.2 Principal Component Analysis – PCA .................... 352
15.11 Time-Sequence Images: Video ............................... 354
15.11.1 Color-Space Transformations for Video Compression ... 355
15.11.2 Multiple-Frame Averaging .............................. 358
15.11.3 Required Number of Bits for image Sums, Averages, and Differences ......................................................... 362
15.12 Image Subtraction for Change Detection .................... 363
15.13 Difference Images as Features ............................... 364
15.13.1 Number of Bits in Difference Image ..................... 364
15.14 "Mask” or “Template” Multiplication ....................... 365
CONTENTS

15.15 Image Division .................................................. 365
  15.15.1 Image Division to Correct Spatial Sensitivity ........ 365
  15.15.2 Image Division to Enhance Low-Contrast Imagery ...... 367

16 Local Operations ................................................. 371
  16.1 Window Operators – Correlation ............................ 372
    16.1.1 Examples of $3 \times 3$ Crosscorrelation Operators .. 373
  16.2 Convolution .................................................... 374
    16.2.1 Evaluating Discrete Convolutions ..................... 376
    16.2.2 Convolutions – Edges of the Image .................. 378
    16.2.3 Convolutions – Computational Intensity ............. 379
    16.2.4 Smoothing Kernels – Lowpass Filtering ............... 380
    16.2.5 Frequency Response of 1-D Averagers ................ 382
    16.2.6 2-D Averagers ........................................... 385
    16.2.7 Differencing Kernels – Highpass Filters ............ 386
    16.2.8 Frequency Responses of 2-D Derivative Operators: ... 394
    16.2.9 Laplacian Operator ..................................... 400
    16.2.10 Discrete “Sharpening” Operators .................... 402
    16.2.11 2-D Gradient ........................................... 410
    16.2.12 Pattern Matching ....................................... 413
    16.2.13 Normalization of Contrast of Detected Features .... 417
  16.3 Nonlinear Filters ............................................. 419
    16.3.1 Median Filter ........................................... 419
    16.3.2 Other Statistical Filters (Mode, Variance, Maximum, Minimum) 428
  16.4 Adaptive Operators ........................................... 429
  16.5 Convolution Revisited – Bandpass Filters ................. 430
  16.6 Implementation of Filtering .................................. 436
    16.6.1 Nonlinear and Shift-Variant Filtering ............... 436
  16.7 Neighborhood Operations on Multiple Images ............. 437
    16.7.1 Image Sequence Processing ........................... 437
    16.7.2 Spectral + Spatial Neighborhood Operations .......... 437

17 Shape-Based Operations .......................................... 439
  17.1 Pixel Connectivity ............................................ 440
  17.2 Image Labeling ............................................... 441
    17.2.1 Example: ................................................. 442
    17.2.2 Border Coding ........................................... 442
  17.3 Border Operations ............................................. 443
    17.3.1 Contraction/erosion and expansion/dilation .......... 443
  17.4 Cascaded Morphological Operations – “Opening” and “Closing” .... 445
  17.5 Applications of Morphological Operations ................... 446
    17.5.1 Noise Removal ........................................... 446
    17.5.2 Medial Axis Transform .................................. 447
  17.6 Binary Morphological Operations ............................ 449
    17.6.1 Horizontal Structuring Element ........................ 449
18 Geometric Operations

18.1 Least-Squares Solution for Transformation ............................. 454
18.2 Common Geometrical Operations ........................................ 456
18.3 Pixel Transfers ................................................................ 457
18.4 Pixel Interpolation ................................................................ 458
  18.4.1 Example: Image Rotation ................................................ 459

19 Global Operations

19.1 Relationship to Neighborhood Operations .............................. 461
19.2 Discrete Fourier Transform (DFT) ......................................... 463
19.3 1-D Fast Fourier Transform (FFT) ........................................ 464
19.4 Multidimensional DFTs ........................................................ 466
19.5 Other Global Operations ....................................................... 470
19.6 Discrete Cosine Transform (DCT) ......................................... 470
  19.6.1 Steps in Forward DCT .................................................. 478
  19.6.2 Steps in Inverse DCT ................................................... 479
19.7 Walsh-Hadamard Transform ................................................ 479
  19.7.1 Interpretation of the W-H Transform ............................... 484

20 Image Compression

20.1 References for Image Compression ........................................ 487
20.2 Image Storage, Transmission, and Compression ....................... 487
20.3 Image Compression ............................................................. 489
  20.3.1 Histograms, Information Theory, and Imaging .................. 489
  20.3.2 Information Content of Images ........................................ 495
  20.3.3 Maximizing Image Information ....................................... 498
  20.3.4 Information Content (Entropy) of Natural Images .............. 499
20.4 Lossless Compression .......................................................... 501
  20.4.1 Run-Length Encoding .................................................... 501
  20.4.2 Huffman Code .............................................................. 501
  20.4.3 Information in Correlated Images – Markov Model ............. 506
  20.4.4 “Vector” Coding (Compression) ...................................... 507
  20.4.5 Other Flavors of Huffman Coding ................................... 514
  20.4.6 Arithmetic Coding ......................................................... 514
  20.4.7 Dictionary-Based Compression ........................................ 518
20.5 Transform Coding ............................................................... 522
  20.5.1 Color-Space Transformations .......................................... 526
  20.5.2 Space-Variant Transformations ....................................... 527
  20.5.3 Block DFT Coding via DCT ............................................ 529
20.6 JPEG Image Compression of Static Images .............................. 538
  20.6.1 Example 1: “Smooth” Region ......................................... 539
  20.6.2 Example 2: “Busy” Image ............................................. 547
Preface

The first course in Basic Principles introduced the concept of an imaging chain and considered the links about the sources of radiation and the materials used to capture the image. This course considers the links that collect the radiation (optics) and process the data.

References

Optics

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

Digital Image Processing

Center for Image Processing in Education (lots of links to software and images) (http://www.evisual.org/homepage.html)
Image J software for image processing and analysis in Java, evolution of NIHImage (http://rsb.info.nih.gov/ij/)
Image2000 image processing software from NASA (http://www.ccpo.odu.edu/SEES/ozone/oz_i2k_soft.htm)
Hypercube Image Analysis Software (http://www.tec.army.mil/Hypercube/)


**Image Compression**


Chapter 1

Review for Optics

1.1 Optics: Introduction and Review

The science of optics is often divided into three classifications based on the scale of the phenomena considered.

I. Geometrical Optics (Ray Optics): Macroscopic-scale Phenomena
    considers light to be a *RAY* that travels in a straight line until it encounters an interface between media. The wavelength $\lambda$ and frequency $\nu$ of the light are assumed to be zero and infinity, respectively: $\lambda \rightarrow 0, \nu \rightarrow \infty$;
    explains *reflection* and *refraction*
    useful for designing imaging systems.
    more difficult to assess the *quality* of the resulting image

II. Physical Optics (Wave Optics): Microscopic-scale Phenomena
    considers light (electromagnetic radiation) to be a *WAVE*;
    the action of light is described by *Maxwell’s equations*;
    light is a traveling wave with wavelength $\lambda$, temporal frequency $\nu$, phase velocity $c$;
    leads to explanations of *reflection*, *refraction*, *diffraction*, *interference*, *polarization*, *dispersion*. Useful for assessing the quality of the images.

III. Quantum Optics: Atomic-scale Phenomena
    light is a *photon*, has both wave-like and particle-like characteristics;
    used to analyze the interaction of light and matter on a sub-microscopic level;
    explains the photoelectric effect, lasers.

Phenomena in the first two of these categories are most relevant to imaging; we will ignore the third.

Referenced Sources: Optics Texts:

Physics Texts:
Chapter 2

Review: Oscillations

Sources: HR §15, C§1

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

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

- e.g., position angle of a pendulum bob in a gravitational field
- position of a mass on a spring
- voltage across the capacitor in an LC circuit

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

Oscillations result from the joint presence of two forces:

1. **Inertia**: displaces the physical quantity (e.g., the position angle of the pendulum or the voltage in the LC oscillator) from its equilibrium value.

2. **Restoring (or return) force**: opposes changes in the physical quantity, acts to return it to equilibrium. The greater the deviation from equilibrium, the larger the restoring force. (acts as *negative feedback* whose restoring force increases with
Oscillations of matter can be either transverse or longitudinal (or some combination):

1. **Longitudinal oscillation**: the vectors describing the opposing forces are parallel, e.g., a mass attached to a spring, restricted to motion toward or away from the spring.

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

### 2.1 Harmonic Oscillations

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

\[
y(t) - y_0 = A_0 \cos \{\Phi[t]\} = A_0 \cos [\omega_0 t + \phi_0]
\]

\[
y(t) = y_0 + A_0 \cos(\omega t + \phi_0)
\]

- \(y\) is the “position” of the characteristic of the medium, e.g., an angle, voltage, etc.
- \(y_0\) — equilibrium value of the characteristic;
- \(A_0\) — amplitude of the oscillation, i.e., maximum displacement from equilibrium, units of \(A_0\) are the same as those of \(y\): \([A] = [y]\);
- \(\omega_0\) — angular temporal frequency of the oscillation, units are \([\omega_0] = \text{radians per second}\);
- \(\nu\) — temporal frequency of the oscillation, units are \([\nu] = \text{cycles per second} = \text{Hertz}\);
- \(\nu = \frac{\omega_0}{2\pi}\);
- \(T\) — period of the oscillation, units are \([T] = \text{seconds}\), \(T = \frac{1}{\nu} = \frac{2\pi}{\omega_0}\);
2.2 Harmonic Oscillations — Energy Considerations

Given the equation of motion of a simple system, e.g., \( y(t) = A_0 \cos(\omega_0 t + \phi_0) \), the velocity, acceleration, and force exerted by the system can be calculated by taking derivatives:

**Velocity:**
\[
\frac{dy}{dt} \equiv \dot{y} = A_0 \frac{d}{dt} \cos(\omega_0 t + \phi_0) = A_0 \cdot (-\omega_0 \sin(\omega_0 t + \phi_0)) = -A_0 \omega_0 \sin(\omega_0 t + \phi_0)
\]

**Acceleration:**
\[
\frac{d^2y}{dt^2} \equiv \ddot{y} = -A_0 \omega_0^2 \cos(\omega_0 t + \phi_0)
\]

**Inertial Force**
\[
ma = m \ddot{y} = -m \omega_0^2 \cos(\omega_0 t + \phi_0) = \omega_0^2 y(t) \equiv -ky
\]

where:
\[
k \equiv m \omega_0^2
\]
is the “Force Constant” of Restoring Force. The force equation may be transposed to:
\[
m \ddot{y} + ky = 0
\]
this is the equation of motion for the simple harmonic oscillator.

From these equations, it is easy to derive the potential and kinetic energies of the harmonic oscillator:

**Kinetic Energy** \( \mathcal{E}_k \):
\[
\mathcal{E}_k[t] = \frac{1}{2} m v^2 = \frac{m}{2} (-\omega_0 A_0 \sin(\omega_0 t + \phi_0))^2
\]

**Potential Energy** \( \mathcal{E}_p \):
\[
\mathcal{E}_p[t] = -\int_0^y F \cdot ds = -\int_0^y (-ky) dy = \frac{ky^2}{2} = \frac{m \omega_0^2 y^2}{2}
\]
Total Energy $\mathcal{E}[t]$ is the sum of kinetic and potential energies:

$$\mathcal{E}[t] = \mathcal{E}_k[t] + \mathcal{E}_p[t] = \frac{mA_0^2\omega_0^2}{2}\sin^2[\omega_0 t + \phi_0] + \frac{mA_0^2\omega_0^2}{2}\cos^2[\omega_0 t + \phi_0]$$

which is a constant over time!

$$\mathcal{E} = \frac{mA_0^2\omega_0^2}{2}$$

Observations:

1. $\mathcal{E}$ is not a function of time, i.e., the total energy is constant
2. $\mathcal{E}_k$ and $\mathcal{E}_p$ are both always greater than 0.
3. $\mathcal{E} \propto A_0^2$, the energy is proportional to the square of the amplitude
4. $\mathcal{E} \propto \omega_0^2$, the energy is proportional to the square of the frequency: higher frequency $\Rightarrow$ more energy
5. $\omega_0^2$ is the return force per unit displacement per unit mass.

### 2.2.1 Anharmonic Oscillations

Oscillations also may be anharmonic, or nonharmonic. This simply means that the characteristic of the physical system varies in a nonsinusoidal manner. For example:
2.3 REPRESENTATIONS OF HARMONIC OSCILLATIONS

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

\[ \frac{df}{dx} = \frac{df_1}{dx} + \frac{df_2}{dx} \]

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

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

2.3 Representations of Harmonic Oscillations

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

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

where the second expression arises because \( \sin \theta = \cos \left( \frac{\pi}{2} - \theta \right) \) and the last expression from the symmetry of the cosine (i.e., \( \cos [-\theta] = \cos [\theta] \)). This description of oscillations is perfectly ok – it leads to all the correct results – but it can be awkward to keep a math handbook handy to recall the necessary expressions for the cosine and/or sine of sums, differences, and/or products of angles. The notation becomes even more complicated when considering the superposition (sum) of many oscillations or waves. For example, how easy is it to find the resultant of the sum of two oscillators, \( y_1[t] + y_2[t] \), where \( y_i = A_i \sin(\omega_i t + \phi_i) \)? You can look this up to find:

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

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

Review: Complex Numbers

H §2

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

Complex numbers arise from imaginary numbers. Since there is no real number solution for $\sqrt{-1}$, the imaginary number $i$ is arbitrarily assigned as the solution, i.e., $i = \sqrt{-1} \Rightarrow i^2 = -1$

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

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

**Complex Conjugate:** The complex conjugate of a complex number $z = a + ib$ is defined as $z^* \equiv a - ib$, i.e., simply replace $i$ with $-i$ wherever it appears!

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

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

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

3. **Multiplication:**

$$z_1 \cdot z_2 = (a_1 + ib_1) \cdot (a_2 + ib_2)$$
$$= a_1a_2 + a_1(ib_2) + a_2(ib_1) + (ib_1)(ib_2)$$
$$= (a_1a_2 - b_1b_2) + i(a_1b_2 + a_2b_1)$$

Re$\{z_1z_2\} = a_1a_2 - b_1b_2$

Im$\{z_1z_2\} = a_1b_2 + a_2b_1$;

4. **Reciprocal:** (use this trick) multiply $z_2$ by 1 in this form:

$$\frac{z_2^*}{z_2} = \frac{a_2 - ib_2}{a_2 - ib_2} = 1,$$ assuming that $z_2^* \neq 0$, and thus that $z_2 \neq 0$.
to obtain the reciprocal of $z_2$:

$$\frac{1}{z_2} \cdot \frac{z_2^*}{z_2^*} = \frac{1}{a_2 + ib_2} \cdot \frac{a_2 - ib_2}{a_2 - ib_2} = \frac{a_2 - ib_2}{a_2^2 + b_2^2}$$

$$\left(\frac{a_2}{a_2^2 + b_2^2}\right) + i \left(\frac{-b_2}{a_2^2 + b_2^2}\right)$$

$$\Rightarrow \text{Re}\left\{\frac{1}{z_2}\right\} = \frac{a_2}{a_2^2 + b_2^2}$$

$$\Rightarrow \text{Im}\left\{\frac{1}{z_2}\right\} = -\frac{b_2}{a_2^2 + b_2^2}$$

The magnitude and phase of the reciprocal are:

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

$$\Phi\left\{\frac{1}{z_2}\right\} = \tan^{-1}\left[\frac{-b_2}{a_2}\right] = -	an^{-1}\left[\frac{b_2}{a_2}\right] = -\Phi\{z_2\}$$

5. Division: Apply multiplication and the reciprocal to obtain:

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

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

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

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

6. The real and imaginary parts of $z$ can be expressed in terms of $z$ and $z^*$

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

$$\text{Im}\{z\} = \frac{1}{2}(z - z^*)$$
The magnitude of \( z \) is defined as

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

### 3.1 Graphical Representation of Complex Numbers

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

For example,

\[
\begin{align*}
z_1 &= 1 + i \\
z_2 &= 2 + i \\
\implies z_3 &= z_1 + z_2 = 3 + 2i \\
\implies z_4 &= z_1 - z_2 = -1
\end{align*}
\]

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

- **magnitude:** \( A_1 = |z_1| \equiv \sqrt{z_1 \cdot z_1^*} \)
  
  \[
  = \sqrt{(a_1 + ib_1)(a_1 - ib_1)} \\
  = \sqrt{a_1^2 + b_1^2}
  \]
phase: $\phi_1 = \tan^{-1} \left( \frac{b_1}{a_1} \right) = \tan^{-1} \left[ \frac{\text{Im}\{z_1\}}{\text{Re}\{z_1\}} \right]$

(n.b., there is a subtle problem with this definition for the phase – the inverse tangent is defined on the interval $[-\frac{\pi}{2}, \frac{\pi}{2})$, i.e., the range is only $\pi$ radians, whereas $\phi$ is defined over a range of $2\pi$ radians.)

$\text{Re}\{z\} = \text{Re}\{a + ib\} = a = A_1 \cos \phi$

$\text{Im}\{z\} = \text{Im}\{a + ib\} = b = A_1 \sin \phi$

Magnitude (a real number) $\equiv |z| = \sqrt{a^2 + b^2} = \sqrt{A_1^2 \cos^2 \phi + A_1^2 \sin^2 \phi} = \sqrt{A_1^2} = A_1$

$z = \text{Re}\{z\} + i \text{Im}\{z\}$

$= A_1 \cos \phi + A_1 (i \sin \phi)$

$= A_1 (\cos \phi + i \sin \phi)$

3.2 Euler Relation — Complex Exponential

H §2 pp.19-21, Schaum’s Outline Complex Variables §1, Schaum’s Outline Optics §1

Complex numbers are very conveniently denoted as exponentials; makes multiplication easy. Represent $z$ in its polar form:
3.2 EULER RELATION – COMPLEX EXPONENTIALS

\[ z = (r, \phi) \]
\[ = r (\cos [\phi] + i \sin [\phi]) \]
\[ \equiv re^{i\phi} \]

This expression arises from the Euler relation:

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

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

\[ z \equiv r (\cos \theta + i \sin \theta) \]
\[ dz = (\cos \theta + i \sin \theta) \, dr + r (\cos \theta + i \sin \theta) \, d\theta \]
\[ = r (\cos \theta + i \sin \theta) \, dr + r (i^2 \sin \theta + \cos \theta) \, d\theta \]
\[ = z \frac{dr}{r} + i \, r (\cos \theta + i \sin \theta) \, d\theta = z \left( \frac{dr}{r} + i d\theta \right) \]
\[ \frac{dr}{r} + id\theta = \frac{dz}{z} \implies \frac{dz}{z} = \int_0^r \frac{dr}{r} + i \int_0^\theta d\theta = \log_e r + i \theta \]

\[ \log_e z = \log_e r + i \theta \implies e^{\log_e z} = e^{\log_e r + i \theta} = e^{\log_e r + e^{i \theta}} \]
\[ \implies z = [r \cos \theta, r \sin \theta] \implies z = re^{i \theta} \]

A different proof of Euler’s relation, for those who know power-series expansions:

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

\[ \cos \theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots \implies \lim_{\theta \to 0} \{\cos \theta\} = 1 \]
\[ \sin \theta = \frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots \implies \lim_{\theta \to 0} \{\sin \theta\} = 0 \]
\[ e^{i\theta} = 1 + i\theta + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \cdots \]
\[ = 1 + i\theta + \frac{i^2 \theta^2}{2!} + \frac{i^3 \theta^3}{3!} + \frac{i^4 \cdot i^2 \cdot \theta^4}{4!} + \cdots \]
\[ = 1 + i\theta - \frac{\theta^2}{2!} - \frac{i \theta^3}{3!} + \frac{\theta^4}{4!} + \frac{i \theta^5}{5!} - \cdots \]
\[ = \left( 1 - \frac{i \theta^2}{2!} + \frac{\theta^4}{4!} - \cdots \right) + i \left( \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots \right) \]
\[ = \cos \theta + i \sin \theta. \]

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

\[
\lim_{\theta \to 0} \{\cos [\theta]\} = 1 \\
\lim_{\theta \to 0} \{\sin [\theta]\} = \theta \\
\lim_{\theta \to 0} \{\tan [\theta]\} = \lim_{\theta \to 0} \left\{ \frac{\sin [\theta]}{\cos [\theta]} \right\} = \theta
\]

Graphs of these three functions are compared in the figure:

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

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

2. addition:
   \[ z_1 + z_2 = A_1 e^{i\phi_1} + A_2 e^{i\phi_2} \]

3. multiplication:
   \[ z_1 z_2 = A_1 e^{i\phi_1} A_2 e^{i\phi_2} = A_1 A_2 e^{i(\phi_1 + \phi_2)} \]

4. division:
   \[ \frac{z_1}{z_2} = \frac{A_1}{A_2} e^{i\phi_1} e^{-i\phi_2} = \frac{A_1}{A_2} e^{i(\phi_1 - \phi_2)} \]

3.3.1 De Moivre’s Theorem

Generalization of multiplication of complex exponentials:
\[ z^n = \left[r(\cos \phi + i \sin \phi)\right]^n = r^n \left[\cos \phi + i \sin \phi\right]^n = r^n \left[\cos n\phi + i \sin n\phi\right] \quad \text{Proof by induction} \]

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

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

Examples:
\[
\phi = 0 \implies e^0 = 1 \text{ because } \cos[0] = 1 \text{ and } \sin[0] = 0
\]
\[
\phi = \frac{\pi}{2} \implies e^{i\frac{\pi}{2}} = \cos \left[ \frac{\pi}{2} \right] + i \sin \left[ \frac{\pi}{2} \right] = i
\]
\[
\phi = \pi \implies e^{i\pi} = \cos[\pi] + i \sin[\pi] = -1
\]
\[
y[t] = A \cos[\omega t + \phi] = \text{Re} \left\{ A e^{i(\omega t + \phi)} \right\} = \text{Re} \left\{ z[t] \right\}
\]

As will be discussed, products and sums of same-frequency harmonic oscillations are easily computed.

### 3.4 Description of Harmonic Oscillations via the Euler relation

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

\[
z[t = 0] = Ae^{i0} = A
\]
\[
z \left[ t = \frac{\pi}{4\omega} = \frac{T}{8} \right] = Ae^{i\frac{\pi}{4}} = A \left( \cos \left[ \frac{\pi}{4} \right] + i \sin \left[ \frac{\pi}{4} \right] \right) = A \left( \frac{1}{\sqrt{2}} \right) (1 + i)
\]
\[
z \left[ t = \frac{\pi}{2\omega} = \frac{T}{4} \right] = Ae^{i\frac{\pi}{2}} = A \left( \cos \left[ \frac{\pi}{2} \right] + i \sin \left[ \frac{\pi}{2} \right] \right) = A \cdot i
\]
\[
z \left[ t = \frac{\pi}{\omega} = \frac{T}{2} \right] = Ae^{i\pi} = A \left( \cos[\pi] + i \sin[\pi] \right) = -A
\]
\[
z \left[ t = \frac{3\pi}{2\omega} = \frac{3T}{4} \right] = Ae^{i\frac{3\pi}{2}} = A \left( \cos \left[ \frac{3\pi}{2} \right] + i \sin \left[ \frac{3\pi}{2} \right] \right) = A \cdot -1 = -A
\]

As \( t \) increases, the complex function describes a circle of radius \( A \) about the origin.
3.5 Oscillations as Projections of Circular Harmonic Motion

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

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

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

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

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

### 3.5 Oscillations as Projections of Circular Harmonic Motion

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

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

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

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

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

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

$$\text{Re} \{ y [t] \} = \cos [\omega t]$$

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

$$\text{Im} \{ y [t] \} = \sin [\omega t] = \cos \left[ \omega t - \frac{\pi}{2} \right]$$
3.6 Phasor Notation for Oscillations

H §7.3

The interpretation of harmonic motion as a projection of uniform circular motion leads to a third method for representing oscillations – the phasor. Its use is quite popular in electrical engineering applications.

The phasor with magnitude $A$ and phase $\Phi [t_0]$ is denoted by the polar vector $(A, \Phi [t_0])$ that describes the instantaneous position of the oscillator on the 2-D plot (Argand diagram). As time progresses, the phasor of an oscillator rotates with period $T = \frac{1}{\nu} = \frac{2\pi}{\omega}$. Generally, the phasor picture portrays the amplitude and phase of the oscillator at a particular time $t_0$ (generally $t_0 = 0$ seconds).

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

H§7.§14

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

3.7.1 Digression: Nonlinear Optics and Second-Harmonic Generation:

To help illustrate linear media and the principle of superposition, we will first consider an example where superposition is not valid. There are situations and media which can generate a resultant that is not a linear combination of the components. This effect has developed into the field of nonlinear optics. For example, a high-energy laser focused on one of a class of crystals (such as quartz or potassium dihydrogen phosphate – KDP) which generate some emerging energy proportional to square of the sum of the incident electric field $E$:

$$E[t] \simeq (E_1 \cos [\omega_1 t] + E_2 \cos [\omega_2 t])^2$$

$$= E_1^2 \cos^2 [\omega_1 t] + E_2^2 \cos^2 [\omega_2 t] + 2E_1 E_2 \cos [\omega_1 t] \cdot \cos [\omega_2 t]$$
As we will shortly demonstrate, the first two terms on the right-hand side can also be written:

\[ E_1^2 \cos^2[\omega_1 t] = E_1^2 \left(1 + \cos[2\omega_1 t]\right) \]
\[ E_2^2 \cos^2[\omega_2 t] = E_2^2 \left(1 + \cos[2\omega_2 t]\right) \]

These individual pieces are the sums sinusoids that oscillate with angular temporal frequencies \(2\omega_1\) and \(2\omega_2\), respectively, and constant terms (cosines that “oscillate” with frequency equal to zero). The third term on the right also may be rewritten:

\[ 2E_1 E_2 \cos[\omega_1 t] \cos[\omega_2 t] = E_1 E_2 \left\{ \cos\left((\omega_1 + \omega_2) t\right) + \cos\left((\omega_1 - \omega_2) t\right) \right\}. \]

In words, the electromagnetic interactions in this nonlinear medium generates sinusoids that oscillate with frequencies that differ from the original frequencies: sinusoids with zero frequency, twice the frequencies of the component functions, and sinusoids with the sum and difference frequencies. If both input beams have the same frequency \(\omega_0\) (and thus the same wavelength \(\lambda_0\)), there will be an output component with frequency \(\omega' = 2\omega, \lambda' = \frac{\lambda}{2}\). For example, a laser rod composed of Yttrium-Aluminum-Garnet doped with Neodymium (a Nd:YAG laser) can lase to make a beam with \(\lambda = 1.06 \mu m\) (in the “near-infrared” region of the spectrum). If the laser beam has sufficient energy and is directed onto a crystal that is has a strong “nonlinear” response, an output beam may be produced at the “doubled wavelength” \(\lambda' = 0.53 \mu m\), i.e., visible green light. Such an effect is called second-harmonic generation and is a very active research area in quantum optics.

Though nonlinear effects are of great interest in optics today, we will just consider situations where the principle of superposition is valid – the output is the sum of the component terms.

### 3.8 Superposition of Same-Frequency Oscillations

#### 3.8.1 Trigonometric Notation

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

\[ y_1[t] = A_1 \cos[\omega t + \phi_1] \]
\[ y_2[t] = A_2 \cos[\omega t + \phi_2] \implies y[t] = y_1[t] + y_2[t] \]

The trigonometric solution of the resultant \(y[t]\) can be found as follows:

\[ y_1 + y_2 = A_1 (\cos[\omega t] \cos[\phi_1] - \sin[\omega t] \sin[\phi_2]) + A_2 (\cos[\omega t] \cos[\phi_2] - \sin[\omega t] \sin[\phi_2]) \]
\[ = \cos[\omega t] (A_1 \cos[\phi_1] + A_2 \cos[\phi_2]) - \sin[\omega t] (A_1 \sin[\phi_1] + A_2 \sin[\phi_2]) \]
3.8 SUPERPOSITION OF SAME-FREQUENCY OSCILLATIONS

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

\[
\text{Re}\{(\mathcal{A}, \phi)\} = \mathcal{A}\cos\phi = A_1\cos[\phi_1] + A_2\cos[\phi_2]
\]

\[
\text{Im}\{(\mathcal{A}, \phi)\} = \mathcal{A}\sin\phi = A_1\sin[\phi_1] + A_2\sin[\phi_2]
\]

The squared magnitude of the result is:

\[
(\mathcal{A}\sin\phi)^2 + (\mathcal{A}\cos\phi)^2 = \mathcal{A}^2 = (A_1\sin[\phi_1] + A_2\sin[\phi_2])^2 + (A_1\cos[\phi_1] + A_2\cos[\phi_2])^2
\]

\[
\implies \mathcal{A} = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_1 - \phi_2)},
\]

and phase:

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

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

Consider some simple cases:

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

\[
A_1\cos[\omega t + \phi_1] + A_1\cos[\omega t + \phi_1] = 2A_1\cos[\omega t + \phi_1] \implies \mathcal{A} = 2A_1, \ \phi = \phi_1
\]

\[
\mathcal{A}^2 = A_1^2 + A_1^2 + 2A_1A_1\cos(\phi_1 - \phi_1) = 2A_1^2 + 2A_1^2\cos(0) = 4A_1^2
\]

\[
A = 2A_1
\]

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

\[
\implies \phi = \phi_1
\]

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

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

\[
\mathcal{A}^2 = A_1^2 + A_1^2 + 2A_1A_1\cos[(\phi_1 - (\phi_1 - \pi)]
\]

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

\[
\implies \mathcal{A} = 0
\]

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

Addition of two oscillations with same amplitude but out of phase by \(\pm\pi\) radians gives zero output, also as expected.
3. $A_1 = A_2, \phi_2 = \phi_1 + \frac{\pi}{2} \implies$ same amplitude, phase difference of $+\frac{\pi}{2}$ radians.

The resultant has magnitude:

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

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

and phase:

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

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

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

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

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

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

which leads to the observation:

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

and the phase of the resultant is:

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

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

### 3.8.2 Phasor Representation

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

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

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

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

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

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

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

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

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

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

Consider the complex representation of two oscillators with the same frequency \( \omega \):

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

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

\( n.b. \), The resultant oscillation has the same frequency as the components frequency \( \omega \).

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

\[
\mathcal{A} = \sqrt{A_1^2 + A_2^2 + 2A_1 A_2 \cos(\phi_1 - \phi_2)} \\
\tan [\phi] = \begin{bmatrix} A_1 \sin(\phi_1) + A_2 \sin(\phi_2) \\ A_1 \cos(\phi_1) + A_2 \cos(\phi_2) \end{bmatrix}
\]

\textit{THE SUPERPOSITION OF TWO SAME-FREQUENCY OSCILLATIONS IS AN OSCILLATION OF THAT FREQUENCY}

(Fourier analysis makes this statement obvious!)

3.9 Superposition of Many Same-Frequency Oscillators

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

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

The resultant oscillation has amplitude \( \mathcal{A} \) and phase \( \Phi \), and hence may be specified by the phasor \((\mathcal{A}, \Phi)\).
3.9 SUPERPOSITION OF MANY SAME-FREQUENCY OSCILLATORS

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

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

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

By the Pythagorean theorem:

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

Look at the square of the real part:

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

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

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

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

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

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

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

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

i.e., \( j = [2, N] \), \( k = [1, N - 1] \)
The treatment for the imaginary part is identical:

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

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

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

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

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

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

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

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

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

3.10 Superposition of Randomly Phased Oscillators

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

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

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

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

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

Since \( \cos [\phi_j - \phi_k] \) is randomly distributed over the interval \([-1, 1] \), we expect that
3.11 SUPERPOSITION OF NONRANDOMLY PHASED OSCILLATORS

many terms likely will sum to zero:

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

so that the amplitude of the resultant should be.

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

$$\implies A = \left(\sqrt{N}\right) A_0$$

The phase of the sum of the random-phase oscillators cannot be predicted, but can be any angle in the interval $[-\pi, +\pi)$

Recall that the energy of the oscillator is proportional to $A^2$, so if the phases are random, the total energy is the sum of the individual energies, as expected. Note that the total amplitude is $\sqrt{N}$ times as large as the individual amplitude. Randomly phased oscillators are said to be incoherent.

Two examples of superposition of randomly phased oscillators, showing resultant magnitudes.

3.11 Superposition of Nonrandomly Phased Oscillators

Special Case II: Amplitudes AND phases are equal, i.e., $A_j = A_k = A_0$ and $[\phi_j] = [\phi_k] = \phi_0$
\[ I = A^2 = N \cdot A_0^2 + 2 \cdot A_0^2 \sum_{j=2}^{N} \cos(\phi_0 - \phi_j) \]

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

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

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

Examples:

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

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

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

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

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

If the phase relationship between the component oscillators is well-defined, the oscillators are \textit{coherent}.

\[ \begin{array}{c}
\text{Real Axis} \\
\text{Imaginary Axis}
\end{array} \]

Two examples of the sum of nonrandomly phased oscillators. In one case (shown in black), the sum yields a “null output” (resulting magnitude is 0). In the other case, the sum is nonzero.
3.12 Superposition of Oscillations with Different Frequencies

3.12.1 Complex Notation

H§7.5, HR §20

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

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

$$y [t] = y_1 [t] + y_2 [t] = A_0 \cos [\omega_1 t] + A_0 \cos [\omega_2 t]$$
$$= A_0 \left( \cos [\omega_1 t] + \cos [\omega_2 t] \right)$$
$$= \text{Re} \left\{ A_0 \left[ e^{i\omega_1 t} + e^{i\omega_2 t} \right] \right\}$$

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

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

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

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

\[ A_0 \left( \sin[\omega_1 t] + \sin[\omega_2 t] \right) = 2A_0 \cos \left( \frac{\omega_1 - \omega_2}{2} t \right) \cdot \sin \left( \frac{\omega_1 + \omega_2}{2} t \right) \]

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

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

we obtain:

\[ y[t] = 2A_0 \cos \left( \Omega_{\text{avg}} t \right) \cos \left[ \Omega_{\text{mod}} t \right] \]

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

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

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

Example:

Consider the product and sum of two harmonic oscillations with angular frequencies \( \nu_1 = \frac{1}{50} \) and \( \nu_2 = \frac{1}{60} \) cycles per unit length, so the corresponding temporal periods are \( T_1 = 50 \) and \( T_2 = 60 \). These are illustrated below:
3.12 SUPERPOSITION OF OSCILLATIONS WITH DIFFERENT FREQUENCIES

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

The sum of these two oscillations is:

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

$$= 2 \cos \left[ 2\pi \left( \frac{1}{10} + \frac{1}{60} \right) t \right] \cdot \cos \left[ 2\pi \left( \frac{1}{10} - \frac{1}{18} \right) t \right]$$

$$\simeq 2 \cos \left[ 2\pi \frac{t}{54.545} \right] \cdot \cos \left[ 2\pi \frac{t}{600} \right]$$

$$\nu_{avg} \simeq \frac{1}{54.545}, \quad \nu_{mod} = \frac{1}{600}$$

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

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

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

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

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

\[
\nu_{\text{sum}} = \frac{1}{27.27} \text{ Hz}, \ \nu_{\text{diff}} = \frac{1}{300} \text{ Hz}
\]

### 3.13 Introduction to Fourier Analysis

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

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

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

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

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

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

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

\[
f_2[t] = 0
\]

\[
f_3[t] = -\frac{1}{3} \cos \left[2\pi \frac{t}{\left(\frac{3}{5}\right)}\right]
\]

\[
f_4[t] = 0
\]

\[
f_5[t] = +\frac{1}{5} \cos \left[2\pi \frac{t}{\left(\frac{3}{5}\right)}\right]
\]

Obviously, \(y[t]\) is becoming less and less harmonic as more terms are added, and in fact is starting to look like a completely different function – a square wave. Especially note that as higher frequency components are added (i.e., larger values of \(n\)), the verticals become “steeper” and the edges become “sharper.” Note also that the summation overshoots when transitioning from horizontal to vertical and vice versa. This is known as the Gibbs phenomenon, and the visibility of this effect diminishes as more
3.13 INTRODUCTION TO FOURIER ANALYSIS

As $N \to \infty$, the function $y[t]$ becomes a periodic square wave, which is quite dissimilar from the component functions.

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

Sum of sinusoids with specific different magnitudes and frequencies to produce a square wave.

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

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

This concept should be quite familiar to you – the spectrum of white light is analogous. White light is a periodic function – it looks the same at all times. Spherical rain droplets act as prisms to disperse white light into its constituent components – the colors of the spectrum. The brightnesses of each color correspond to the energy of the component – brighter $\Rightarrow$ more energy. The droplet prisms act as Fourier transformers since they derive the spectrum of the function as Newton showed, the spectrum can be transformed back to white light with another prism.
Chapter 4

Review: Traveling Waves

4.1 Introduction

To date, we have considered oscillations, i.e., periodic, often harmonic, variations of a physical characteristic of a system. The system at one time is indistinguishable from the system observed at a later time if the time difference is an integral number of temporal periods. To maintain oscillatory behavior, the energy of the oscillator must remain within the system, i.e., there can be no losses of energy. We will now extend this picture to oscillations that travel from the source and thus transport energy away. Energy must be continually added to the system to maintain the oscillation and the transported energy can do work on other systems at a distance.

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

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

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

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

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

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

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

$$\Phi[z, t] = \alpha z + \beta t$$

$$\implies \alpha z_0 + \beta t_0 = \alpha z_1 + \beta t_1.$$

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

$$\Phi[z, t] = |\alpha| z - |\beta| t$$

By dimensional analysis, we know that $|\alpha| z - |\beta| t$ has the “dimensionless dimensions” of angle [i.e., measured in the unitless quantity of radians]. We have already identified $\beta = \omega_0$, the angular frequency of the oscillation. Similarly, if $[z] = \text{mm}$ must have dimensions of radians/mm, i.e., $\alpha$ tells how many radians of oscillation exist per millimeter (or other unit of length) – the angular spatial frequency of the wave, commonly denoted by $k$:

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

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

### 4.1.1 2-D Plot of 1-D Traveling Wave

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

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

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

### 4.2 Notation and Dimensions for Waves in a Medium

Trigonometric Notation:

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

Complex Notation:

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

\(y\) = position of the characteristic of the medium, e.g., \([y]\) = angle, voltage, ... ;
\[ y_0 = \text{equilibrium value of the characteristic}; \]
\[ A_0 = \text{amplitude of the wave, i.e., maximum displacement from equilibrium, } [A_0] = [y]; \]
\[ z, t = \text{spatial and temporal coordinates}, [z] = \text{length (e.g., mm)}, [t] = \text{s}; \]
\[ T = \text{period of the wave}, [T] = \sec [s]; T = \frac{1}{\nu_0} = \frac{2\pi}{\omega_0}; \]
\[ \lambda_0 = \text{wavelength}, [\lambda_0] = \text{mm} \]
\[ \omega_0 = \text{angular temporal frequency of the wave}, \omega_0 = \frac{2\pi}{T}, [\omega_0] = \text{radians per s}; \]
\[ k_0 = \text{angular spatial frequency of the wave}, k_0 = \frac{2\pi}{\lambda_0}, [k_0] = \text{radians per mm}; \]
\[ \nu_0 = \text{temporal frequency of the wave}, [\nu_0] = \text{cycles per second} \left[ \frac{\text{cycles}}{s} \right] = \text{Hz}, \]
\[ \nu_0 = \frac{\omega_0}{2\pi}; \]
\[ \Phi = \text{phase angle of the wave}, [\Phi] = \text{radians}, (\text{in this case, } \Phi \text{ is linear in time and space}); \]
\[ \phi_0 = \text{initial phase of the wave, i.e., phase angle at } t = 0 \text{ and } z = 0, [\phi_0] = \text{radians}. \]
\[ \sigma_0 = \text{wavenumber}, \sigma_0 = \frac{1}{\lambda_0}, \text{number of wavelengths per unit length}, [\sigma_0] = \text{mm}^{-1}. \]

Relations between the phase and the temporal frequencies

\[ \omega_0 = -\frac{\partial \Phi}{\partial t} \]
\[ \nu_0 = \frac{\omega_0}{2\pi} = -\frac{1}{2\pi} \frac{\partial \Phi}{\partial t} \]

### 4.3 Velocity of Traveling Waves

The phase velocity \( v_\phi \) of a wave is the speed of travel of a point of constant phase. A definition for phase velocity can be derived by dimensional analysis: \([v_\phi] = \text{mm per s}; [\omega_0] = \text{radians per s}; [k] = \text{radians per mm}:\]

\[
\mathbf{\longrightarrow} \begin{bmatrix} \omega_0 \\ k_0 \end{bmatrix} = \begin{bmatrix} \text{radians per second} \\ \text{radians per mm} \end{bmatrix} = \begin{bmatrix} \text{radian-mm} \\ \text{radian-s} \end{bmatrix} = \begin{bmatrix} \text{mm} \\ \text{s} \end{bmatrix}
\]

Slightly more rigorously, we can find the phase velocity of a wave by taking derivatives of the equation for the wave:

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

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

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

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

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

\[ z_1 = b' + \left( \frac{\omega_0}{k_0} \right) t_1 \]
\[ z_2 = b' + \left( \frac{\omega_0}{k_0} \right) t_2 \]

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

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

4.4 Superposition of Traveling Waves

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

\[ y_1 [z, t] = A_0 \cos [k_1 z - \omega_1 t] \]
\[ y_2 [z, t] = A_0 \cos [k_2 z - \omega_2 t]. \]

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

\[ \Omega_1 \equiv \frac{k_1 z}{t} - \omega_1 \]
\[ \Omega_2 \equiv \frac{k_2 z}{t} - \omega_2 \]

\[ y [z, t] = y_1 [z, t] + y_2 [z, t] = A_0 \{ \cos [k_1 z - \omega_1 t] + \cos [k_2 z - \omega_2 t] \} \]
\[ = A_0 \left\{ \cos \left[ \left( \frac{k_1 z}{t} - \omega_1 \right) t \right] + \cos \left[ \left( \frac{k_2 z}{t} - \omega_2 \right) t \right]\right\} \]
\[ = A_0 \{ \cos [\Omega_1 t] + \cos [\Omega_2 t]\} \]
\[ = 2A_0 \cos \left( \frac{\Omega_1 + \Omega_2}{2} \right) t \cos \left( \frac{\Omega_1 - \Omega_2}{2} \right) t \]

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

\[ \left( \frac{\Omega_1 + \Omega_2}{2} \right) t = \left( \frac{k_1 z}{t} - \omega_1 + \frac{k_2 z}{t} - \omega_2 \right) t = \left( \frac{k_1 + k_2}{2} \right) z - \left( \frac{\omega_1 + \omega_2}{2} \right) t \equiv k_{avg} z - \omega_{avg} t \]

where $k_{avg} \equiv \frac{k_1 + k_2}{2}$, $\omega_{avg} \equiv \frac{\omega_1 + \omega_2}{2}$.
\[
\left(\frac{\Omega_1 - \Omega_2}{2}\right)t = \left(\frac{k_1}{t} - \omega_1 - \frac{k_2}{t} + \omega_2\right)t = \frac{k_1 - k_2}{2}z - \frac{\omega_1 - \omega_2}{2}t \equiv k_{\text{mod}}z - \omega_{\text{mod}}t
\]

where \(k_{\text{mod}} = \frac{k_1 - k_2}{2}\), \(\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2}\).

### 4.5 Standing Waves

Consider the superposition of two waves with the same amplitude \(A_0\), temporal frequency \(\nu_0\), and wavelength \(\lambda_0\), but that are traveling in opposite directions:

\[
\begin{align*}
    f_1[z, t] + f_2[z, t] &= A_0 \cos [k_0 z - \omega_0 t] + A_0 \cos [k_0 z + \omega_0 t] \\
    &= 2A_0 \cos \left[\frac{k_0 z - \omega_0 t}{2} + \frac{k_0 z + \omega_0 t}{2}\right] \cdot \cos \left[\frac{k_0 z - \omega_0 t}{2} - \frac{k_0 z + \omega_0 t}{2}\right] \\
    &= 2A_0 \cos \left[\frac{k_0 z + k_0 z}{2} + \frac{-\omega_0 t + \omega_0 t}{2}\right] \cdot \cos \left[\frac{k_0 z - k_0 z}{2} + \frac{-\omega_0 t - \omega_0 t}{2}\right] \\
    &= 2A_0 \cos [k_0 z] \cdot \cos [-\omega_0 t] \\
    &= 2A_0 \cos [k_0 z] \cdot \cos [\omega_0 t], \quad \text{because} \quad \cos [-\theta] = + \cos [+\theta] \\
    &= 2A_0 \cos \left[2\pi \frac{z}{\lambda_0}\right] \cdot \cos [2\pi \nu_0 t]
\end{align*}
\]

This is the product of a spatial wave with wavelength \(\lambda_0\) and a temporal oscillation with frequency \(\nu_0\).

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

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

\[ y(z,t) = \sum_{n=1}^{\infty} y_n = \sum_{n=1}^{\infty} A_n \cos (k_n z - \omega_n t + \phi_n), \]

where \( A_n, k_n, \) and \( \omega_n \) are the amplitude, angular spatial frequency, and angular spatial frequency of the \( n^{th} \) wave. Therefore, we can define the phase velocity of the \( n^{th} \) wave as:

\[ (v_\phi)_n = \frac{\omega_n}{k_n}. \]

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

What if the phase velocities are different, i.e., if \((v_\phi)_1 \neq (v_\phi)_2\)? In this case, points of constant phase on the two waves will move at different velocities, and therefore the distance between points of constant phase will change as a function of position or time. Therefore the shape of the superposition wave will change as a function of time; these waves are \textit{dispersive}.

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

4.7 Average Velocity and Modulation (Group) Velocity

We added two traveling waves of different frequencies and obtained the same result we saw when adding two oscillations: the sum of two harmonic waves yields the product of two harmonic waves with modulation and average spatial \textit{and} temporal frequencies. Using the new terms: \( k_{\text{avg}}, k_{\text{mod}}, \omega_{\text{avg}}, \) and \( \omega_{\text{mod}} \), we can define the phase
velocities of the average and modulation waves:
\[ v_{\text{avg}} \equiv \frac{\omega_{\text{avg}}}{k_{\text{avg}}} = \left( \frac{\omega_1 + \omega_2}{2} \right) = \frac{\omega_1 + \omega_2}{k_1 + k_2} \]
\[ v_{\text{mod}} \equiv \frac{\omega_{\text{mod}}}{k_{\text{mod}}} = \left( \frac{\omega_1 - \omega_2}{2} \right) = \frac{\omega_1 - \omega_2}{k_1 - k_2} \]

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

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

In a nondispersive medium, the phase velocity is constant over frequency (or wavelength), i.e.,
\[ (v_\phi)_1 = \frac{\omega_1}{k_1} = (v_\phi)_2 = \frac{\omega_2}{k_2}. \]

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

\[ v_{\text{avg}} = \frac{\omega_{\text{avg}}}{k_{\text{avg}}} = \frac{\omega_1 + \omega_2}{k_1 + k_2} = \frac{\omega_1}{k_1} \left( 1 + \frac{k_2}{k_1} \right) \]

Since \(\frac{\omega_1}{k_1} = \frac{\omega_2}{k_2}\) for nondispersive waves \(\Rightarrow \frac{\omega_2}{\omega_1} = \frac{k_2}{k_1}\) and:

\[ v_{\text{avg}} = \frac{\omega_1}{k_1} \left( 1 + \frac{k_2}{k_1} \right) = \frac{\omega_1}{k_1} = v_1 = v_2 = v_{\text{avg}} \]

Similarly for the velocity of the modulation wave:

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

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

\[ v_{\text{mod}} = \frac{\omega_1}{k_1} \left( 1 - \frac{k_2}{k_1} \right) = \frac{\omega_1}{k_1} = v_1 = v_2 = v_{\text{mod}} = v_{\text{avg}} \]

Note also that \(\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\Delta \omega}{\Delta k} \Rightarrow \frac{d \omega}{dk} = \omega_{\text{mod}} \)

*In a nondispersive medium, all waves (all spatial and temporal frequencies and all*
modulation and average waves) travel at the same velocity.

4.8 Dispersion Relation for Nondispersive Traveling Waves

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

$$\omega = \omega(k) = (v_\phi) \cdot k$$

where $v_\phi$ is constant, so that $\omega \propto k$.

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

\[\text{Dispersion Relation for Nondispersive Waves, Two types of wave with different velocities } (v_\phi)_1 > (v_\phi)_2.\]

4.9 Dispersive Traveling Waves

The more general, more common, and more important case is that of dispersive waves. Here, the phase velocity $v_\phi = \frac{c}{n}$ is not constant; $v_\phi$ varies with frequency. This is the normal state of affairs for light traveling in a medium such as glass. The common specification of the phase velocity of light in medium is the refractive index $n$:

$$n = \frac{c}{v_\phi}$$
where $v_\phi$ is the phase velocity of light in the medium. In a dispersive medium, we can interpret group velocity in another way:

$$\omega(k) = k \cdot v_\phi$$

$$\implies v_{\text{mod}} = \frac{d\omega}{dk} = \frac{d}{dk} (k \cdot v_\phi)$$

$$= \left(\frac{dk}{dk}\right) v_\phi + k \cdot \left(\frac{dv_\phi}{dk}\right) = v_\phi + k \left(\frac{dv_\phi}{dk}\right).$$

In other words, the group velocity is the sum of the phase velocity $v_\phi$ and a term proportional to $\frac{dv_\phi}{dk}$, which is the change in phase velocity with wavenumber:

$$\frac{dv_\phi}{dk} > 0 \implies v_{\text{mod}} > v_\phi$$

$$\frac{dv_\phi}{dk} < 0 \implies v_{\text{mod}} < v_\phi.$$ 

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

**4.9.1 Example: Dispersive Traveling Waves**

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

$$\omega(k) = \alpha k^\ell$$

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

$$v_{\text{avg}} = \frac{\omega}{k} = \frac{\alpha(k^\ell)}{k} = \alpha k^{\ell-1}$$

$$v_{\text{mod}} = \frac{d\omega}{dk} = \frac{d}{dk} (\alpha k^\ell) = \ell (\alpha k^{\ell-1}) = \ell \cdot v_{\text{avg}}.$$ 

So if $\ell > 1$, then $v_{\text{mod}} > v_{\text{avg}}$, and if $\ell < 1$, $v_{\text{mod}} < v_{\text{avg}}$. The first relation corresponds to anomalous dispersion and the second to normal dispersion. The dispersion relation for normal dispersion is nonlinear and concave down, while that for anomalous dispersion is nonlinear and concave up. Of course, for nondispersive waves the dispersion relation is linear.
Phase and modulation (group) velocities on the dispersion plot $\omega(k)$. The phase velocity at wavenumber $k_1$ is $\frac{\omega_1}{k_1}$, while the velocity of the modulation wave is the slope of the dispersion curve evaluated at $k_1$, $v_{\text{mod}} = \frac{d\omega}{dk}|_{k=k_1}$.

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

Recall that an anharmonic, though periodic, oscillation can be expressed as a sum of harmonic terms of different frequencies, i.e., as a Fourier series. We can therefore find the effect of dispersion on an anharmonic traveling wave by decomposing it into its Fourier series of harmonic terms and propagating each separately at its own velocity. The resultant is found by resumming the resulting components. For example, if:

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

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

As we’ve already seen, \(f[z, t]\) is the sum of the first three terms of a square wave. The wave at the source is a “blurry square wave,” as shown in (a), where the three wavelengths of the three waves are respectively 4 units, \(\frac{4}{3}\) units, and \(\frac{4}{5}\) units. In the nondispersive case, \(\lambda_1 = 3\lambda_2 = 5\lambda_3\) and \(k_1 = \frac{k_2}{3} = \frac{k_3}{5}\), which means in turn that \(\frac{\omega_1}{k_1} = \frac{\omega_2}{k_2} = \frac{\omega_3}{k_3}\) and \(v_1 = v_2 = v_3\). Since all components in the waveform propagate at the same velocity, then the relative phase difference is maintained throughout and the “shape” of the wavefront doesn’t change as it propagates as shown in (b):
Propagation of a waveform in a nondispersive medium: (a) sum of three sinusoids to produce a “blurry” square wave (lowpass filtered); (b) resulting waveform after propagating all three sinusoids with the same phase velocity, so that all components travel the same distance. The waveform is undistorted.

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

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

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

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

4.9.3 Energy and Information Transmission in Nondispersive and Dispersive Media

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

\[
 f[t] = s[t] \cdot r[t] = s[t] \cdot \cos(\omega_{\text{carrier}} \cdot t) = s[t] \cdot \cos(2\pi \nu_{\text{carrier}} \cdot t)
\]

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

\[ v_{\text{average}} = \nu_{\text{average}} \cdot \lambda_{\text{average}} = \frac{\omega_{\text{carrier}}}{k_{\text{carrier}}} \]

\[ v_{\text{mod}} = \nu_{\text{mod}} \cdot \lambda_{\text{mod}} \approx \frac{d\omega}{dk} \bigg|_{k_{\text{carrier}}} \]

The information (speech or music) is carried by the modulation and travels through the medium at the modulation velocity, which we know to be less than the average velocity in a normally dispersive medium. An example is shown as “snapshots” of the product of a long-period sinusoidal modulation with frequency \( \nu_1 \) and period \( \lambda_1 \) (taken to be \( \frac{8}{7} \) units in this example) and a short-period carrier wave with frequency \( \nu_2 \) and \( \lambda_2 \) (\( \frac{8}{9} \) units). The phase velocity of the higher-frequency is assumed to be \( \frac{12}{13} \approx 93\% \) of the velocity of the lower-frequency wave. The periods of the average and modulation waves are:

\[ \lambda_{\text{avg}} = 2 \frac{\lambda_1 \cdot \lambda_2}{\lambda_1 + \lambda_2} = 1 \text{ unit} \]

\[ \lambda_{\text{mod}} = 2 \frac{\lambda_1 \cdot \lambda_2}{|\lambda_1 - \lambda_2|} = 8 \text{ units} \]

The snapshots are taken at increments of \( T = \frac{T_{\text{avg}}}{4} \), so that the average wave propagates by one-quarter period between images. A point of constant phase on the average wave is denoted in each image by the black dot, which is seen to travel faster than a point of constant phase on the modulation wave; in this case, the ratio of modulation velocity to average velocity is approximately:

\[ \frac{v_{\text{mod}}}{v_{\text{avg}}} \approx 0.69 \]

Thus the “information” travels about 70% as fast as the average wave.
Illustration of normal dispersion of two waves shown at increments of $\Delta t = \frac{T_{avg}}{4}$. The black dot marks a point on the average wave with the same phase, which moves faster than the corresponding point on the modulation wave.
Chapter 5

Review: Doppler Effect

5.1 Transition from Acoustic Waves to Electromagnetic Waves

HR § 40

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

5.1.1 Acoustic Doppler Effect, Source at Rest

CASE I Motion of Receiver, Source and Medium at Rest

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

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

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

$$= \nu \left(1 \pm \frac{\nu_o}{v}\right) = \nu'$$

The + sign means that the receiver approaches the source.

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

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

5.1.2 Acoustic Doppler Effect – Source in Motion

CASE II Source Moves in Medium, Receiver Stationary
5.1 TRANSITION FROM ACOUSTIC WAVES TO ELECTROMAGNETIC WAVES

Doppler effect for sound waves if the source moves towards the observer. The circles represent wavefronts emitted by the source at different times, showing that they do not have a common center of symmetry.

Again, this is a snapshot of the wavefronts emitted by a source moving toward the receiver with velocity \( v_s \). The wavefronts emitted at later times have less distance to travel to the observer. The distance between adjacent wavefronts in the medium is actually shortened on one side and lengthened on the other, i.e.,

\[
\lambda' = \lambda \mp \Delta \lambda = \lambda \mp \frac{v_s}{v}, \text{ where the negative sign } \implies \text{ source approaching observer.}
\]

Therefore:

\[
\nu' = \frac{v}{\lambda'} = \frac{v}{\lambda \mp \frac{v_s}{v}} = \frac{v}{\frac{\lambda v + v_s}{v}} = \nu \left[ \frac{v}{\nu - v_s} \right] = \nu'
\]

where \( v \) is the velocity of the wave in the medium. For example with \( \nu = 1000 \text{ Hz} \), \( v_s = 60 \text{ m.p.h.} \approx \frac{26.8 \text{ m}}{\text{s}} \) toward observer

\[
\nu' = 1000 \text{ Hz} \cdot \left[ \frac{330 \text{ m}}{303.2} \right] \approx 1000 \text{ Hz} \cdot \left[ \frac{330}{303.2} \right] = 1088 \text{ Hz} > 1000 \text{ Hz}
\]

In the case of the source moving in the medium, the frequency is significantly different than for the case of the observer moving (1088 Hz vs. 1081 Hz); this difference can be detected to determine whether the observer or the source is moving relative to the medium.

5.1.3 Acoustic Doppler Effect — Both Source and Receiver Moving

Case III Both Source and Receiver Moving, Medium at Rest
CHAPTER 5 REVIEW: DOPPLER EFFECT

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

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

upper signs \(\implies\) source and receiver approach each other

5.2 Doppler Effect for Light

5.2.1 Difference between Light and Sound

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

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

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

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

\textit{“If not instantaneous, light is extraordinarily rapid.”}

A surprisingly good measurement of \(c\) was made by Ole Römer in 1675. The Keplerian laws of planetary motion enabled Römer to predict the times of eclipse of Jupiter’s Galilean satellites. He found that the measured times did not agree with prediction – when Jupiter was closest to earth, the times of eclipse were early, and when Jupiter was distant the times were late. Römer ascribed the difference to a finite velocity of light, and computed a value of \(c = 2 \cdot 10^8 \frac{m}{s}\). The largest source
of error was Römer’s lack of knowledge about the earth’s orbital velocity. When corrected for this error, Römer’s method yields a very accurate value of $3 \cdot 10^8 \text{m/s}$.

Besides its velocity, light had been demonstrated to have the character of a wave by Newton’s demonstration of dispersion by a prism and by the polarization experiments of Fresnel. These characteristics led to Fresnel’s hypothesis of the “aether” — the medium of transmission for light, which is analogous to air for propagation of sound. If it exists, the aether must be present everywhere, including in vacuum. From the calculations of the Doppler effect, the frequency shift of light must depend on whether the source or the observer is moving.

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

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

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

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

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

Einstein’s result is:

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

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

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

In this case, the series solution is:

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

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

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

In the case of light its velocity $c >> v$, we can dispense with the terms with orders
larger than unity:
\[ \nu' \simeq \nu \left(1 - \frac{\nu}{c}\right) = \nu - \nu \frac{v}{c} \text{ for } v << c. \]

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

Maxwell’s Equations for Electromagnetic Waves

6.1 Vector Operations

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

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

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

To minimize any confusion resulting from the use of the symbol “$x$” to represent both a vector and a particular component of a vector, a normal-faced “$x_i$” with a subscript will be used to indicate the $i^{th}$ component of the vector $\mathbf{x}$, while the bold-faced subscripted symbol “$\mathbf{x}_i$” denotes the $i^{th}$ member of a set of vectors. Other notations also will be employed during certain aspects of the discussion, but these cases will be explicitly noted.
Definitions of the algebraic operations of vectors will be essential to this discussion. For example, the sum of two N-D vectors \( \mathbf{x} \) and \( \mathbf{y} \) is generated by summing the pairs of corresponding components:

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

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

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

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

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

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

\[
\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} = \tilde{\mathbf{x}}
\]

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

\[
\sum_{n=1}^{N} (x_n)^2 \equiv |\mathbf{x}|^2
\]
The result is the squared magnitude of the vector. The vector’s length, or norm, is the square root of Eq. (3.5), as shown in the figure and thus also is real valued.

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

Length, or “norm”, of 2-D vector with real-valued components.

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

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

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

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

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

### 6.1.1 Scalar Product of Two Vectors

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

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

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

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

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

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

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

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

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

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

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

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

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

$$ \mathbf{f}_1 \cdot \mathbf{f}_2 = x_1 x_2 + y_1 y_2 $$

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

$$ = |\mathbf{f}_1| |\mathbf{f}_2| \cos \theta_1 \cos \theta_2 + |\mathbf{f}_1| |\mathbf{f}_2| \sin \theta_1 \sin \theta_2 $$

$$ = |\mathbf{f}_1| |\mathbf{f}_2| \cos (\theta_1 - \theta_2) = |\mathbf{f}_1| |\mathbf{f}_2| \cos \theta $$

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

$$ \mathbf{a} \cdot \mathbf{x} = |\mathbf{a}| |\mathbf{x}| \cos \theta = |\mathbf{a}| |\mathbf{x}| \cos \theta $$

where $\theta$ represents the “included” angle between the two N-D vectors. This angle is measured in the 2-D plane defined by the two vectors. If we consider the 3-D analogy of two vectors from the origin to the surface of a sphere, then the angle $\theta$ represents the angle along the “great circle” that connects the two vector tips.

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

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

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

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

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

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

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

### 6.1.2 Cross Product

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

The area of \( |\mathbf{A}| \cdot |\mathbf{B}| \sin \theta \) may be computed as a 3-D vector that points perpendicular to the two component vectors with length equal to the area; the calculation
6.1 VECTOR OPERATIONS

The "cross product" of the two 3-D vectors. Given the two component vectors:

\[ \mathbf{A} = \mathbf{\hat{x}}A_x + \mathbf{\hat{y}}A_y + \mathbf{\hat{z}}A_z \]
\[ \mathbf{B} = \mathbf{\hat{x}}B_x + \mathbf{\hat{y}}B_y + \mathbf{\hat{z}}B_z \]

the cross product is defined:

\[ \mathbf{A} \times \mathbf{B} = \det \begin{bmatrix} \mathbf{\hat{x}} & \mathbf{\hat{y}} & \mathbf{\hat{z}} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{bmatrix} = \mathbf{\hat{x}}(A_yB_z - A_zB_y) + \mathbf{\hat{y}}(A_zB_x - A_xB_z) + \mathbf{\hat{z}}(A_xB_y - A_yB_x) \]

In the example given, \( \mathbf{A} = \mathbf{\hat{x}}|\mathbf{A}|, \mathbf{B} = \mathbf{\hat{x}}(|\mathbf{B}| \cos [\theta]) + \mathbf{\hat{y}}(|\mathbf{B}| \sin [\theta]) \), so that \( A_z = B_z = 0 \)

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

It is easy to see that:

\[ \mathbf{B} \times \mathbf{A} = -\mathbf{A} \times \mathbf{B} \]

Note that the cross product is defined for 3-D vectors ONLY (though we can apply the definition to 2-D vectors by considering their component in the third direction to be zero, see example for curl that follows later).

6.1.3 Triple Vector Product

The "triple vector product" is the cross product of two 3-D vectors (call them \( \mathbf{A} \) and \( \mathbf{B} \)) crossed with a third vector (\( \mathbf{C} \)). The result may be evaluated by straightforward...
(yet tedious!) calculation and produces the result:

\[ \mathbf{A} \times \mathbf{B} \times \mathbf{C} = \mathbf{B} (\mathbf{C} \cdot \mathbf{A}) - \mathbf{A} (\mathbf{B} \cdot \mathbf{C}) \]

\[ = \mathbf{B} (\mathbf{C} \cdot \mathbf{A}) - \mathbf{A} (\mathbf{C} \cdot \mathbf{B}) \]

where the fact that the scalar product commutes for vectors with real-valued components has been used. The triple vector product yields the difference of two scaled replicas of the first two vectors, where the scaling factors are the scalar products of \( \mathbf{C} \) with \( \mathbf{A} \) and \( \mathbf{B} \). The “output” is a vector, as it must be.

We will use this expression for the triple vector product to evaluate the “curl of the curl” shortly.

### 6.2 Vector Calculus

Feynman, *Lectures on Physics II*, Chapter 2,3

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

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

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

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

If the displacement is the differential element \( d\mathbf{s} = \hat{x}dx + \hat{y}dy + \hat{z}dz \), then the scalar
6.2 VECTOR CALCULUS

Product yields a differential element of work

\[ dW = \mathbf{F} \cdot d\mathbf{s} \]

and the work resulting by the action of the force from point \( a \) to point \( b \) is:

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

Note that no work is performed if the force acts at right angles to the displacement; the work is “positive” if the force acts in the direction of the displacement (e.g., a weight dropping in a gravitational field); the work is “negative” if the force acts in opposition to the displacement.

The work can be evaluated via:

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

\[ = \int F_x \, dx + \int F_y \, dy + \int F_z \, dz = T + c \]

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

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

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

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

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

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

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

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

so that:

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

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

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

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

\[ \frac{\partial}{\partial y} F_x = \frac{\partial}{\partial y} \left( -\frac{\partial V}{\partial x} \right) = -\frac{\partial^2 V}{\partial x \partial y} \]
\[ \frac{\partial}{\partial x} F_y = \frac{\partial}{\partial x} \left( -\frac{\partial V}{\partial y} \right) = -\frac{\partial^2 V}{\partial y \partial x} = \frac{\partial}{\partial y} F_x \]

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

The same pattern of operations leads to two other relations:

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

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

We can then write:

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

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

\[ \nabla = \left[ \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right] \]
It also may be written in explicit vector form as:

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

where \( \hat{x}, \hat{y}, \) and \( \hat{z} \) are the unit vectors along the \( x, y, \) and \( z \) axes respectively. Thus we can write:

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

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

\[ \nabla (a + b) = \nabla a + \nabla b \]

\[ \nabla (\alpha a) = \alpha \nabla a \]

where \( a, b \) are scalar fields (functions) and \( \alpha \) is a numerical constant.

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

### 6.3 Gradient

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

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

\[ \nabla f [x, y, z] \equiv \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{bmatrix} = \hat{x} \frac{\partial f}{\partial x} + \hat{y} \frac{\partial f}{\partial y} + \hat{z} \frac{\partial f}{\partial z} = \text{a vector} \]

As implied by its name, the gradient vector at \( [x, y, z] \) points “uphill” in the direction of maximum rate-of-change of the field; the magnitude of the gradient \( |\nabla f| \) is the slope of the scalar field.\( x: \)
Gradient of the scalar field is a vector field. At each coordinate \([x, y]\), the vector points “uphill” and its length is equal to the slope.

As an aside, the operator \(\nabla\) is only defined in Cartesian coordinates; general expressions in cylindrical and spherical coordinates do not exist.

### 6.4 Divergence

*Derives a Scalar Field \(\nabla \cdot \mathbf{g}\) from a Vector Field \(\mathbf{g}\)*
\[ \nabla \cdot \mathbf{g}(x, y, z) \equiv \left[ \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right] \cdot [g_x, g_y, g_z] \]
\[ = \frac{\partial g_x}{\partial x} + \frac{\partial g_y}{\partial y} + \frac{\partial g_z}{\partial z} = \text{a scalar} \]

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

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

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

\[ dF = \mathbf{g} \cdot d\mathbf{a} \]
\[ \implies F = \iint_{\text{surface area } A} \mathbf{g} \cdot d\mathbf{a} \]

The divergence of the vector field describes the total flux out of a volume \( V \), which is equivalent to the flux through the macroscopic surface area \( A \) of the volume \( V \), which is in turn built up from all of the differential surface elements \( d\mathbf{a} \) enclosing the volume.

\[ F = \iiint_{\text{volume } V} (\nabla \cdot \mathbf{g}) \, dV \]
\[ = \iint_{\text{surface area } A} \mathbf{g} \cdot d\mathbf{a} \]
\[ = \iint_{\text{surface area } A} \mathbf{g} \cdot \hat{n} \, da \]

where \( \hat{n} \) is the unit vector normal to the differential element of area \( d\mathbf{a} \). If there are no net “sources” or “sinks” of flux within the volume (points from which the flux in the field “diverges out of” or “converges into”), then the divergence of flux through that surface must be zero:

\[ F = \iiint_{\text{volume } V} (\nabla \cdot \mathbf{g}) \, dV = \iint_{\text{surface area } A} \mathbf{g} \cdot d\mathbf{a} = 0 \]

if no “source” or “sink” of vector field within \( V \)
If there is a source of flux in the volume, then the divergence of the vector field is proportional to that source “strength”:

\[
\iint_{\text{surface area } A} \mathbf{g} \cdot d\mathbf{a} > 0
\]

A vector field with nonzero divergence has a disparity between input and output flux.

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

A vector field whose divergence is zero used to be called “solenoidal;” magnetic fields are solenoidal.

### 6.4.1 Gauss’ Theorem for Divergence

Consider the 6-sided cube with one face (#1) located in the \(y-z\) plane at \(x = 0\), and the “opposite” face (#2) parallel to the \(y-z\) plane at \(x = \Delta x\), as shown in the figure.
We want to evaluate the flux of the vector field \( \mathbf{g} [x, y, z] \) through each of the 6 faces. The “inward” flux through face #1 is the area integral of \( g_x [0, y, z] \):

\[
g_1 = \int \int g_x [0, y, z] \ dy \ dz
\approx g_x [0, y, z] \cdot \Delta y \cdot \Delta z
\]

The “inward flux” through face #2 (opposite #1) is the area integral:

\[
g_2 = \int \int -g_x [\Delta x, y, z] \ dy \ dz = -\int \int g_x [\Delta x, y, z] \ dy \ dz
\approx \left( -g_1 + \frac{\partial g_1}{\partial x} \cdot \Delta x \right) \cdot \Delta y \cdot \Delta z
\]

The sum of the two fluxes is:

\[
g_1 + g_2 = \left[ \frac{\partial g_x}{\partial x} \cdot \Delta x \right] \cdot \Delta y \cdot \Delta z
\]

Similarly, the sum of the fluxes through opposite faces are:

\[
g_3 + g_4 = \left[ \frac{\partial g_y}{\partial y} \cdot \Delta y \right] \cdot \Delta x \cdot \Delta z
\]
\[
g_5 + g_6 = \left[ \frac{\partial g_z}{\partial z} \cdot \Delta z \right] \cdot \Delta x \cdot \Delta y
\]
The sum of the 6 fluxes is easy to evaluate:

\[
\sum_{n=1}^{6} g_n = \frac{\partial g_x}{\partial x} \cdot \Delta x \cdot \Delta y \cdot \Delta z + \frac{\partial g_y}{\partial y} \cdot \Delta x \cdot \Delta y \cdot \Delta z + \frac{\partial g_z}{\partial z} \cdot \Delta x \cdot \Delta y \cdot \Delta z
\]

\[
= \left( \frac{\partial g_x}{\partial x} + \frac{\partial g_y}{\partial y} + \frac{\partial g_z}{\partial z} \right) \cdot \Delta x \cdot \Delta y \cdot \Delta z
\]

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

This is the sum of the normal vectors of the surface fluxes through the sums of the surfaces:

\[
(\nabla \cdot \mathbf{g}) \Delta V = \int \int_{\text{surface of differential volume}} (\mathbf{g}[x, y, z] \cdot \hat{n}) \, da
\]

If we construct a macroscopic volume by summing up these differential volumes, we obtain the integral formula:

\[
\int \int \int_{\text{macroscopic volume}} (\nabla \cdot \mathbf{g}) \, dV = \int \int_{\text{surface of macroscopic volume}} (\mathbf{g}[x, y, z] \cdot \hat{n}) \, da
\]

### 6.5 Curl

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

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

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

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

\[
\nabla \times \mathbf{g}[x, y, z] = \det \begin{bmatrix}
\hat{x} & \hat{y} & \hat{z} \\
\frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
g_x & g_y & g_z
\end{bmatrix}
\]

\[
= \hat{x} \left( \frac{\partial g_z}{\partial y} - \frac{\partial g_y}{\partial z} \right) + \hat{y} \left( \frac{\partial g_z}{\partial x} - \frac{\partial g_x}{\partial z} \right) + \hat{z} \left( \frac{\partial g_y}{\partial x} - \frac{\partial g_x}{\partial y} \right) = \text{a vector}
\]

To visualize curl, imagine a vector field that describes motion of a fluid (e.g., water or wind). If a paddle-wheel placed at some location in the fluid does not revolve, the field at that location has no curl. If the wheel does revolve about some axis, then the
curl is nonzero. The direction of the curl vector is that of the axis of the paddlewheel for the maximum rotation in the counterclockwise direction; the magnitude of the curl vector is that rotation “rate.” The algebraic sign of the curl is determined by the direction of rotation (counterclockwise \( \Rightarrow \) positive curl). The paddle will rotate only if the vector field is spatially nonuniform, e.g., it will not rotate in a vector field where all vectors have the same length and point in the same direction. A vector field with zero curl used to be called “irrotational;” electrostatic fields (i.e., NOT traveling electric waves) are irrotational.

Note that some points in the field can have zero curl while others have nonvanishing curl. Both vector fields shown in the examples of divergence have zero curl, since a paddle wheel placed at any point in either field will not rotate.

\[
\begin{align*}
\mathbf{f}_1[x,y] & \neq 0 \\
\mathbf{f}_2[x,y] & \neq 0
\end{align*}
\]

*Two vector fields with nonzero curl. The “paddlewheel” rotates in both cases.*

### 6.5.1 Example of Function with Large Curl

Consider the 3-D field composed of vectors that satisfy:

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

The vectors in this field lie in the \( x - y \) plane and those located on the \( x \) or \( y \) axes are oriented perpendicular to the axes and get longer with increasing distance from the origin.
The vector field \( \mathbf{g}[x, y, z] = [-y, x, 0] \)

This drawing is not very complete and so might be misleading; consider the vector located at \([x, y] = [1, 1]:\)

\[ \mathbf{g}[1, 1, 0] = -\hat{x} + \hat{y} + 0\hat{z} \]

and so “points” in the diagonal direction (towards \( \theta = +\frac{3\pi}{4} \) or towards about 10:30 if you remember what an analog clock face looks like!). Therefore the vectors in this field define a counterclockwise “flow” where the velocity of the flow increases with radial distance. It is a 1-D analogue of the “bathtub drain vortex.” The magnitudes and azimuth angles of the vectors in this field may be evaluated:

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

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

Now evaluate the partial derivatives of the vectors:

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

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

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

The curl of the field is obtained by substitution:

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

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

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

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

If the divergence

\section{6.6 Laplacian}

The divergence of the gradient often appears in problems in electromagnetic theory and in imaging. It may be applied to scalar functions via:
\[ \nabla \cdot \nabla f = \nabla^2 f \]
\[ = \left( \hat{x} \frac{\partial}{\partial x} + \hat{y} \frac{\partial}{\partial y} + \hat{z} \frac{\partial}{\partial z} \right) \cdot \left( \hat{x} \frac{\partial f}{\partial x} + \hat{y} \frac{\partial f}{\partial y} + \hat{z} \frac{\partial f}{\partial z} \right) \]
\[ = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \]

where the scalar operator \( \nabla^2 \) is:
\[ \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \]

Since the Laplacian is a scalar operator, it also may be applied to a vector field to produce the sum of the Laplacians of the three component functions along the three directions:
\[ (\nabla \cdot \nabla) \mathbf{g} = \nabla^2 \mathbf{g} \]
\[ = \hat{x} \frac{\partial^2 g}{\partial x^2} + \hat{y} \frac{\partial^2 g}{\partial y^2} + \hat{z} \frac{\partial^2 g}{\partial z^2} \]

The Laplacian in other coordinate systems (cylindrical or spherical) also may be defined, but produces more complicated expressions (that we probably don’t need!)

\[ \nabla^2 f [x, y, z] = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \text{ (Cartesian)} \]
\[ \nabla^2 f (\rho, \phi, z) = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial f}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2} \text{ (cylindrical)} \]
\[ \nabla^2 f (r, \theta, \phi) = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \text{ (spherical)} \]
\[ = \frac{\partial^2 f}{\partial r^2} + \frac{2}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r^2 \sin^2 \phi} \frac{\partial^2 f}{\partial \phi^2} \text{ (spherical)} \]

The Laplacian is the spatial derivative in the 3-D wave equation, which will be considered in more detail shortly:
\[ \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \mu \frac{\partial^2 f}{\partial t^2} \]
\[ \implies \nabla^2 f = \frac{1}{\nu^2} \frac{\partial^2 f}{\partial t^2} \]

### 6.6.1 Curl of Curl

The curl of the curl may be evaluated via the vector triple product that was presented earlier:
\[ \nabla \times (\nabla \times \mathbf{g}) = \nabla (\nabla \cdot \mathbf{g}) - (\nabla \cdot \nabla) \mathbf{g} = \nabla (\nabla \cdot \mathbf{g}) - \nabla^2 \mathbf{g} \]

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

### 6.7 Electric and Magnetic Fields

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

\[
\mathbf{F} \propto Q \cdot \mathbf{E} \quad \Rightarrow \quad \frac{\mathbf{F}}{Q} \propto \mathbf{E} \text{ measured in } \left[ \frac{\text{kg} - \text{m}}{\text{s}^2 - \text{C}} \right]
\]

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

#### 6.7.1 A Note on Units

If you consult other books, you will likely see many differences in the equations due to the different systems of units used in electromagnetics (and thus in optics); many students (including the author!) find it difficult to cut through the seeming morass of differences. For example, two of the well-known physics texts on the subject, by Lorrain and Corson and by Jackson, use different systems; the former uses the rationalized MKS system (meter, kilogram, second), the latter uses CGS units (centimeter, gram, second), which includes many factors of \( 4\pi \). The systems evolved from Coulomb’s law that evaluates the force between two electrical charges \( Q_1 \) and \( Q_2 \):

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

The constant of proportionality may be called \( k \):

\[
\mathbf{F} = k \frac{Q_1 Q_2}{r_{12}^2} \mathbf{\hat{r}}_{12}
\]
If the charges are measured in electrostatic units (esu) (also called statcoulombs), the distance in centimeters, and the force in dynes \( (g \cdot cm^2 \cdot s^{-2}) \), then \( k = 1 \). This means that two charges of 1 esu separated by 1 cm produces a force of 1 dyn. But what if the charges are measured in coulombs \([C]\), the distance in meters \([m]\), and the force in newtons \((1 N = 1 kg \cdot m \cdot s^{-2})\)? The value of \( k \) is determined from the knowledge that there are \( 10^5 \) dyn per N, \( 2.998 \cdot 10^9 \) esu per C, and \( 10^{-2} \) cm per m:

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

The force between two charges of 1 C separated by 1 m is nearly \( 10^{10} \) N \( \approx 4.5 \cdot 10^{10} \) pounds of force \([\text{lbf}]\), or about 1,100,000 tons (of force)! The constant \( k \) generally is normalized by a factor of \( 4\pi \):

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

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

where 1 Farad \([F]\) is equivalent to:

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

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

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

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

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

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

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

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

6.7.2 Magnetic Fields

The concept of a magnetic field is seemingly somewhat less intuitive, so we’ll consider it in somewhat more detail. The magnetic field is measured in terms of the “flux”
(often labelled by $\phi$), which is a term arising from the original concept of “lines” of magnetic flux emanating from the magnetic “poles.” In fact, the original CGS unit for magnetic flux was called the “line” (now called the “maxwell,” Mx). The flux emanating from a unit field of 1 gauss is $4\pi$ lines because the area of the sphere is $4\pi r^2$. The MKS unit of magnetic flux is the “weber” (1 Wb = $10^8$ Mx), which was defined as the amount of flux which, when changing uniformly in one second, induces 1 volt in 1 turn of a conductor. In electromagnetism, the more important quantity is the magnetic flux density, labeled $B$ and measured in gauss (CGS) or tesla (MKS). One gauss is one line (maxwell) through an area of 1 cm$^2$ and one tesla is 1 Wb per m$^2$:

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

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

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

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

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

where $\varepsilon$ and $\mu$ are the electric permittivity and magnetic permeability of the material, respectively. These are measures of the ability of the electric and magnetic fields to “permeate” the medium; if $\varepsilon$ is increased, then a larger electric field exists within the material, if $\mu$ is larger, then the magnetic field $\mathbf{H}$ does not penetrate as far into the medium.

Since we will consider propagation of light only in vacuum, $\mathbf{D} = \mathbf{E}$ and $\mathbf{B} = \mathbf{H}$. In vacuum, $\mu$ and $\varepsilon$ are denoted $\mu_0$ and $\varepsilon_0$ and both are set to unity in CGS units. In MKS units, the quantities are:
\[ \mu_0 = 4\pi \cdot 10^7 \frac{N}{A^2} \quad (\text{newton per square ampere}) \]

\[ \epsilon_0 = 8.85 \cdot 10^{-12} \frac{F}{m} \quad (\text{farads per meter}) \]

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

\[
v = \frac{1}{\sqrt{\mu \epsilon}}
\]

\[
n = \frac{c}{v} = \frac{\sqrt{\mu \epsilon}}{\sqrt{\mu_0 \epsilon_0}}
\]

### 6.8 Maxwell’s Equations

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

#### 6.8.1 Gauss’ Law for Electric Fields

Gauss’ law relates the flux of the electric field over a closed surface to the total charge enclosed by the surface. In its simplest terms, Gauss’ law states that the existence of electrical charges within a volume produces electric fields that pass through the surface of the volume. The flux of the field through the surface is proportional to the “amount” of charge within the volume. If the volume is enlarged, then so is the surface area, so the flux density through the surface must decrease at the same rate that the surface area increases. Also note that if there is no charge within the volume, there still can be flux through the enclosing surface, but the ingoing and outgoing parts of the flux cancel out.

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

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

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

\[ \iiint_{\text{volume}} d\Phi = \iiint_{\text{surface}} \mathbf{E} \cdot d\mathbf{a} \]

\[ = \frac{Q}{\varepsilon} = \frac{1}{\varepsilon} \iiint_{\text{volume}} \rho[x, y, z] \, dV \]

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

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

\[ = \iiint_{\text{volume}} (\nabla \cdot \mathbf{E}[x, y, z, t]) \, dV \]

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

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

In words, the divergence of the vector electric field is a scalar that is proportional to the total charge within the volume.

### 6.8.2 Gauss’ Law for Magnetic Fields

Since there are no magnetic analogues for “charges”, the volume cannot enclose a magnetic analogue of \( \rho \) or \( Q \), which leads to the particularly simple forms for Gauss’s law for the magnetic flux density:

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

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

In other words, the flux of the magnetic field through any enclosed surface ALWAYS is zero. This is often interpreted by the statement that there are no magnetic analogues of charges (called magnetic “monopoles”). An “electric monopole” is an electron or proton that carries a net charge of one “sign.”

### 6.8.3 Faraday’s Law of Magnetic Induction

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

The mathematical formulation of Faraday’s magneto-electric induction is called Faraday’s law, which states that the rate of change of a magnetic field through a 2-D surface is proportional to the circulation of the electric field around the 1-D perimeter of the 2-D surface. In mathematical terms, the time derivative of the magnetic field is proportional to the particular spatial derivative (the curl) of the electric field:

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

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

### 6.8.4 Ampere’s Law

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

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

where the additional source term \( \mathbf{J} \) is the “current density” of the electric field (measured in amperes per unit volume, equivalent to coulombs per second per unit volume). Note the difference in the algebraic sign in the analogous expressions of Faraday’s law and Ampere’s law.

We have already seen that:

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

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

There are two “source” terms in Maxwell’s equations: the “static” charge density \( \rho \) and the “dynamic” current density \( \mathbf{J} \). These can only be nonzero within media (such as copper wire) and thus vanish in vacuum. If we consider the propagation of light only in a vacuum, neither electric charges nor conductors are present and both source terms vanish.

### 6.8.5 Maxwell’s Equations

(Jackson, *Classical Electrodynamics*, §6)

In 1864, James Clerk Maxwell collected these four equations and derived the form
of the fields that simultaneously satisfy them in some simple cases. In rationalized MKS units, the differential forms of the equations (after adding the correction term to Ampere’s law) are:

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

These four coupled first-order differential equations can be solved directly in many cases; we will do a simple one now that leads to propagation of electromagnetic plane waves.

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

\[
\begin{align*}
\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} &= 0 \quad \text{Gauss’ Law for \( \mathbf{E} \) if no sources present} \quad (1) \\
\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} &= 0 \quad \text{Gauss’ Law for \( \mathbf{B} \)} \quad (2) \\
-\frac{\partial B_x}{\partial t} &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \quad \text{Faraday’s Law} \quad (3) \\
-\frac{\partial B_y}{\partial t} &= \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \quad (4) \\
-\frac{\partial B_z}{\partial t} &= \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \quad (5) \\
+\mu \frac{\partial E_x}{\partial t} &= \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} \quad \text{Ampere’s Law} \quad (6) \\
+\mu \frac{\partial E_y}{\partial t} &= \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} \quad (7) \\
+\mu \frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} \quad (8)
\end{align*}
\]

### 6.9 Wave Equation

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

\[
\nabla \times (\nabla \times \mathbf{E}) = \nabla (\nabla \cdot \mathbf{E}) - (\nabla \cdot \nabla) \mathbf{E} = \nabla (\nabla \cdot \mathbf{E}) - \nabla \cdot \nabla \mathbf{E} = 0 - \nabla^2 \mathbf{E}
\]

where Gauss’ law applicable in “sourcefree” regions has been used in the last step (the enclosed charge is zero). The right side of the equation may be rewritten by applying Ampere’s law:

\[
\nabla \times \left( -\frac{\partial \mathbf{B}}{\partial t} \right) = -\frac{\partial}{\partial t} (\nabla \times \mathbf{B}) = -\epsilon \mu \frac{\partial}{\partial t} \left( \frac{\partial \mathbf{E}}{\partial t} \right) = -\frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2}
\]

After equating the two sides of the equation:

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

where \( c \) is the velocity of light. This expression relates the spatial and temporal second derivatives of the electric field and is called the wave equation. It was first introduced by d’Alembert in 1747. It assumes that no energy of the wave is lost as it propagates, as would occur if friction or damping forces are present.

The general wave equation may be written:

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

where \( v \) is the velocity of a point of constant phase: our old friend the phase velocity. The wave equation may be rigorously derived for a transverse wave on a string — you probably saw this in a classical mechanics course.

The wave equation for electric fields confirms our earlier observation:

\[
\epsilon \mu = \frac{1}{c^2} \implies c = \sqrt{\frac{1}{\epsilon \mu}}
\]

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

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

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

Any differential equation is linear, so that if \( \psi_1 [z, t] \) and \( \psi_2 [z, t] \) are solutions to the
equation, so is $a\psi_1[z,t] + b\psi_2[z,t]$. The linearity property means that light beams can pass “through” each other and that waves can constructively or destructively interfere.

The wave equation has the simple solution:

$$\psi [z,t] = f [z \pm vt]$$

where $f [u]$ is any function that may be differentiated twice.

**Proof.**

Define $u \equiv z \pm vt \implies \frac{\partial u}{\partial z} = 1, \frac{\partial u}{\partial t} = \pm v$

Apply the chain rule:

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial u} \cdot 1 \implies \frac{\partial^2 f}{\partial z^2} = \frac{\partial^2 f}{\partial u^2} \cdot v^2$$

and $\frac{\partial f}{\partial t} = \frac{\partial f}{\partial u} \cdot (\pm v) \implies \frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial u^2} \cdot v^2$

Substitute into wave equation:

$$\frac{\partial^2 f}{\partial z^2} = \frac{1}{v^2} \frac{\partial^2 f}{\partial t^2} \implies \frac{\partial^2 f}{\partial u^2} = \frac{1}{v^2} \left( \frac{\partial^2 f}{\partial u^2} \cdot v^2 \right) = \frac{\partial^2 f}{\partial u^2}$$

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

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

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

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

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

thus agreeing with our previous expression for phase velocity.

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

$$\psi [z,t] = f [z - vt]$$

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

$$\psi [z,t] = g [z + vt]$$

which moves towards $z = -\infty$. 
The spatial derivative of the corresponding 3-D wave equation is the sum of the three second partial derivatives:

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

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

The 3-D wave may still be a sinusoid with argument in radians, so we must be more careful about how the 3-D function becomes a 1-D function. The \(x, y,\) and \(z\) dependencies all have associated “wavelengths” that may be defined by their corresponding “wavenumber” \(k_x, k_y, k_z\) that may be written as a “wavevector” \(\mathbf{k}_0\):

\[
\psi [x, y, z, t] = A \cos \left[ \Phi [x, y, z, t] \right] \\
= A \cos \left[ k_x x + k_y y + k_z z \pm \omega_0 t - \phi_0 \right] \\
= A \cos \left[ \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} \pm \omega_0 t - \phi_0 \right] \\
= A \cos \left[ \mathbf{k}_0 \cdot \mathbf{r} \pm \omega_0 t - \phi_0 \right]
\]

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

### 6.9.1 Electromagnetic Waves from Maxwell’s Equations

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

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

We will now solve these equations for a single specific case: an infinite plane electric field wave propagating in vacuum toward \(z = +\infty\). The locus of points of constant phase (often called a wavefront) of a plane wave is (obviously) a plane. The vector electric field \(\mathbf{E}\) is assumed to be constant as a function of \(x\) and/or \(y\) at a particular value of \(z\), but its vector amplitude can vary with \(z\); this variation will be shown to be sinusoidal. This constraint means that derivatives of the components “transverse” to the direction of travel must vanish:

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

and thus the expression for the 4-D vector field \(\mathbf{E}[x, y, z, t]\) can be simplified; the components are no longer functions of \(x\) or \(y\):

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

\[(10)\]
From (9) and Gauss’ law for electric fields (1), we find that:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0$$

$$\Rightarrow 0 + 0 + \frac{\partial E_z}{\partial z} = 0 \Rightarrow \frac{\partial E_z}{\partial z} = 0$$

(11)

Since the derivative of $E_z$ with respect to $z$ vanishes, then the $z$-component of the electric field $E_z$ must be constant, and its amplitude is arbitrary so we select it to be 0:

$$E_z[x, y, z] = \text{constant} \to 0$$

(12)

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

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

(13)

i.e., the only existing electric field is perpendicular (transverse) to $z$! This is a significant result in and of itself. We can simplify eq.(13) by rotating the coordinate system about the $z$ axis to align $\mathbf{E}$ with the $x$-axis, so that:

$$E_y[z, t] = 0 \text{ by assumption}$$

(14)

The resulting expression for the electric field is now very simple:

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

(15)

We can substitute this expression into Faraday’s Law (eqs. 3,4,5) to find the magnetic field:

$$-\frac{\partial B_x}{\partial t} = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = 0 - \frac{\partial E_y}{\partial z} \implies \frac{\partial B_x}{\partial t} = \frac{\partial E_y}{\partial z} = 0 \implies B_x[t] \text{ is constant with time}$$

(3)

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = \frac{\partial E_y}{\partial z} \implies \frac{\partial B_y}{\partial t} = -\frac{\partial E_x}{\partial z}$$

(4)

$$-\frac{\partial B_z}{\partial t} = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = 0 \implies B_z[t] \text{ is constant with time}$$

(5)

We can arbitrarily set the constant terms $B_x[t] = B_z[t] = 0$, so the only remaining equation is:

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z}$$

(4)

which says that the time derivative of the magnetic field $B_y$ is equal to to the negative of the space derivative of $E_x$. We can now find a relation between $B_y$ and $E_x$ by standard solution techniques of differential equations. For example, we can use the
6.9 WAVE EQUATION

wave equation:

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

\[ \rightarrow \frac{\partial^2 E_x [z, t]}{\partial z^2} = \frac{1}{c^2} \frac{\partial^2 E_x [z, t]}{\partial t^2} \]

Assume that \( E_x [z, t] \) varies sinusoidally with \( z \):

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

It is easy to show that this satisfies the wave equation if the phase velocity of the wave is what we would predict:

\[ \frac{\partial^2 E_x [z, t]}{\partial x^2} = \frac{\partial^2 E_x [z, t]}{\partial y^2} = 0 \]

\[ \frac{\partial^2 E_x [z, t]}{\partial z^2} = -(k_0^2) E_0 \cos [k_0 z - \omega_0 t] \]

\[ \frac{\partial^2 E_x [z, t]}{\partial t^2} = -(-\omega_0)^2 E_0 \cos [k_0 z - \omega_0 t] \]

\[ \implies v_0 = \frac{\omega_0}{k_0} \]

Now substitute this to solve for \( B_y \) by integration:

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

\[ \implies -B_y [z, t] = \int \frac{\partial E_x}{\partial z} dt = -(k_0 E_0) \int \sin [k_0 z - \omega_0 t] dt \]

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

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

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

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

where \( v_\phi \) is the phase velocity of the electromagnetic wave that we have already defined.

\[ B_y = \frac{E_0}{v_\phi} \cos [k_0 z - \omega_0 t] = \frac{E_x}{v_\phi} \implies E_x = v_\phi B_y \]
CHAPTER 6 MAXWELL’S EQUATIONS FOR ELECTROMAGNETIC WAVES

Note that the only existing component of \( B \) is \( B_y \), which is perpendicular to the component \( E_x \) of \( E \), and that both \( E \) and \( B \) are transverse to the direction of propagation (in the \(+z\) direction). Also note that the sinusoidal variations of \( E \) and \( B \) have the same arguments, which means that they oscillate “in phase”. The amplitude of the magnetic field is smaller by the factor of the phase velocity \( v_\phi = c \), so the effect of the magnetic field on observations is generally much smaller and often may be ignored in physical situations, though it is essential for the electric field to propagate.

6.9.2 Poynting Vector

Given transverse electric and magnetic fields \( E \) and \( B \), the light propagates in the mutually orthogonal direction \( k \):

\[
\mathbf{s} \equiv c^2 \mathbf{E} \times \mathbf{B}
\]

This is “Poynting’s vector” and has dimensions of power per unit area (e.g., \( \text{W/m}^2 \)). It measures the total flow of energy through unit area in unit time.

In this case with \( E[x,y,z,t] = \hat{x}E_x[z,t] \) and \( B[x,y,z,t] = \hat{y}B_y[z,t] \), the propagation direction is:

\[
\begin{align*}
E[x,y,z,t] &= \hat{x}E_x[z,t] \\
B[x,y,z,t] &= \hat{y}B_y[z,t] = \frac{\hat{y}E_x}{v}[z,t]
\end{align*}
\]

\[
\mathbf{s} = (\hat{x} \times \hat{y}) \cdot v^2 \frac{E_x^2}{E_v} = \hat{z} (v^2 \varepsilon E_x^2)
\]

along the \(+z\) direction, as predicted.

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

\[
I[x,y,z,t] = \langle \mathbf{s}[x,y,z,t] \rangle \equiv \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} \mathbf{s}[x,y,z,t'] \, dt'
\]

\[
= \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} c^2 \varepsilon \left( \hat{x}E_0 \cos[k_0z - \omega_0t] \right) \times \hat{y} \left( \frac{E_0}{c} \cos[k_0z - \omega_0t] \right) \, dt'
\]

\[
= \frac{\hat{z}(\varepsilon E_0^2)}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} c^2 \varepsilon \frac{E_0^2}{c} \cos^2[k_0z - \omega_0t] \, dt'
\]

\[
= \hat{z} (c \varepsilon E_0^2) \cdot \frac{1}{\Delta T} \int_{t-\frac{\Delta T}{2}}^{t+\frac{\Delta T}{2}} \cos^2[k_0z - \omega_0t] \, dt'
\]

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

\[
|\mathbf{s}|_{\text{average}} = \frac{1}{2} \frac{E_0^2}{v}
\]
6.9 WAVE EQUATION

6.9.3 Redux on Phase Velocity of Electromagnetic Waves

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

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

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

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

\[
\Rightarrow \mu \varepsilon E_0 (\omega_0 \sin [k_0 z - \omega_0 t]) = -\frac{E_0}{v_\phi} (\omega_0 - k_0 \sin [k_0 z - \omega_0 t])
\]

\[
\Rightarrow \mu \varepsilon \omega_0 E_0 = \frac{E_0 k_0}{v_\phi}
\]

\[
\Rightarrow \mu = \frac{k_0}{\omega_0 v_\phi} = \frac{1}{v_\phi^2} \Rightarrow v_\phi^2 = \left( \frac{\omega_0}{k_0} \right)^2 = \frac{1}{\mu \varepsilon}
\]

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

which we already knew from the wave equation. In vacuum, \( \mu \equiv \mu_0 \), \( \varepsilon \equiv \varepsilon_0 \), \( v_\phi \equiv c \), and \( c = \sqrt{\frac{1}{\mu_0 \varepsilon_0}} \). The permittivity and permeability of free space (vacuum) can be measured in laboratory experiments, thus allowing a calculation of the phase velocity of electromagnetic waves. The permeability in vacuum is:

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

and the permittivity is:

\[
\varepsilon_0 \equiv 8.85 \times 10^{-12} \text{ farads} \text{m}^{-1}
\]
These values produce the result:

$$\mu_0 \epsilon_0 = \left( 8.85 \cdot 10^{-12} \frac{\text{coul}}{\text{v-m}} \right) \left( 1.26 \cdot 10^{-6} \frac{\text{J}}{\text{amp}^2 \text{m}} \right)$$

$$= 1.11 \cdot 10^{-17} \frac{\text{coul}}{\text{v-m}} \cdot \frac{\text{J}}{\text{v-m} \left( \frac{\text{coul}^2 \cdot \text{m}}{\text{s}^2} \right)}$$

$$= 1.11 \cdot 10^{-17} \frac{1}{\text{v-m}} \cdot \frac{\text{J-s}^2}{\text{coul-m}}$$

$$= 1.11 \cdot 10^{-17} \frac{s^2}{\text{m}^2}$$

$$\Rightarrow c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} = 2.99 \cdot 10^8 \frac{\text{m}}{\text{s}}$$, which agrees with experiment.

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

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

So therefore

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

**Index of Refraction**

Of course, the dimensionless ratio of the velocity of light in vacuum to that in the medium is the _index of refraction_ $n$:

$$n \equiv \frac{c}{v} \geq 1$$

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

$$\Rightarrow n^2 = \frac{\epsilon}{\epsilon_0}$$

For metals and absorptive materials, the index of refraction is complex valued and the permeabilities may not be equal. The complex refractive index often is denoted by $\tilde{n}$ and its imaginary part by $\kappa$:

$$\tilde{n}^2 \equiv (n + i\kappa)^2 = \frac{\mu \epsilon}{\mu_0 \epsilon_0}$$
so that
\[ |k| = k = \frac{n \omega_0}{c} = (n + i\kappa) \frac{\omega_0}{c} \]

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

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

If we assume that the direction of propagation \( \hat{s} \) is in the direction of \( \mathbf{r} \) (as in a plane wave), then
\[
\hat{s} \cdot \mathbf{r} = |\hat{s}| |\mathbf{r}| \cos(0) = |\mathbf{r}| = r
\]

so that the electric field may be simplified to:

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

where the second term decays with increasing \( r \); it is attenuated. The amplitude decreases by the factor of \( e^{-1} \approx 0.368 \) of the incident value when \( r \) is equal to the skin depth \( \delta \):

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

where \( \lambda_0 \) is the wavelength measured in vacuum. The skin depth \( \delta \) is a measure of the distance that light will penetrate through the attenuating medium. In a metal, the imaginary part \( \kappa \) of the index of refraction can be large, which means that the skin depth is small and the electric field “lies” on the surface of the metal; little field penetrates to the interior.

If \( n \) (and thus \( k \)) is real valued (i.e., \( \kappa = 0 \), as is true for optically transparent media), then the electric field is

\[
E_0 \exp \left[ i \omega_0 \left( \frac{n}{c} r - t \right) \right] = E_0 \exp \left[ i \omega_0 \left( \frac{r}{v} - t \right) \right] = E_0 \exp \left[ i (nk_0 r - \omega_0 t) \right]
\]

which confirms that the velocity is

\[
v = \frac{c}{n}
\]

and that the wavelength in the medium is:

\[
\lambda_0' = \frac{\lambda_0}{n}
\]
In the case of complex-valued refractive index, the magnetic field is obtained from the electric field via:
\[
\vec{B} = \frac{k \times \vec{E}}{\omega} = \frac{\tilde{n}}{c} (\hat{s} \times \vec{E})
\]
where \(\hat{s}\) is the unit vector in the direction of propagation of the light (the Poynting vector). Note that some authors write \(\tilde{n} = n (1 + ik)\), where \(k\) is the attenuation index.

Values of the refractive index for common materials include:

<table>
<thead>
<tr>
<th>Medium</th>
<th>(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>1.0 (by definition)</td>
</tr>
<tr>
<td>air</td>
<td>(\approx 1.00027)</td>
</tr>
<tr>
<td>water</td>
<td>1.33</td>
</tr>
<tr>
<td>“crown” glass</td>
<td>1.5</td>
</tr>
<tr>
<td>“flint” glass</td>
<td>1.7</td>
</tr>
<tr>
<td>diamond</td>
<td>2.417</td>
</tr>
<tr>
<td>germanium</td>
<td>(\approx 4.0) (only transparent for (\lambda \gtrsim 2\mu m))</td>
</tr>
</tbody>
</table>

### 6.10 Consequences of Maxwell’s Equations

1. **Copropagation of \(\vec{E}\) and \(\vec{B}\):** The wave travels in a direction mutually perpendicular to both \(\vec{E}\) and \(\vec{B}\), and in fact the propagation direction is defined by the direction:
\[
\hat{s} \propto \vec{E} \times \vec{B} \text{ – the Poynting vector}
\]
In other words, the wave requires both electric and magnetic fields to propagate, and they copropagate. I like to interpret this result as meaning that the magnetic field provides the medium for propagation of the electric field and vice versa.

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

3. In vacuum, \(\vec{E}\) and \(\vec{B}\) are in-phase, which means that the phases of the sinusoidal variation of \(\vec{E}\) and \(\vec{B}\) are identical (the phases of the fields often are out of phase in some types of matter).

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

5. Energy is carried by both the electric and magnetic fields, and the magnitude of the energy \(\mathcal{E} \propto \vec{E}_0^2\).
6. There is no limitation on the possible frequencies of the waves, i.e., \([0 \leq \omega \leq \infty]\), which implies the allowed wavelengths are in the interval \([\infty \geq \lambda \geq 0]\).

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

\[
I [x, y, z, t] = \frac{1}{2} (cE_0^2)
\]

Relationship between \(E\) and \(B\) for a linearly polarized wave traveling from left to right; the fields are in phase.

### 6.11 Dispersion Redux

Earlier we considered the effects of dispersion on traveling waves from a simple point of view where we just assumed that waves with different temporal frequencies might travel at different speeds. We called the dispersion “normal” if the velocity of waves with longer wavelengths exceeds that of waves with shorter wavelengths, which also means that the index of refraction decreases with increasing wavelength. At this point, we will try to understand why this is so and also determine the conditions that are necessary for anomalous dispersion (where the index of refraction increases with increasing wavelength).

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

\[
v_\phi \propto \frac{\sqrt{\varepsilon}}{\rho}
\]

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

\[
n \cong A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4}
\]
where \( A, B, \) and \( C \) are constants determined from measurements of \( n \) at three wavelengths. This expression gives a pretty good approximation of \( n \). For example, the Cauchy formula for the index of refraction of air is:

\[
n \approx 1.000287566 + \frac{1.3412 \text{nm}^2}{\lambda^2} + \frac{0}{\lambda^4}
\]

which decreases with increasing wavelength.

The phenomenon of anomalous dispersion may have been observed (but not pursued) by William Fox Talbot. The first significant study was performed by Le Roux in 1862, who discovered that the index of refraction of a prism containing iodine vapor was 1.020 for red light and 1.019 for blue light, so that \( n_{\text{blue}} < n_{\text{red}} \). Within 10 years, Christiansen noted similar behavior in an analine dye that exhibited a strong absorption of green light, normal refraction for red, orange, and yellow light, but a smaller refraction angle for blue light. The refractive index of a medium that exhibited strong absorption was seen to increase rapidly as the wavelength was decreased approaching the absorption band. It took some time to produce a theory of matter and light that explained this effect.

The reason for dispersion is due to the interaction of light with matter. Light can be absorbed by matter, where the energy of the light wave is converted to energy of some form in the matter (e.g., it may increase the thermal energy in the matter. Light also can be scattered, where the electric charges in the matter (protons in the nuclei or, more usually, electrons in the atomic shells) absorb and then re-emit electromagnetic waves. If the scattering is elastic, then no energy is transferred to the medium, all of it stays in the electromagnetic wave. If inelastic, some energy is transferred to the medium. Scattering generally occurs when the waves encounter a structure or obstacle whose dimension is smaller than a wavelength, and the wave is re-emitted into a new direction. Electromagnetic waves are scattered by electric charges that may be bound in an atom or free (unbound). A material like glass contains many charges, and the macroscopic effect is the sum of the effects from each individual charge.

The interactions of light with matter are characterized by two numerical factors: the “absorption coefficient” \( \alpha \) and the (possibly complex valued) “refractive index” \( n \), which both may be functions of the frequency of the incident light. The absorption is due to transfer of energy from the light to the medium; at frequencies where the absorption coefficient is small, the light can penetrate the matter to a significant depth, and thus the matter is “transparent.”

Light impinging on a medium causes the charged particles (protons in atomic nuclei or electrons in atomic shells) to vibrate at the oscillation frequency of the light. These accelerated charges emit light of that same oscillation frequency in turn (this is the scattered light). The relative phase of the incident light and the re-emitted light determines much of the effect of the medium on the incident light. For example, if the incident and scattered waves are out of phase by \( \approx 180^\circ \) in some direction, then the light beam propagating in that direction will be attenuated.

The oscillations damp out because the electrons are influenced by other forces,
including the electric forces induced by neighboring charges. Thus the interaction of light with the medium acts like a damped harmonic oscillator that is driven by the sinusoidal force induced by the light wave. We can think of the interaction of light with charged particles in matter as a so-called “driven” or “forced” harmonic oscillator whose amplitude decreases with time due to the damping. The charges are “bound” to fixed equilibrium locations and can oscillate with one or more resonant frequencies determined by the internal forces due to neighboring atoms. The electric charges can absorb and re-emit the light (i.e., scatter it) in ANY direction. If the frequency of the light is close to the oscillation frequency of one of the resonant states, then some of the electromagnetic energy is retained by the charge and not scattered; it instead increases the energy of the charge. The wave then loses some amplitude when scattered. If the frequency of the light is far from a resonant frequency of the charges, then the scattered light constructively interferes along the same direction as the incident light and it can pass through the medium; in other words, the medium is transparent for light with those frequencies. Even in transparent media, the phase of the re-emitted light generally is shifted by the interaction of the medium.

The first question to consider is the reason why the light is “forward scattered” in transparent media. We’ve actually already given the answer; the light emitted in this direction by all atoms interferes constructively and light emitted in other directions interferes destructively. The charges in matter that scatter the light may be viewed as uniformly distributed and quite close together (separated by fractions of nanometers, significantly less than visible wavelengths). The phase “lag” of the forward-scattered light is equivalent to a “slowing down” of the light wave, hence the index of refraction.

6.11.1 Feynman’s Model for Refractive Index

From my point of view, the best discussion of refraction and dispersion was given by Richard Feynman in his famous Lectures on Physics, Volume I, Chapter 31. This discussion is derived from his treatment.

Consider an electromagnetic wave incident on a thin plate of glass. The source point is assumed to be a large distance away (to the left) and the observation point also is a large distance to the right (thus the figure is “not to scale!”).
CHAPTER 6 MAXWELL’S EQUATIONS FOR ELECTROMAGNETIC WAVES

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

The electric field at the measurement point is the sum of the original electric field plus a correction term $E_2$ due to emission by the electric charges in the glass plate. We can assume that the charges are electrons, since the protons are much more massive. The electrons in the glass oscillate under the influence of the incoming electric field from the source and thus emit their own electric fields. The observed field includes contributions from these charges. These modifications occur in a way that makes the field inside the glass “appear” to be moving at a different phase velocity — this is the reason why the index of refraction of glass is larger than one.

Mathematical Picture of Phase Change of Light Due to Glass

Consider one electron in the glass; it “feels” the effect of the incident field and the fields generated by all of the other charges in the glass, and the motions of all of these other charges are influenced by that one electron that we are observing. To simplify the problem, we assume that the influences of the other atoms are small relative to the effect of the source, so that the total field at the observation point is little affected by the motions of the other charges. In effect, we are assuming that the index of refraction of the glass is very small (close to one). The calculation will produce a field that travels in the same direction as the incident field ($E_2$) and a field that travels in the opposite direction ($E_1$ - the “reflected” field), but the latter is small because $n \approx 1$.

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

$$E_s = E_0 \exp \left[ +i (k_0z - \omega t) \right] = E_0 \exp \left[ +i \frac{\omega}{c} (\frac{z}{c} - t) \right]$$

where the factoring conveniently leaves the distance $z$. Here $\omega$ is the angular temporal frequency of the driving force, and thus is a variable in the problem (rather than a
parameter of the wave), hence it does not have a subscript. Assume that \( z = 0 \) at the “front” (input) side of the plate and \( z = \Delta z \) at the back side, so that the phase of the electric field at the front of the plate is:

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

If there were no glass, then the phase at the back of the plate observed would be identical to that at the front of the plate observed at an earlier time \( t' \):

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

where the phase increment is due to the delay during the transfer of the extra distance \( \Delta z \). If the glass “slows down” the light so that the velocity in the glass is \( v < c \), then the phase at the rear of the glass will include an additional factor

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

If we substitute the index of refraction, then the distance parts of the phase are proportional

\[
n = \frac{c}{v}
\]

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

Thus the “additional” phase due to the extra time to travel through the glass is:

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

The electric field at the back of the glass is:

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

Thus the contribution to the electric field due to the glass plate may be interpreted as an additive contribution to the phase instead of an additive contribution to the amplitude. The term with \( (n - 1) \) is the change due to the glass and will be evaluated
by physical arguments in a moment. It may be expanded in a Taylor series:

\[ \exp [+i\theta] = \frac{(+i\theta)^0}{0!} + \frac{(+i\theta)^1}{1!} + \frac{(+i\theta)^2}{2!} + \cdots + \frac{(+i\theta)^N}{N!} + \cdots \]

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

If we assume that the glass is thin, so that \( \Delta z \approx 0 \), then we can ignore all terms of order two or larger:

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

and thus the electric field at a large distance “behind” the glass plate is approximately:

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

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

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

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

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

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

We know that the leading factor \( +i = \exp \left[ +i \frac{\pi}{2} \right] \), which indicates that the electric field from the charges in the glass is out of phase with the original electric field by \( +\frac{\pi}{2} \) radians, as shown on the vector (phasor) diagram of the contributions of the two fields:

![Argand diagram of phasor contributions from incident field \( E_0 \) and field due to charges in the glass \( E_2 \), which is oriented approximately perpendicular to \( E_0 \) and](image-url)
Physical Picture of Electron Oscillations

If the field $E_2$ thus evaluated can be expressed in terms of the oscillating charges in the glass, then we will have explained the behavior of the refractive index. Again, we assume that the incident field has the form of a plane wave:

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

At the “front edge” of the glass ($z = 0$), the field is the same as given before:

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

The electrons in the glass “feel” this electric field and are driven in the same direction by the force:

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

(there also is a magnetic field, but its effect on the electrons is so much smaller that it can be ignored). The electrons have mass $m$ and act as though bound to the atoms by little springs that exert restoring forces proportional to the distance of the electron from its equilibrium position. The restoring force on the electron position has the form:

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

The electron oscillates at its “normal frequency,” which is:

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

this is a parameter of the electron + spring system (and thus has a subscript). The equation of motion of the electrons in the glass is:

$$m \frac{d^2 x}{dt^2} + m \omega_0^2 x = F = e E_0 \exp [-i\omega t]$$

where the last term is the “driving force” due to the electric field with the variable angular temporal frequency $\omega$. We solve this equation by standard methods of differential equations; we assume that the position $x$ of each electron also oscillates about its equilibrium point at the same rate. The amplitude of the oscillation is assumed to be $x_0$:

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

The derivatives needed in the equation of motion are

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

$$\frac{d^2 x}{dt^2} = (-i\omega)^2 x_0 \exp [-i\omega t] = -\omega^2 x$$
So the equation of motion is simplified to:

\[ m \left( -\omega^2 x_0 \right) \exp[-i\omega t] + m \omega_0^2 x_0 \exp[-i\omega t] = eE_0 \exp[-i\omega t] \]

\[ \implies \left( -m\omega^2 + m\omega_0^2 \right) x_0 = eE_0 \]

And thus the amplitude of the electron displacement from equilibrium may be expressed in terms of the “normal” oscillation frequency, mass, electric charge, and the amplitude and oscillation frequency of the incident electric field:

\[ x_0 = \frac{eE_0}{m (\omega_0^2 - \omega^2)} \]

Therefore the motions of EACH individual charge in the thin glass plate due to the incident electric field is the simple expression:

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

Note that this does not include the initial positions of the charges, which are (obviously) different for each.

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

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

The electric field at large distances from the charge that oscillates perpendicularly to the distance decreases approximately as the reciprocal of the distance and includes the time delay for the field to arrive:

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

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

The total field at the observation point is the vector sum of the contributions from the individual electrons in the thin glass plate. These contributions may be integrated in polar coordinates. If \( \eta \) is the number density of the electrons per unit area in the glass, then the electric field is:

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

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

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

Note that the term \( \exp \left[ +i \cdot \infty \right] \) oscillates, and rapidly, so we can assume it to be zero:
\[
E_{all} = 2\pi \eta e \left( -\omega^2 x_0 \right) \exp \left[ -i \omega t \right] \cdot -\frac{c}{i \omega} \exp \left[ +i \omega \frac{z}{c} \right]
\]
\[ = (2\pi \eta e x_0) \cdot (-i \omega) \exp \left[ -i \omega \left( t - \frac{z}{c} \right) \right]
\]
\[ = (2\pi \eta e) \cdot (-i \omega) x_0 \exp \left[ -i \omega t \right] \exp \left[ +i \omega \frac{z}{c} \right]
\]

which shows that the field measured at the observation point due to all of the oscillating charges is out of phase by \( \frac{\pi}{2} \) radians and delayed.

All we need to do now is substitute the formula for \( x_0 \) from the driven harmonic oscillator that was derived in the previous section:
\[
E_2 = (2\pi \eta e) \cdot (-i \omega) \frac{cE_0}{m(\omega_0^2 - \omega^2)} \exp \left[ -i \omega t \right] \exp \left[ +i \omega \frac{z}{c} \right]
\]
\[
E_2 = -i \omega \frac{(2\pi \eta e^2)E_0}{m(\omega_0^2 - \omega^2)} \exp \left[ +i \omega \left( \frac{z}{c} - t \right) \right]
\]

In words, the electrons in the glass that oscillate due to the incident field emit a wave that travels in the same direction (towards \( z = +\infty \)). The amplitude of the wave is proportional to the number of atoms \( \eta \) and to the strength of the source field \( E_0 \). This field resembles that evaluated before (in the earlier box)
\[ E_2 \cong -i \omega \frac{(n-1) \cdot \Delta z}{c} \left( E_0 \exp \left[ +i \omega \left( \frac{z}{c} - t \right) \right] \right) \]

These two are equal if we identify that some of the factors are identical:
\[
\frac{(n-1) \cdot \Delta z}{c} = \frac{(2\pi \eta e^2)}{m(\omega_0^2 - \omega^2)}
\]
\[ \implies (n-1) \Delta z = \frac{(2\pi \eta e^2 c^2)}{m(\omega_0^2 - \omega^2)}
\]
\[ \implies n = 1 + \frac{(2\pi e^2 c^2)}{m(\omega_0^2 - \omega^2)} \cdot \frac{\eta}{\Delta z} \]
CHAPTER 6 MAXWELL’S EQUATIONS FOR ELECTROMAGNETIC WAVES

We now define $N$ to be the number density of electrons per unit volume in the glass, which can be expressed as the product of the number density $\eta$ per unit area in the thin sheet and the thickness $\Delta z$:

$$N = \eta \cdot \Delta z \implies \frac{\eta}{\Delta z} = N$$

Thus the index of refraction can be written as:

$$n = 1 + \frac{2\pi Ne^2}{m(\omega_0^2 - \omega^2)} = 1 + \frac{\alpha}{\omega_0^2 - \omega^2}$$

where the constant $\alpha$ includes the contributions from the number density, the mass, the charge on the electron, and the velocity of light. This is the frequency dependent index of refraction in the simple model of a thin sheet of glass with oscillating bound electrons. This expression is graphed for the case $\omega_0 = 2$ and $\alpha = 1$, showing that the index is largest for light with temporal frequency $\omega \approx \omega_0$, i.e., in the vicinity of the frequency of the “resonant” oscillations of the electrons due to the restoring forces. In this simple model, the index of refraction can be less than unity, which indicates that this picture is incomplete.

Graph of $n$ vs. $\omega$ assuming that $\omega_0 = 2$ and $\frac{2\pi Ne^2}{m} = 1$, showing that the index generally increases with increasing $\omega$, which means that $n$ decreases with increasing wavelength; this is normal dispersion.

In thick glass, the fields from different thin sheets interact which increases the mean refractive index. Also the oscillations of the electrons are actually “damped” out by other forces in the glass. The sampling pushes the extrema of $n$ towards its mean value. To derive this behavior, we could continue the discussion in the same vein (most books do, including Feynman and Hecht). We could follow this line of reasoning, but we can demonstrate a useful connection to linear systems theory and thus simplify the derivation.
We can think of the action of light on matter as a linear shift-invariant system with a “causal” impulse response, i.e., a temporal system whose impulse response is zero for \( t < 0 \) and thus cannot “respond” until stimulated. An appropriate impulse response for a “causal” damped oscillating system is:

\[
h [t] = A_0 \exp [-\gamma_0 t] \cdot STEP [t] \cdot \sin [2\pi \nu_0 t] \\
= A_0 \exp [-\gamma_0 t] \cdot STEP [t] \cdot \cos \left[ 2\pi \nu_0 t - \frac{\pi}{2} \right]
\]

where \( \gamma_0 \) is the damping coefficient (i.e., the reciprocal of the time required for the output response to decrease by \( e^{-1} \approx 0.368 \)), \( \nu_0 \) is the natural oscillating frequency of the charged particle (the electron), and \( \phi_0 \) is the initial phase. This function \( h [t] \) measures the “response” of the system, which in our case is the position of the electron relative to its equilibrium position. The STEP function ensures that \( h [t < 0] = 0 \), meaning that the electron “sits” at its equilibrium position until disturbed; this forces the system to be causal. When stimulated by a “pulse of light,” (modeled by a Dirac delta function \( \delta [t] \)), the position of the electron increases from zero following a sinusoidal curve, but the amplitude of the sine wave decreases with time due to the decaying exponential, as shown:

\[
\text{Impulse response of damped harmonic oscillator:} \\
h [t] = \exp [-\gamma t] \cdot STEP [t] \cdot \sin [2\pi \nu t] \quad \text{where} \quad A_0 = 1, \ \nu_0 = 2, \ \text{and} \ \gamma_0 = \frac{1}{2}.
\]

The response of this system is characterized by the transfer function of the impulse response, which is its 1-D temporal Fourier transform. We use the known Fourier
transforms:

\[ F_1 \{ e^{-t} \cdot \text{STEP} [t] \} = \frac{1}{1 + 2\pi i \nu} = \frac{1 + i(-2\pi \nu)}{1 + (2\pi \nu)^2} \]

\[ F_1 \{ \sin[2\pi \nu_0 t] \} = i \cdot \frac{1}{2|\nu_0|} (\delta[\nu + \nu_0] - \delta[\nu - \nu_0]) \]

and the scaling theorem:

\[ F_1 \{ f[t] \} = F[\nu] \implies F_1 \{ f[\gamma_0 t] \} = \frac{1}{|\gamma_0|} F\left[\frac{\nu}{\gamma_0}\right] \]

to derive the transfer function of this system:

\[ H[\nu] = A_0 F_1 \{ \sin[2\pi \nu_0 t] \cdot (\exp[-\gamma_0 t] \cdot \text{STEP}[t]) \} \]

\[ = A_0 F_1 \{ \sin[2\pi \nu_0 t] \cdot (\exp[-\gamma_0 t] \cdot \text{STEP}[\gamma_0 t]) \} \]

\[ = i \left( \frac{A_0}{2|\gamma_0|} \right) (\delta[\nu + \nu_0] - \delta[\nu - \nu_0]) * \frac{1}{1 + 2\pi i \frac{\nu}{\gamma_0}} \]

\[ = i \left( \frac{A_0}{2|\gamma_0|} \right) (\delta[\nu + \nu_0] - \delta[\nu - \nu_0]) * \frac{1 - 2\pi i \frac{\nu}{\gamma_0}}{1 + \left(\frac{2\pi \nu}{\gamma_0}\right)^2} \]

\[ \text{Re} \{ H[\nu] \} = \frac{A_0 \pi}{\gamma_0} \left( \frac{\nu + \nu_0}{1 + \left(\frac{\nu + \nu_0}{\gamma_0}\right)^2} - \frac{\nu - \nu_0}{1 + \left(\frac{\nu - \nu_0}{\gamma_0}\right)^2} \right) \]

\[ = \frac{A_0}{2\gamma_0^2} \left( \frac{\omega + \omega_0}{\gamma_0^2 + (\omega + \omega_0)^2} - \frac{\omega - \omega_0}{\gamma_0^2 + (\omega - \omega_0)^2} \right) \]

\[ \text{Im} \{ H[\nu] \} = \frac{A_0}{2\gamma_0} \left( \frac{1}{\gamma_0^2 + (\omega + \omega_0)^2} - \frac{1}{\gamma_0^2 + (\omega - \omega_0)^2} \right) \]

The transfer function acts as a frequency-dependent "scale factor" applied to the amplitude of the electron oscillation. The graphs of the real part, imaginary part, magnitude, and phase are shown below, where the domain is assumed to include negative temporal frequencies. In this example, \( \nu_0 = 2 \) and \( \gamma_0 = 0.5 \).
In words, the transfer function measures the “response” of the system, i.e., the amplitude of the oscillation of the charged particle, which we called $x[t]$ before.

$$x[t] \propto x_0 \cdot (\text{Re}\{H[\nu]\} + i \text{Im}\{H[\nu]\}) \cdot \exp[-i\omega t]$$

$$= \frac{A_0 x_0}{2\gamma_0^2} \exp[-i\omega t] \left[ \left( \frac{\omega + \omega_0 + i\gamma}{\gamma_0^2 + (\omega + \omega_0)^2} - \frac{\omega - \omega_0 - i\gamma}{\gamma_0^2 + (\omega - \omega_0)^2} \right) \right]$$

The amplitude is again a function of the temporal frequency of the incident light. Note that the phase of the transfer function is approximately 0 radians for $|\nu| < |\nu_0|$, approximately $-\pi$ radians for $\nu > \nu_0$, and $-\frac{\pi}{2}$ radians if the frequency of the incident light is $+\nu_0$. This means that the system response (the oscillation of the charged particle) is “in phase” if the frequency of the incident light is less than the “resonant frequency” of the oscillating charge. The oscillation of the charged particle is “out of phase” if the frequency of the incident light is larger than the resonant frequency. Also note from the magnitude that the system response is quite large near resonance, which means that the oscillation of the charged particle is large. Of course the meaning of light with a negative temporal frequency is not very clear in this context, and can be rectified by recognizing that the real and imaginary part of the response of the system are related due to causality, a fact reflected in the Kramers-Kronig equations, which are beyond the scope of this discussion.

The index of refraction is proportional to this amplitude plus a bias, so we can use the graphs of $H[\nu]$ to understand the frequency behavior of $n$. As mentioned previously, the index of refraction decreases with increasing wavelength (increases...
with increasing temporal frequency), thus the phase velocity in a medium of light
with longer wavelengths is larger and the dispersion is normal. The phase velocity
of the modulation wave (the group velocity) is less than the phase velocity of the
average wave, and messages travel more slowly than the carrier wave that conveys
the message.

In the vicinity of an absorption due to the resonance of charged particles in the
medium, the index of refraction increases with increasing wavelength over a small
range, which means that shorter wavelengths travel faster and the dispersion is anom-
alous. In this region, the phase velocity of the modulation wave is larger than the
phase velocity of the average wave. This implies that messages can travel faster than
the velocity of light. HOWEVER, since this only happens where light is absorbed,
the message cannot propagate.

Real and imaginary parts of refractive index in vicinity of a “weak” absorption.

Real and imaginary parts of refractive index for multiple absorptions. Note that
anomalous dispersion only occurs in vicinity of absorptions, so that light cannot
6.12 Dual Nature of Light: Photons

In many contexts, the “particle” picture of light is more appropriate. In imaging, for example, consider images created with different exposure times. Photographs taken in with shorter exposures generally look grainier:

![Images created from increasing numbers of photons, showing increase in signal-to-noise ratio.](image)

A “particle of light” is the photon, whose energy is proportional to the temporal frequency:

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

where \( h \) is Planck’s constant, which is often normalized by a factor of \( 2\pi \), called “h-bar:”

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

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

If \( \lambda = 550 \text{ nm} \), the energy per photon is only:

\[
E = (6.625 \times 10^{-34} \text{ J} \cdot \text{s}) \cdot \frac{3 \times 10^8 \text{ m}}{550 \text{ nm}} \approx 3.6 \times 10^{-19} \text{ J}
\]

The “photon flux” is the number of photons per second in a light beam:

\[
\Phi = \frac{P}{h\nu}, \text{ where } P \text{ is the power}
\]

Typical fluxes per unit area for some sources are shown in the table.
The pattern of photon arrivals tells something about the source. Random (incoherent) light sources (such as light bulbs) emit photons with random arrival times and a Bose-Einstein distribution. Coherent light sources, on the other hand, emit photons with a Poisson distribution, which is more uniform but still random.

<table>
<thead>
<tr>
<th>Light Source</th>
<th>$\Phi/A$ in photons/sec-$m^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>focused laser</td>
<td>$10^{26}$</td>
</tr>
<tr>
<td>unfocused laser</td>
<td>$10^{21}$</td>
</tr>
<tr>
<td>bright sunlight</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td>indoor light</td>
<td>$10^{16}$</td>
</tr>
<tr>
<td>twilight</td>
<td>$10^{14}$</td>
</tr>
<tr>
<td>moonlight</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>starlight</td>
<td>$10^{10}$</td>
</tr>
</tbody>
</table>

6.12.1 Momentum of Photons

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

$$p = \frac{h}{\lambda} = h\kappa$$

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

\[ y(z, t) = A_0 \cos[kz \mp \omega t] = A_0 \cos \left( 2\pi \left( \frac{z}{\lambda} \mp \nu t \right) \right) = A_0 \Re \{ e^{i(kz \mp \omega t)} \} \]

We see electromagnetic radiation with detectors, i.e., devices which respond in some way to incident electromagnetic radiation. The human eye is sensitive only to visible light, i.e., light with wavelengths in the range 400 nm ≤ \( \lambda \) ≤ 700 nm. This is not the case for all life, however. The pit viper can see radiation emitted by humans at a wavelength of about 10\( \mu \)m; it needs special receptors on the sides of its head to do this.

As shown in the plot of the electromagnetic spectrum, the frequencies of visible wavelengths are quite large: \( \nu \approx 10^{15} \) Hz. The temporal period of an optical wave is therefore \( T = \nu^{-1} \approx 10^{-15} \) s. Human visual receptors cannot respond fast enough to detect the periodic oscillation of the wave amplitude; we see an invariant brightness. Note that this limitation exists for all detectors of visible radiation (e.g., photographic film, light meters, etc.); they all respond to the average brightness. The same is true for hearing; your ear cannot detect the variation of sound pressure due to the oscillation at frequencies above a few Hz. Because water waves have a much lower frequency, the amplitude and phase of the wave can be measured. Similarly, the phase can be measured of electromagnetic waves that have a much smaller temporal frequency, e.g., radio waves.

The average amplitude of a sinusoidal wave is:

\[
\langle y(z, t) \rangle = \frac{1}{T_d} \int_0^{T_d} y(z, t) \, dt = \frac{1}{T_d} \int_0^{T_d} A_0 \cos[kz - \omega t] \, dt = -\frac{A_0}{\omega T_d} \sin[kz - \omega t] \big|_{t=0}^{t=T_d}
\]

Since \( y(z, t) \) is sinusoidal, the average value of the wave will tend to zero unless \( T_d \) is smaller than the wave’s temporal period. However, the intensity (squared-magnitude)
of the wave does not average to zero:

\[ E \propto \langle y^2 [z, t] \rangle = \frac{1}{T_d} \int_0^{T_d} y^2 [z, t] \, dt \]

\[ = \frac{1}{T_d} \int_0^{T_d} A_0^2 \cos^2 [kz - \omega t] \, dt \]

\[ = \frac{A_0^2}{T_d} \int_0^{T_d} \cos^2 [kz - \omega t] \, dt \]

\[ = \frac{A_0^2}{T_d} \cdot \frac{T_d}{2} = \frac{A_0^2}{2} \]

\[ \Rightarrow E \propto \langle y^2 [z, t] \rangle \Rightarrow \frac{A_0^2}{2} \quad \text{if } T_d >> \nu^{-1} \]

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

Detectors of visible light are sensitive to time-averaged intensity, not amplitude.
Chapter 7

Propagation of Light Waves

7.1 Wavefronts

7.1.1 Plane Waves

The form of any wave (matter or electromagnetic) is determined by its source and described by the shape of its wavefront, i.e., the locus of points of constant phase. If a traveling wave is emitted by a planar source, then the points of constant phase form a plane surface parallel to the face of the source. Such a wave is called a plane wave, and travels in one direction (ideally). Since energy is conserved, the total energy in the wave must equal the energy emitted by the source, and therefore the energy density (the energy passing through a unit area), is constant for a plane wave. Recall that in a wave of amplitude $A$ and frequency $\omega$, the energy $E \propto A^2 \omega^2$. Therefore, for a plane wave, the amplitude is constant; the wave does not attenuate.

Plane wave toward $z = +\infty$ at velocity $v_\phi = \frac{\omega}{k}$, wavelength $\lambda = \frac{2\pi}{k}$, frequency $\nu = \frac{\omega}{2\pi}$, amplitude $A_0$:

$$f[x, y, z, t] = A_0 \cos [kz \mp \omega t]$$

(n.b., no variation in $y$ or $z$)

General 3-D plane wave traveling in a direction $\mathbf{k} = [k_x, k_y, k_z]$, $\mathbf{r} = [x, y, z]$ and the definition of the scalar product (dot product):

$$f[\mathbf{r}, t] = A_0 \cos [\mathbf{k} \cdot \mathbf{r} - \omega t] \implies \mathbf{k} \cdot \mathbf{r} = k_x x + k_y y + k_z z$$
### 7.1.2 Cylindrical Waves

If a wave is emitted from a line source, the wavefronts are *cylindrical*. Since the wave expands to fill a cylinder of radius \( r_0 \), the wavefront crosses a cylindrical area that grows as \( \text{Area} = 2\pi rh \propto r \). Therefore, since energy is conserved, the energy per unit area must decrease as \( r \) increases:

\[
\frac{\mathcal{E}}{\text{Area}} = \text{constant} = \frac{\mathcal{E}}{2\pi rh} \propto \frac{\mathcal{E}}{r} \propto \frac{A_0^2}{r} = \text{constant}
\]

\[\implies \text{amplitude} \propto \frac{A_0}{\sqrt{r}}\]

The equation for a cylindrical wavefront emerging from (or collapsing into) a line source is:

\[
f[x, y, z, t] = A[r] \cos [kr \mp \omega t]
\]

\[= \frac{A_0}{\sqrt{r}} \cos [kr \mp \omega t])\]

\[r = \sqrt{x^2 + y^2} > 0\]

“−” \(\implies\) emerging

“+” \(\implies\) collapsing

\(A_0 = \text{amplitude at } r = 0\)
7.1 WAVEFRONTS

7.1.3 Spherical Waves

The wavefront emerging from (or collapsing into) a point is spherical. The area the wave must cross increases as \( x^2 + y^2 + z^2 = r^2 \) (area of sphere is \( 4\pi r^2 \)). Therefore the energy density drops as \( r^2 \) and the amplitude of the wave must decrease as \( \frac{1}{r} \). The equation for a spherical wave is

\[
 f \left[ x, y, z, t \right] = f \left[ r, t \right] = A[r] \cos [kr \mp \omega t] = A_0 \frac{r}{r} \cos [kr \mp \omega t], \text{ where } r > 0
\]

“−” \( \Rightarrow \) emerging
“+” \( \Rightarrow \) collapsing

\( A_0 = \) amplitude at \( r = 0 \)

Note the pattern for the amplitude of plane, cylindrical, and spherical waves:

plane wave \( \Rightarrow \) 2-D source (plane) \( \Rightarrow \) amplitude \( A[r] \propto r^{-0} = 1 \)
cylindrical wave \( \Rightarrow \) 1-D source (line) \( \Rightarrow \) \( A[r] \propto r^{-\frac{1}{2}} \)
spherical wave \( \Rightarrow \) 0-D source (point) \( \Rightarrow \) \( A[r] \propto r^{-1} \)
7.2 Huygens’ Principle

The spherical wave is the basic wave for light propagation using Huygens’ principle. In 1678, Christiaan Huygens theorized a model for light propagation that claimed that each point on a propagating wavefront (regardless of “shape”) could be assumed to be a source of a new spherical wave. The sum of these secondary spherical “wavelets” produced the subsequent wavefronts. Huygens’ principle had the glaring disadvantage that these secondary spherical wavefronts propagated “backwards” as well as forwards. This problem was later solved by Fresnel and Kirchhoff in the 19th century. With that correction, the Huygens’ model provides a very useful model for light propagation that naturally leads to expressions for “diffracted” light.
Chapter 8

Interaction of Light and Matter

8.1 Electromagnetic Waves at an Interface

A beam of light (implicitly a plane wave) in vacuum or in an isotropic medium propagates in the particular fixed direction specified by its Poynting vector until it encounters the interface with a different medium. The light causes the charges (electrons, atoms, or molecules) in the medium to oscillate and thus emit additional light waves that can travel in any direction (over the sphere of $4\pi$ steradians of solid angle). The oscillating particles vibrate at the frequency of the incident light and re-emit energy as light of that frequency (this is the mechanism of light “scattering”). If the emitted light is “out of phase” with the incident light (phase difference $\approx \pm \pi$ radians), then the two waves interfere destructively and the original beam is attenuated. If the attenuation is nearly complete, the incident light is said to be “absorbed.” Scattered light may interfere constructively with the incident light in certain directions, forming beams that have been reflected and/or transmitted. The constructive interference of the transmitted beam occurs at the angle that satisfies Snell’s law; while that after reflection occurs for $\theta_{\text{reflected}} = \theta_{\text{incident}}$. The mathematics are based on Maxwell’s equations for the three waves and the continuity conditions that must be satisfied at the boundary. The equations for these three electromagnetic waves are not difficult to derive, though the process is somewhat tedious. The equations determine the properties of light on either side of the interface and lead to the phenomena of:

1. Equal angles of incidence and reflection;
2. Snell’s Law that relates the incident and refracted wave;
3. Relative intensities of the three waves;
4. Relative phases of the three light waves; and
5. States of polarization of the three waves.

For simplicity, we consider only plane waves, so that the different beams are specified by single wavevectors $\mathbf{k}$, that are valid at all points in a medium and that
point in the direction of propagation. The lengths of the wavevectors are determined:

$$|\mathbf{k}_n| = \frac{2\pi}{\lambda_n} = \frac{2\pi}{\lambda_0} n$$

where $\lambda_0$ is the wavelength in vacuum and $\lambda_n$ is the wavelength in the medium. The interface between the media is assumed to be the $x-y$ plane located at $z = 0$. The incident wavevector $\mathbf{k}_0$, the reflected vector $\mathbf{k}_r$, the transmitted vector $\mathbf{k}_t$ and the unit vector $\hat{n}$ normal to the interface are shown:

The $\mathbf{k}$ vectors of the incident, reflected, and “transmitted” (refracted) wave at the interface between two media of index $n_1$ and $n_2$ (where $n_2 > n_1$ in the example shown).

The angles $\theta_0$, $\theta_r$, and $\theta_t$ are measured from the normal, so that $\theta_0, \theta_t > 0$ and $\theta_r < 0$ as drawn.

The incident and reflected beams are in the same medium (with $n = n_1$) and so have the same wavelength:

$$|\mathbf{k}_0| = |\mathbf{k}_r| = \frac{\omega_0}{v_1} = \frac{2\pi n_1}{\lambda_0}$$

$$\lambda_1 = \frac{2\pi n_1}{|\mathbf{k}_0|} = \frac{2\pi n_1}{|\mathbf{k}_r|}$$

The wavelength of the transmitted beam is different due to the different index of refraction:

$$\lambda_2 = \frac{2\pi n_2}{|\mathbf{k}_t|}$$
As drawn, the normal to the surface is specified by the unit vector perpendicular to the interface; in this case, it points in the direction of the positive z-axis:

\[ \hat{n} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \]

(we could have defined \( \hat{n} \) in the opposite direction).

The incident electric field is a sinusoidal oscillation that may be written in complex notation:

\[ E_{\text{incident}} = E_0 \exp \left[ +i (k_0 \cdot r - \omega_0 t) \right] \]

where \( r = [x, y, z] \) is the position vector of the location where the phase \( k_0 \cdot r - \omega_0 t \) is measured; note that the phases measured at all positions in a plane perpendicular to the incident wavevector \( k_0 \) must be equal (because this is a plane wave).

The reflected and transmitted waves have the general forms:

\[ E_{\text{reflected}} = E_r \exp \left[ +i (k_r \cdot r - \omega_r t + \phi_r) \right] \]
\[ E_{\text{transmitted}} = E_t \exp \left[ +i (k_t \cdot r - \omega_t t + \phi_t) \right] \]

where we have yet to demonstrate that \( \omega_r = \omega_t = \omega_0 \). The constants \( \phi_r \) and \( \phi_t \) are the (perhaps different) initial phases of the reflected and transmitted waves.

### 8.1.1 Snell’s Law for Reflection and Refraction

One boundary condition that must be satisfied is that the phases of all three waves must match at the interface \( (z = 0) \) at all times.

\[ (k_0 \cdot r - \omega_0 t) \big|_{z=0} = (k_r \cdot r - \omega t + \phi_r) \big|_{z=0} = (k_t \cdot r - \omega t + \phi_t) \big|_{z=0} \]

This equivalence immediately implies that the temporal frequencies of the three waves must be identical \( (\omega_0) \), because otherwise the phases would change by different amounts as functions of time. In words, the temporal frequency is invariant with medium, or the “color” of the light does not change as the light travels into a different medium. Therefore the spatial vectors must satisfy the conditions:

\[ (k_0 \cdot r) \big|_{z=0} = (k_r \cdot r + \phi_r) \big|_{z=0} = (k_t \cdot r + \phi_t) \big|_{z=0} \]

Since the scalar products of the three wavevectors with the same position vector \( r \) must be equal, then the three vectors \( k_0, k_r \) and \( k_t \) must all lie in the same plane (call it the x-z plane, as shown in the drawing). The number of waves per unit length at any instant of time must be equal at the boundary for all three waves, as shown:

\[ (k_0)_x = (k_r)_x = (k_t)_x \]
The $x$-components of the three wavevectors (for the incident, reflected, and transmitted refracted waves) must match at the interface to ensure that each produces the same number of waves per unit length.

From the definitions of the vectors we can also see that:

$$ (k_0)_x = |k_0| \cos \left[ \frac{\pi}{2} - \theta_0 \right] = |k_0| \sin [\theta_0] $$

$$ (k_r)_x = |k_r| \cos \left[ \frac{\pi}{2} - \theta_r \right] = |k_r| \sin [-\theta_r] $$

where the factor of $-1$ on the reflected angle is because the angle measured from the normal is clockwise, and hence negative. The equality of the lengths of the incident and reflected wavevectors immediately demonstrates that:

$$ (k_0)_x = (k_r)_x = |k_0| \sin [\theta_0] = |k_r| \sin [-\theta_r] $$

$$ \implies |k_0| \sin [\theta_0] = |k_r| \sin [-\theta_r] $$

$$ \implies \sin [\theta_0] = \sin [-\theta_r] $$

$$ \implies \theta_0 = -\theta_r $$

In words, the angle of reflection is equal to the negative of the angle of incidence. We usually ignore the sign of the angle and say that the angles of incidence and reflection are equal.

Now make the same observation for the transmitted wave:

$$ (k_0)_x = |k_0| \sin [\theta_0] = \frac{2\pi n_1}{\lambda_0} \sin [\theta_0] $$

$$ (k_t)_x = |k_t| \cos \left[ \frac{\pi}{2} - \theta_t \right] = |k_t| \sin [\theta_t] = \frac{2\pi n_2}{\lambda_0} \sin [\theta_t] $$
We equate these to derive the relationship of the angles of the incident and transmitted wavevectors:

\[
\frac{2\pi n_1}{\lambda_0} \sin [\theta_0] = \frac{2\pi n_2}{\lambda_0} \sin [\theta_t] \\
\implies n_1 \sin [\theta_0] = n_2 \sin [\theta_t]
\]

We recognize this to be (of course) *Snell’s law* for refraction.

The reflection law may be cast into the form of Snell’s refraction law by assuming that the index of refraction is negative for the reflected beam:

\[
n_1 \sin [\theta_0] = -n_1 \sin [\theta_r] \\
\implies \sin [\theta_r] = -\sin [\theta_0] \\
\implies \theta_r = -\theta_0
\]

Note that these laws were derived without having to consider the vector nature of the electric and magnetic fields, but rather just the spatial frequencies of the waves at the boundaries. The next task is not quite this simple.....

### 8.1.2 Boundary Conditions for Electric and Magnetic Fields

We’ve determined the angles of the reflected and transmitted (refracted) plane waves in the form of Snell’s law(s). We also need to evaluate the “quantity” of light reflected and refracted due to the boundary. Since the geometries of the fields will depend on the directions of the electric field vectors, we will have to consider this aspect in the derivations. In short, this discussion will depend on the “polarization” of the electric field (different from the “polarizability” of the medium). We will again have to match appropriate boundary conditions at the boundary, but these conditions apply to the vector components of the electric and magnetic fields on each side of the boundary. We use the same notation as before for amplitudes of the electric fields of the incident, reflected, and transmitted (refracted) waves. Faraday’s and Ampere’s laws (the Maxwell equations involving curl) for plane waves can be recast into forms that are more useful for the current task:

\[
\nabla \times \mathbf{E} \propto -\frac{\partial \mathbf{B}}{\partial t} \\
\nabla \times \mathbf{B} \propto +\frac{\partial \mathbf{E}}{\partial t}
\]

We need the constants of proportionality in this derivation. Recall that they depend on the system of units. We will use the MKS system here:

\[
\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\
\nabla \times \mathbf{B} = +\varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}
\]
where \( \epsilon \) and \( \mu \) are the permittivity and permeability of the medium, respectively and the phase velocity of light in the medium is:

\[
v_\phi = \sqrt{\frac{1}{\epsilon \mu}}
\]

The incident field is assumed to be a plane wave of the form already mentioned:

\[
E_{\text{incident}}[x, y, z, t] = E_0 \exp\left[i \left(k_0 \cdot \mathbf{r} - \omega_0 t\right)\right]
\]

We know that \( E_0 \perp \mathbf{k}_0 \). In our coordinate system, the incident wave vector lies in the \( x - z \) plane (the plane defined by \( \mathbf{k}_0 \) and \( \hat{n} \)), so that \( k_{0y} = 0 \):

\[
E_{\text{incident}}[x, y, z, t] = E_0 \exp\left[i \left(k_0 \cdot \mathbf{r} - \omega_0 t\right)\right] = E_0 \exp\left[i \left(k_{0x} x + k_{0y} y + k_{0z} z - \omega_0 t\right)\right]
\]

The boundary conditions that must be satisfied by the electric fields and by the magnetic fields at the boundary are perhaps not obvious. Consider the figure on the left:

The boundary conditions on the electric and magnetic fields at the boundary are established from these situations.

We assume that there is no charge or current on the surface and within the cylinder that straddles the boundary. If the height of the cylinder is decreased towards zero, then Gauss’ laws establish that the flux of the electric and magnetic fields through the top and bottom of the cylinder (the \( z \) components in this geometry) must cancel:

\[
\epsilon_1 E_{1z} \cdot \hat{n} - \epsilon_2 E_{2z} \cdot \hat{n} = 0
\]

\[
\Rightarrow \epsilon_1 E_{1z} = \epsilon_2 E_{2z}
\]

\[
B_{1z} \cdot \hat{n} - B_{2z} \cdot \hat{n} = 0
\]

\[
\Rightarrow B_{1z} = B_{2z}
\]
The flux of the electric field in a medium is the so-called “displacement” field $\mathbf{D} = \varepsilon \mathbf{E}$ and the flux of the magnetic field is the field $\mathbf{B}$. Thus Gauss’ law determines that the normal components of $\mathbf{D}$ and of $\mathbf{B}$ are continuous across the boundary of the medium.

The figure on the right is a rectangular path (a “loop”) that also straddles the boundary. The unit vector $\hat{\mathbf{t}} \perp \hat{\mathbf{n}}$ points along the surface. If the “height” of the loop $dh \to 0$, then the circulations of the electric and magnetic fields must cancel:

\[
\mathbf{E}_1 \cdot \hat{\mathbf{t}} - \mathbf{E}_2 \cdot \hat{\mathbf{t}} = 0
\]
\[
\implies E_{1x} = E_{2x}
\]

\[
\frac{\mathbf{B}_1}{\mu_1} \cdot \hat{\mathbf{t}} - \frac{\mathbf{B}_2}{\mu_2} \cdot \hat{\mathbf{t}} = 0
\]
\[
\implies \frac{B_{1x}}{\mu_1} = \frac{B_{2x}}{\mu_2}
\]

We now want to solve Maxwell’s equations for an incident plane wave, which will depend on the incident angle $\theta_0$ and on the vector direction of the electric field. It is convenient to evaluate these conditions in two cases of linearly polarized waves: (1) where the polarization is perpendicular to the plane of incidence defined by $\hat{\mathbf{n}}$ and $\mathbf{k}_0$ (the so-called “s” polarization or transverse electric (TE) waves), which also means that the electric field vector is “parallel” to the interface, and (2) the polarization is parallel to the plane of incidence defined by $\hat{\mathbf{n}}$ and $\mathbf{k}_0$ (the so-called “p” polarization or transverse magnetic (TM) waves). The two cases are depicted below:

The electric field perpendicular to the plane of incidence: this is the TRANSVERSE ELECTRIC field (TE, also called the “s” polarization).
CHAPTER 8 INTERACTION OF LIGHT AND MATTER

The electric field is parallel to the plane of incidence; this is the TRANSVERSE MAGNETIC field (TM, also called the “p” polarization).

8.1.3 Transverse Electric Waves, s Polarization

In the TE case in our geometry, the electric field is oriented along the $y$ direction and the wavevector has components in the $x$ and $z$ directions:

$$
E_{\text{incident}}[x, y, z, t] = (\hat{x} \cdot 0 + \hat{y} \cdot |E_0| + \hat{z} \cdot 0) \exp[+i (k_{0x}x + k_{0z}z - \omega_0t)] \\
= \hat{y}E_0 \exp[+i (k_{0x}x + k_{0z}z - \omega_0t)]
$$

The magnetic field is derived from the relation:

$$
B = \frac{n}{c}k_0 \times E
$$

$$
B_{\text{incident}}[x, y, z, t] = \left( -\cos[\theta_0] \cdot n_1 \frac{|E_0|}{c} \hat{x} + 0\hat{y} + \left[ +\sin[\theta_0] \cdot n_1 \frac{|E_0|}{c} \right] \hat{z} \right) \\
\cdot \exp[+i (k_{0x}x + k_{0z}z - \omega_0t)]
$$

The reflected fields are:

$$
E_{\text{reflected}}[x, y, z, t] = \hat{y} \cdot |E_0| \exp[+i (k_{r_x}x + k_{r_z}z - \omega_0t)]
$$

$$
B_{\text{reflected}}[x, y, z, t] = \left( +\cos[-\theta_0] \cdot n_1 \frac{|E_0|}{c} \hat{x} + \left[-\sin[-\theta_0] \cdot n_1 \frac{|E_0|}{c} \right] \hat{z} \right) \\
\cdot \exp[+i (k_{0x}x + k_{0z}z - \omega_0t)]
$$

$$
= \left( +\cos[\theta_0] \cdot n_1 \frac{|E_0|}{c} \hat{x} + \left[ \sin[\theta_0] \cdot n_1 \frac{|E_0|}{c} \right] \hat{z} \right) \\
\cdot \exp[+i (k_{0x}x + k_{0z}z - \omega_0t)]
$$
and the transmitted (refracted) fields are:

\[
\mathbf{E}_{\text{transmitted}}[x, y, z, t] = \hat{y} \cdot |\mathbf{E}_t| \exp \left[ +i \left( k_{tx} x + k_{tz} z - \omega_0 t \right) \right]
\]

\[
\mathbf{B}_{\text{transmitted}}[x, y, z, t] = \left( \begin{array}{c} -\cos [\theta_t] \cdot n_2 \frac{\mathbf{E}_t}{c} \\ \sin [\theta_t] \cdot n_2 \frac{\mathbf{E}_t}{c} \end{array} \right) \hat{x} + \left( \begin{array}{c} \sin [\theta_t] \cdot n_2 \frac{\mathbf{E}_t}{c} \\ -\cos [\theta_t] \cdot n_2 \frac{\mathbf{E}_t}{c} \end{array} \right) \hat{z} \cdot \exp \left[ +i \left( k_0 x + k_0 z - \omega_0 t \right) \right]
\]

The only components of the electric field at the interface are transverse, so the only boundary conditions to be satisfied are the tangential electric field:

\[
E_0 + E_r = E_t \implies 1 + \frac{E_r}{E_0} = \frac{E_t}{E_0}
\]

This is typically expressed in terms of the reflection and transmission coefficients for the amplitude of the waves (not the power of the waves; these are the reflectance \( R \) and transmittance \( T \) of the interface, which will be considered very soon):

\[
r_{TE} = \frac{E_r}{E_0} \quad t_{TE} = \frac{E_t}{E_0}
\]

where the subscripts denote the transverse electric polarization. The boundary condition for the normal magnetic field yields the expression:

\[
\frac{n_1}{\mu_1 c} \sin[\theta_0] (E_0 + E_r) = \frac{n_2}{\mu_2 c} \sin[\theta_t] E_t
\]

while that for the tangential magnetic field:

\[
\frac{n_1}{\mu_1 c} \cos[\theta_0] (E_0 - E_r) = \frac{n_2}{\mu_2 c} \cos[\theta_t] E_t
\]

These may be solved simultaneously for \( r \) and \( t \) to yield expressions in terms of the indices, permeabilities, and angles:

\[
\text{Reflectance Coefficient for TE Waves}
\]

\[
r_{TE} = \frac{E_r}{E_0} = \frac{\frac{n_1}{\mu_1} \cos[\theta_0] - \frac{n_2}{\mu_2} \cos[\theta_t]}{\frac{n_1}{\mu_1} \cos[\theta_0] + \frac{n_2}{\mu_2} \cos[\theta_t]}
\]

\[
|r_{TE}| = \frac{n_1 \cos[\theta_0] - n_2 \cos[\theta_t]}{n_1 \cos[\theta_0] + n_2 \cos[\theta_t]} \quad \text{if} \quad \mu_1 = \mu_2 \quad (\text{usual case})
\]
Transmission Coefficient for TE Waves

\[ t_{TE} = \frac{E_t}{E_0} = \frac{+2\frac{n_1}{\mu_1} \cos [\theta_0]}{\frac{n_1}{\mu_1} \cos [\theta_0] + \frac{n_2}{\mu_2} \cos [\theta_t]} \]

Again, these are the amplitude coefficients; the reflectance and transmittance of light at the surface relate the energies or powers. These measure the ratios of the reflected or transmitted power to the incident power. The power is proportional to the product of the magnitude of the Poynting vector and the area of the beam. The areas of the beams before and after reflection are identical, which means that the reflectance is just the ratio of the magnitudes of the Poynting vectors. This reduces to the square of the amplitude reflection coefficient:

\[ R = r^2 \]

which reduces to this expression for the TE case:

\[ R_{TE} = \left( \frac{n_1 \cos [\theta_0] - n_2 \cos [\theta_t]}{n_1 \cos [\theta_0] + n_2 \cos [\theta_t]} \right)^2 \]

The transmission \( T \) is a bit more complicated to compute, because the refraction at the interface changes the “width” of the beam in one direction (along the \( x \)-axis in this example), so that the area of the transmitted beam is different from that of the incident beam. This is illustrated in the figure for a case with \( n_1 > n_2 \):

\[ \text{Demonstration that the areas of the beams differ in the two media. This must be accounted for in the calculation of the power transmission } T. \]

The magnitude of the Poynting vector is proportional to the product of the index of
refraction and the squared magnitude of the electric field:

\[ |s_1| \propto n_1 |E_0|^2 \]
\[ |s_2| \propto n_2 |E_t|^2 \]

The ratio of the powers is:

\[ T = \frac{|s_2| A_2}{|s_1| A_1} = \frac{n_2 |E_t|^2 A_2}{n_1 |E_0|^2 A_1} = \frac{n_2}{n_1} \cdot t^2 \cdot \frac{A_2}{A_1} \]

The area of the transmitted beam changes in proportion to the dimension along the \( x \)-axis in this case, which allows us to see that:

\[ \frac{A_2}{A_1} = \frac{w_2}{w_1} = \frac{\sin \left[ \frac{\pi}{2} - \theta_t \right]}{\sin \left[ \frac{\pi}{2} - \theta_0 \right]} = \frac{\cos [\theta_t]}{\cos [\theta_0]} \]

which leads to the final expression for the transmission at the interface:

\[ T = \frac{n_2}{n_1} \cdot t^2 \cdot \left( \frac{\cos [\theta_t]}{\cos [\theta_0]} \right) \]

\[ T = \left( \frac{n_2 \cos [\theta_t]}{n_1 \cos [\theta_0]} \right) \cdot t^2 \]

Snell’s law gives a relationship between the incident and transmitted angles:

\[ n_1 \sin [\theta_0] = n_2 \sin [\theta_t] \implies \sin [\theta_t] = \frac{n_1}{n_2} \sin [\theta_0] \]

\[ \implies \cos [\theta_t] = \sqrt{1 - \sin^2 [\theta_t]} = \sqrt{1 - \left( \frac{n_1}{n_2} \sin [\theta_0] \right)^2} \]

Thus we can write down the transmittance \( T \) in terms of the refractive indices and the incident angle:

\[ T = \left( \frac{\sqrt{n_2^2 - n_1^2 \sin^2 [\theta_0]}}{n_1 \cos [\theta_0]} \right) \cdot t^2 \]

For the TE case, the transmission is:

\[ T_{TE} = \left( \frac{\sqrt{n_2^2 - n_1^2 \sin^2 [\theta_0]}}{n_1 \cos [\theta_0]} \right) \cdot \left( \frac{2n_1 \cos [\theta_0]}{n_1 \cos [\theta_0] + n_2 \cos [\theta_t]} \right)^2 \]

These will be plotted for some specific cases after we evaluate the coefficients for TM waves.
8.1.4 Transverse Magnetic Waves (p polarization)

In the TM case in our geometry, the electric field is in the x-z plane and the wavevector has components in the x and z directions:

\[
\mathbf{E}_{\text{incident}}[x, y, z, t] = (\hat{x} \cdot |\mathbf{E}_0| \cos \theta + \hat{z} \cdot |\mathbf{E}_0| \sin \theta) \exp \left[ +i (k_{0x}x + k_{0z}z - \omega t) \right]
\]

The magnetic field is in the y-direction:

\[
\mathbf{B}_{\text{incident}}[x, y, z, t] = \left( n_1 \frac{|\mathbf{E}_0|}{c} \hat{y} \right) \exp \left[ +i (k_{0x}x + k_{0z}z - \omega t) \right]
\]

The reflected fields are:

\[
\mathbf{E}_{\text{reflected}}[x, y, z, t] = (\hat{x} - |\mathbf{E}_0| \cos \theta - \hat{z} \cdot |\mathbf{E}_0| \sin \theta) \exp \left[ +i (k_{0x}x + k_{0z}z - \omega t) \right]
\]

\[
\mathbf{B}_{\text{reflected}}[x, y, z, t] = \left( n_1 \frac{|\mathbf{E}_0|}{c} \hat{y} \right) \exp \left[ +i (k_{0x}x + k_{0z}z - \omega t) \right]
\]

and the transmitted (refracted) fields are:

\[
\mathbf{E}_{\text{transmitted}}[x, y, z, t] = (\hat{x} - |\mathbf{E}_0| \cos \theta - \hat{z} \cdot |\mathbf{E}_0| \sin \theta) \exp \left[ +i (k_{tx}x + k_{tz}z - \omega t) \right]
\]

\[
\mathbf{B}_{\text{transmitted}}[x, y, z, t] = \left( n_2 \frac{|\mathbf{E}_0|}{c} \hat{y} \right) \exp \left[ +i (k_{tx}x + k_{tz}z - \omega t) \right]
\]

In the case, the boundary condition on the normal component of \( \mathbf{B} \) is trivial, but the other components are:

\[
\mu_1 \sin \theta \left( E_0 + E_r \right) = \mu_2 \sin \theta \left( E_t \right)
\]

\[
\cos \theta \left( E_0 - E_r \right) = \cos \theta \left( E_t \right)
\]

\[
\frac{n_1}{\mu_1 c} \left( E_0 + E_r \right) = \frac{n_2}{\mu_1 c} E_t
\]

These are solved for the reflection and transmission coefficients:

\[
R_{TM} = \frac{+n_2 \cos \theta - n_1 \cos \theta}{+n_2 \cos \theta + n_1 \cos \theta}
\]

which simplifies if the permeabilities are equal (as they usually are):

\[
R_{TM} = \frac{+n_2 \cos \theta - n_1 \cos \theta}{+n_2 \cos \theta + n_1 \cos \theta}
\]

if \( \mu_1 = \mu_2 \)
The corresponding reflectance is:

\[ R_{TM} = \left( \frac{+n_2 \cos \theta_0 - n_1 \cos \theta_t}{+n_2 \cos \theta_0 + n_1 \cos \theta_t} \right)^2 \]

The amplitude transmission coefficient evaluates to:

\[ t_{TM} = \frac{2 \frac{n_1}{\mu_1} \cos \theta_0}{+n_2 \frac{1}{\mu_2} \cos \theta_0 + \frac{n_1}{\mu_1} \cos \theta_t} \]

again, if the permeabilities are equal, this simplifies to:

\[ t_{TM} = \frac{2n_1 \cos \theta_0}{+n_2 \cos \theta_0 + n_1 \cos \theta_t} \] if \( \mu_1 = \mu_2 \)

The corresponding transmittance function is:

\[ T_{TM} = \left( \frac{\sqrt{n_2^2 - n_1^2 \sin^2 \theta_0}}{n_1 \cos \theta_0} \right) \cdot \left( \frac{2n_1 \cos \theta_0}{+n_2 \cos \theta_0 + n_1 \cos \theta_t} \right)^2 \]

### 8.1.5 Comparison of Coefficients for TE and TM Waves

We should compare the coefficients for the two cases of TE and TM waves. The reflectance coefficients are:

\[ r_{TE} = \frac{n_1 \cos \theta_0 - n_2 \cos \theta_t}{n_1 \cos \theta_0 + n_2 \cos \theta_t} \]

\[ r_{TM} = \frac{+n_2 \cos \theta_0 - n_1 \cos \theta_t}{+n_2 \cos \theta_0 + n_1 \cos \theta_t} \]

where the angles are also determined by Snell’s law:

\[ n_1 \sin \theta_0 = n_2 \sin \theta_t \]

\[ \implies \cos \theta_t = \sqrt{1 - \left( \frac{n_1}{n_2} \sin \theta_0 \right)^2} \]

Note that angles and the indices for the TE case are “in” the same media, i.e., the index \( n_1 \) multiplies the cosine of \( \theta_0 \), which is in the same medium. The same condition holds for \( n_2 \) and \( \theta_t \). The opposite is true for the TM case: \( n_1 \) is applied to \( \cos \theta_t \) and \( n_2 \) to \( \cos \theta_0 \). These same observations also apply to the corresponding transmission
coefficients:

\[
t_{TE} = \frac{+2n_1 \cos \theta_0}{n_1 \cos \theta_0 + n_2 \cos \theta_t} \\
t_{TM} = \frac{+2n_1 \cos \theta_0}{+n_2 \cos \theta_0 + n_1 \cos \theta_t}
\]

**Normal Incidence (\(\theta_0 = 0\))**

In the case of normal incidence where \(\theta_0 = \theta_r = \theta_t = 0\), then the TE and TM equations evaluate to:

\[
\begin{align*}
R_{TE}|_{\theta_0=0} &= \frac{n_1 - n_2}{n_1 + n_2} \\
R_{TM}|_{\theta_0=0} &= \frac{+n_2 - n_1}{+n_2 + n_1} = - (R_{TE}|_{\theta_0=0}) \\
t_{TE}|_{\theta_0=0} &= \frac{+2n_1}{n_1 + n_2} \\
t_{TM}|_{\theta_0=0} &= \frac{+2n_1}{n_1 + n_2} = t_{TE}|_{\theta_0=0}
\end{align*}
\]

cases are identical. Also, the areas of the incident and transmitted waves are identical so there is no area factor in the transmittance. The resulting formulas for reflectance and transmittance reduce to:

\[
\text{normal incidence (}\theta_0 = 0\)) \\
R_{TE} (\theta_0 = 0) = R_{TM} (\theta_0 = 0) \equiv R = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 \\
T = \frac{4n_1 n_2}{(n_1 + n_2)^2}
\]

**Example: Rare-to-Dense Reflection**  If the input medium has a smaller refractive index \(n\) (a rarer medium) than the second (denser) medium, so that \(n_1 < n_2\), then the coefficients are:
\begin{align*}
n_1 &= 1.0 \\
n_2 &= 1.5 \\
\implies r_{TE} &= \frac{1.0 - 1.5}{1.0 + 1.5} = -0.2 = 0.2e^{+i\pi} \\
\implies r_{TM} &= \frac{1.5 - 1.0}{1.0 + 1.5} = +0.2 \\
\implies t_{TE} &= t_{TM} = \frac{2 \cdot 1.0}{1.0 + 1.5} = +0.8 \\
\implies R_{TE} &= R_{TM} = 0.04 \\
\implies T_{TE} &= T_{TM} = 0.96
\end{align*}

In words, the phase of the reflected light is changed by \( \pi \) radians = 180\(^\circ\) if reflected at a “rare-to-dense” interface such as the usual air-to-glass case.

**Example: Dense-to-Rare Reflection**  
If the input medium is “denser” \((n_1 > n_2)\), then these values are obtained:

\begin{align*}
n_1 &= 1.5 \\
n_2 &= 1.0 \\
\implies r_{TE} &= \frac{1.5 - 1.0}{1.5 + 1.0} = +0.2 \\
\implies r_{TM} &= \frac{1.0 - 1.5}{1.0 + 1.5} = -0.2 = 0.2e^{+i\pi} \\
\implies R_{TE} &= R_{TM} = 0.04 \\
\implies T_{TE} &= T_{TM} = 0.96
\end{align*}

There is no phase shift of the reflected amplitude in “dense-to-rare” reflection, commonly called “internal” reflection.

**8.1.6 Angular Dependence of Reflection and Transmittance at “Rare-to-Dense” Interface**

Consider the graphs of these coefficients for the cases of the “rare-to-dense” interface \((n_1 = 1 < n_2 = 1.5)\). The reflection coefficients are plotted vs. incident angle measured in degrees from 0\(^\circ\) (normal incidence) to 90\(^\circ\) (grazing incidence).
Amplitude reflectance and transmittance coefficients for $n_1 = 1.0$ (air) and $n_2 = 1.5$ (glass) for both TE and TM waves, plotted as functions of the incident angle from $\theta_0 = 0^\circ$ (normal incidence) to $\theta_0 = 90^\circ$ (grazing incidence). The reflectance coefficient $r_{TE} < 0$ for all $\theta$, which means that there is a phase shift upon reflection, whereas $r_{TM} > 0$ for $\theta_0 < \theta_B$ (Brewster’s angle). Also note that the transmittance coefficients are very similar functions.

**Brewster’s Angle – Angle of Complete Polarization**

Note that $r_{TM} = 0$ at one particular angle ($\approx 60^\circ$) in the TM case (parallel polarization), which means that no amplitude of this wave is reflected if incident at this angle. In other words, any light reflected at this angle must be the TE wave which is completely polarized perpendicular to the plane of incidence. This is Brewster’s angle, the angle such that the reflected wave and the refracted wave are orthogonal (i.e., $\theta_0 + \theta_t = \frac{\pi}{2} \implies \theta_t = \frac{\pi}{2} - \theta_0$). In this case, the electrons driven in the plane of the incidence will not emit radiation at the angle required by the law of reflection. This is sometimes called the angle of complete polarization. Note that the transmitted light contains both polarizations, though not in equal amounts.
Polarization of reflected light at Brewster’s angle. The incident beam at θ₀ = θ_B is unpolarized. The reflectance coefficient for light polarized in the plane (TM waves) is 0, and the sum of the incident and refracted angle is 90° = π/2. Thus

\[ \theta_B + \theta_t = \frac{\pi}{2} \implies \theta_t = \frac{\pi}{2} - \theta_B. \]

From Snell’s law, we have:

\[ n_1 \sin [\theta_1] = n_2 \sin [\theta_2] \]

At Brewster’s angle,

\[ n_1 \sin [\theta_B] = n_2 \sin \left[ \frac{\pi}{2} - \theta_B \right] \]
\[ = n_2 \left( \sin \left[ \frac{\pi}{2} \right] \cos [\theta_B] - \cos \left[ \frac{\pi}{2} \right] \sin [\theta_B] \right) \]
\[ = + n_2 \cos [\theta_B] \]

\[ n_1 \sin [\theta_B] = n_2 \cos [\theta_B] \]

\[ \implies \frac{n_2}{n_1} = \frac{\sin [\theta_B]}{\cos [\theta_B]} = \tan [\theta_B] \]
\[ \implies \theta_B = \tan^{-1} \left[ \frac{n_2}{n_1} \right] \]

If \( n_1 = 1 \) (air) and \( n_2 = 1.5 \) (glass), then \( \theta_B \simeq 56.3^\circ \). For incident angles larger than about 56°, the reflected light is plane polarized parallel to the plane of incidence. If the dense medium is water (\( n_2 = 1.33 \)), then \( \theta_B \simeq 52.4^\circ \). This happens at the interface with any dielectric. The reflection at Brewster’s angle provides a handy means to determine the polarization axis of a linear polarizer – just look through a linear polarizer at light reflected at a shallow angle relative to the surface (e.g., a waxed floor).

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

The reflectance and transmittance the two polarizations with \( n_1 = 1.0 \) and \( n_2 = 1.5 \) as functions of the incident angle \( \theta_0 \) show the zero reflectance of the TM wave at
Brewster’s angle.

Reflectance and transmittance for \( n_1 = 1.0 \) and \( n_2 = 1.5 \) for TE and TM waves. Note that \( R_{TM} = 0 \) and \( T_{TM} = 1 \) at one angle.

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

At a “glass-to-air” interface where \( n_1 > n_2 \), the reflectance of the TM wave (s polarization) is:

\[
r = \frac{-n_2 \cos \theta_0 + n_1 \cos \theta_t}{n_2 \cos \theta_0 + n_1 \cos \theta_t}
\]

The numerator evaluates to zero for a particular incident angle that satisfies:

\[
n_2 \cos \theta_0 = n_1 \cos \theta_t
\]

\[
\frac{n_1}{n_2} = \frac{\cos \theta_t}{\cos \theta_0}
\]

This corresponds to the situation where Snell’s law requires that:

\[
\sin \left( \theta_0 = \frac{\pi}{2} \right) = 1 = \frac{n_1}{n_2} \sin \theta_0 \implies \sin \theta_0 = \frac{n_2}{n_1}
\]

If \( n_1 = 1.5 \) and \( n_2 = 1.0 \), then

\[
\sin \theta_0 = \frac{2}{3} \implies \theta_0 \approx 0.73 \text{ radians} \approx 41.8^\circ \equiv \theta_c
\]

If the incident angle exceeds this value \( \theta_c \), the critical angle, then the amplitude reflectance coefficients \( r_{TE} \) and \( r_{TM} \) are both unity, and thus so are the reflectances \( R_{TE} \) and \( R_{TM} \). This means that light incident for \( \theta_0 \geq \theta_c \) is totally reflected. This is
the source of total internal reflectance (“internal” because the reflection is from glass back into glass). The phenomenon of TIR is the reason for the usefulness of optical fibers in communications.

The angular dependences of the amplitude reflection coefficients for the case $n_1 = 1.5$ (glass) and $n_2 = 1.0$ (air) are shown. Brewster’s angle in this case satisfies:

$$\theta_B = \tan^{-1} \left( \frac{n_2}{n_1} \right) = \tan^{-1} \left( \frac{1}{1.5} \right) \approx 33.7^\circ$$

Amplitude reflectance coefficients for TE and TM waves if $n_1 = 1.5$ (glass) and $n_2 = 1.0$ (air). Both coefficients rise to $r = +1.0$ at the “critical angle” $\theta_c$, for which $\theta_t = 90^\circ = \frac{\pi}{2}$. Also noted is Brewster’s angle, where $r_{TM} = 0$. The situation for $\theta_0 > \theta_c$ can be interpreted as producing complex-valued $r_{TE}$ and $r_{TM}$.

### 8.1.8 Practical Applications for Fresnel’s Equations

The 4% normal reflectance of one surface of glass is the reason why windows look like mirrors at night when you’re in the brightly lit room. Lasers incorporate end windows oriented at Brewster’s angle to eliminate reflective losses at the mirrors (and also thus producing polarized laser light). Optical fibers use total internal reflection. Hollow fibers use high-incidence-angle near-unity reflections.

### 8.2 Index of Refraction of Glass

We have already stated that the index of refraction $n$ relates the phase velocity of light in vacuum with that in matter:

$$n = \frac{c}{v_p} \geq 1.$$
In a dispersive medium, the index $n$ decreases with increasing $\lambda$, which ensures that the phase velocity $\frac{\omega}{k}$ (of the average wave) is larger than the group velocity $\frac{d\omega}{dk}$ (of the modulation wave).

Refraction is the result of the interaction of light with atoms in the medium and depends on wavelength because the refractive index is also; recall that the index decreases with increasing wavelength:

$$n = 1.5$$

$$n = 1.0$$

Typical dispersion curve for glass showing the decrease in $n$ with increasing $\lambda$ and the three spectral wavelengths used to specify “refractivity”, “mean dispersion”, and “partial dispersion”.

To a first approximation, the index of refraction varies as $\lambda^{-1}$, which allows us to write an empirical expression for the refractivity of the medium $n - 1$:

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

where $a$ and $b$ are parameters determined from measurements. The observation that the index decreases with increasing $\lambda$ determines that $b > 0$. Cauchy came up with an empirical relation for the refractivity more free parameters:

$$n[\lambda] - 1 \cong A \left( 1 + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \cdots \right)$$

Again, the behavior of normal dispersion ensures that $A$ and $B$ are both positive. Yet a better formula was proposed by Hartmann:

$$n[\lambda] \cong n_0 + \frac{\alpha}{(\lambda - \lambda_0)^{1.2}}$$

where $\alpha > 0$. The refractive properties of the glass are approximately specified by the refractivity and the measured differences in refractive index at the three Fraunhofer wavelengths $F$, $D$, and $C$: 

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

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

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

**For the crown glass:**
- refractivity: $n_D - 1 = 0.51666 \approx 0.517$
- Abbé number: $\nu = \frac{1.51666 - 1}{1.52225 - 1.51418} \approx 64.0$

**Glass number = 517640**

**For the flint glass:**
- refractivity: $n_D - 1 = 0.70100 \approx 0.701$
- Abbé number: $\nu = \frac{0.70100 - 1}{1.71748 - 1.69427} \approx 30.2$

**Glass number = 701302**
8.2.1 Optical Path Length

Because the phase velocity of light in a medium is less than that in vacuum, light takes longer to travel through a given thickness of material than through the same “thickness” of vacuum. For a fixed distance $d$, we know that:

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

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

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

$$\Rightarrow t_1 = \frac{t_2}{n} \Rightarrow t_2 > t_1$$

In the time $t_2$ required for light to travel the distance $d$ in a material of index $n$, light would travel a longer distance $nd = ct_2$ in vacuum. The distance $nd$ traveled in vacuum in the equivalent time is the optical path length in the medium.

8.3 Polarization

Maxwell’s equations demonstrated that light is a transverse wave (as opposed to longitudinal waves, e.g., sound). Both the $\mathbf{E}$ and $\mathbf{B}$ vectors are perpendicular to the direction of propagation of the radiation. Even before Maxwell, Thomas Young inferred the transverse character of light in 1817 when he passed light through a calcite crystal (calcium carbonate, $\text{CaCO}_3$). Two beams emerged from the crystal, which Young brilliantly deduced were orthogonally polarized, i.e., the directions of the $\mathbf{E}$ vectors of the two beams are orthogonal. The two components of an electromagnetic wave are the electric field $\mathbf{E} \left[ \frac{\text{V}}{\text{m}} \right]$ and the magnetic field $\mathbf{B} \left[ \text{tesla} = \frac{\text{webers}}{\text{m}^2} \right]$.

The polarization of radiation is defined as the plane of vibration of the electric vector $\mathbf{E}$, rather than of $\mathbf{B}$, because the effect of the $\mathbf{E}$ field on a free charge (an electron) is much greater than the effect of $\mathbf{B}$. This is seen from the Lorentz equation, or the Lorentz force law:

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

$q_0 =$ charge $\left[ \text{coulombs} \right]$  
$\mathbf{F} =$ force on the charge $\left[ \text{newtons}, \ 1 \text{N} = 1 \frac{\text{kg} \cdot \text{m}}{\text{s}^2} \right]$  
$\mathbf{v} =$ velocity of the charge $q_0$, measured in $\left[ \frac{\text{m}}{\text{s}} \right]$  
$c =$ velocity of light $\left[ 3 \cdot 10^8 \frac{\text{m}}{\text{s}} \right]$

The factor $c^{-1}$ ensures that the force on the electron due to the magnetic field is usually much smaller than the electric force.

8.3.1 Plane Polarization = Linear Polarization

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

$$\mathbf{E} = \mathbf{E} [\mathbf{r}, t] = [\hat{x} E_x + \hat{y} E_y] \cos [kz - \omega t]$$

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

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

For a wave of amplitude $E_0$ polarized at an angle $\theta$ relative to the $x$-axis:

$$E_x = E_0 \cos [\theta]$$

$$E_y = E_0 \sin [\theta]$$

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

\[\text{Electric field vector } \mathbf{E} \text{ and magnetic field vector } \mathbf{H} \text{ of a plane-polarized wave}\]

### 8.3.2 Circular Polarization

If the $\mathbf{E}$-vector describes a helical (i.e., screw-like) motion in space, the projection of the $\mathbf{E}$-vector onto a plane normal to the propagation direction $\mathbf{k}$ exhibits circular motion over time, hence the polarization is circular:
Circular polarization occurs when the electric fields along orthogonal axes have the same amplitude by their phases differ by $\pm \frac{\pi}{2}$ radians.

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

$$motion = \hat{x}\cos[\omega t] + \hat{y}\cos\left[\omega t \mp \frac{\pi}{2}\right] = \hat{x}\cos[\omega t] \pm \hat{y}\sin[\omega t]$$

For a traveling wave:

$$\mathbf{E} = [E_x, E_y] = \left[E_0 \cos[kz - \omega t], E_0 \cos\left[kz - \omega t \mp \frac{\pi}{2}\right]\right]$$

$$= \left[E_0 \cos[kz - \omega t], \pm E_0 \sin[kz - \omega t]\right]$$

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

### 8.3.3 Nomenclature for Circular Polarization

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

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

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

### 8.3.4 Elliptical Polarization, Reflections

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

### 8.3.5 Change of Handedness on Reflection

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

![Change in “handedness” of a circularly polarized wave upon reflection by a mirror.](image)

Natural Light

The superposition of emissions from a large number of thermal source elements (as in a light bulb) has a random orientation of polarizations. The state of polarization of the resulting light changes direction randomly over very short time intervals ($\approx 10^{-8}$ seconds). The radiation is termed unpolarized, even though it is polarized when viewed within this short time period. Natural light is neither totally polarized nor totally unpolarized; rather, we speak of partial polarization.
8.4 Description of Polarization States

8.4.1 Jones Vector

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

\[ \mathbf{E} = \text{Re}\{E_0 e^{i(kz - \omega t)}\} = \text{Re}\{E_x e^{i(kz - \omega t)} \}, \text{Re}\{E_y e^{i(kz - \omega t - \delta)}\} \]
\[ = \text{Re}\{[E_x, E_y e^{-i\delta}]e^{i(kz - \omega t)}\} \]
\[ \implies \text{Jones Vector} \quad \mathbf{E} = \begin{bmatrix} E_x \\ E_y e^{-i\delta} \end{bmatrix} \]

Examples:

1. Plane-polarized light along x-axis

\[ \mathbf{E} = \begin{bmatrix} E_0 \\ 0 \end{bmatrix} \]

2. Plane-polarized light along y-axis:

\[ \mathbf{E} = \begin{bmatrix} 0 \\ E_0 \end{bmatrix} \]

3. Plane-polarized light at angle \( \theta \) to x-axis:

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

4. RHCP

\[ \mathbf{E} = \hat{x} E_0 \cos [kz - \omega t] + \hat{y} E_0 \sin [kz - \omega t] \]
\[ = \hat{x} E_0 \cos [kz - \omega t] + \hat{y} E_0 \cos \left[kz - \omega t - \frac{\pi}{2}\right] \]
\[ = \text{Re}\left\{ \begin{bmatrix} E_0 \\ E_0 \exp \left[-\frac{in\pi}{2}\right] \end{bmatrix} e^{i[kz - \omega t]} \right\} \implies \mathbf{E} = \text{Re}\left\{ \begin{bmatrix} 1 \\ \exp \left[-\frac{in\pi}{2}\right] \end{bmatrix} e^{i[kz - \omega t]} \right\} \]

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

### 8.5 Generation of Polarized Light

#### 8.5.1 Selective Emission:

If all emitting elements of a source (e.g., electrons in a bulb filament), vibrate in the same direction, the radiated light will be polarized in that direction. This is difficult to achieve at optical frequencies ($\Delta t \lesssim 10^{-14} \text{ s} \implies \nu \gtrsim 10^{14} \text{ Hz}$), but is easy at radio or microwave frequencies ($\nu \lesssim 10^8 \text{ Hz}$) by proper design of the antenna that radiates the energy. For example, a radio-frequency oscillator attached to a simple antenna forces the free electrons in the antenna to oscillate along the long (vertical) dimension of the antenna. The emitted radiation is therefore mostly oscillating in the vertical direction; it is vertically polarized.

"Light" (electromagnetic radiation) emitted by a "dipole" radiator is polarized in the direction of motion of the emitting electrons (vertical, in this case).

Rather than generating polarized light at the source, we can obtain light of a selected polarization from natural light by removing unwanted states of polarization. This is the mechanism used in the next section.

#### 8.5.2 Selective Transmission or Absorption

A man-made device for selecting a state of polarization by selective absorption is *Polaroid*. This operates like the microwave-polarizing skein of wires. The wires are parallel to the $y$-axis in the figure. Radiation incident on the wires drives the free electrons in the wires in the direction of polarization of the radiation. The electrons driven in the $y$-direction along the surface of the wire and strike other such electrons, thus dissipating the energy in thermal collisions. What energy that is reradiated by such electrons is mostly directed back toward the source (reflected). The $x$-component
of the polarization is not so affected, since the electrons in the wire are constrained against movement in that direction. The $x$-component of the radiation therefore passes nearly unaffected.

Common Polaroid sheet acts as a skein of wires for optical radiation. It is made from clear polyvinyl acetate which has been stretched in one direction to produce long chains of hydrocarbon molecules. The sheet is then immersed in iodine to supply lots of free electrons.

Polarization by “skein of wires” – the radiation polarized parallel to the direction of the wires in the skein is absorbed, so the radiation polarized perpendicular to the wires is transmitted.

8.5.3 Generating Polarized Light by Reflection – Brewster’s Angle

The two polarizations of light reflected from an interface between two different dielectric media (i.e., media with different real refractive indices) see the same configuration of the interface only with normal incidence (i.e., the light is incident perpendicular to the surface). Thus the two polarizations must be identically reflected. However, if the light is incident obliquely, one polarization “sees” the bound electrons of the surface differently and therefore is reflected differently. The reflected wave is polarized to some extent; the amount of polarization depends on the angle of incidence and the index of refraction $n$. The polarization mechanism is simply pictured as a forced electron oscillator. The bound electrons in the dielectric material are driven by the incident oscillating electric field of the radiation $\mathbf{E} \exp [i (k_0 z_0 \pm \omega_0 t)]$, and hence vibrate at frequency $\nu_0 = \frac{\omega_0}{2\pi}$. Due to its acceleration, the vibrating electron reradiates radiation at the same frequency $\nu$ to produce the reflected wave. The state of polarization of the reflected radiation is a function of the polarization state of the incident wave, the angle of incidence, and the indices of refraction on either side of the interface. If the reflected wave and the refracted wave are orthogonal (i.e., $\theta_0 + \theta_t = 90^\circ \implies \theta_t = \frac{\pi}{2} - \theta_0$), then the reflected wave is completely plane polarized parallel to the surface (and thus polarized perpendicular to the plane of incidence). This angle appeared in the discussion of the reflectance coefficients in the previous section. In this case, the electrons driven in the plane of the incidence will not emit radiation at the angle required by the law of reflection. This angle of complete polarization is called Brewster’s Angle $\theta_B$, which we mentioned earlier during
8.5 GENERATION OF POLARIZED LIGHT

the discussion of the Fresnel equations.

Brewster’s angle: the incident beam at $\theta_0 = \theta_B$ is unpolarized. The reflectance coefficient for light polarized in the plane (TM waves) is 0, and the sum of the incident and refracted angle is $90^\circ = \frac{\pi}{2}$. Thus $\theta_B + \theta_t = \frac{\pi}{2} \implies \theta_t = \frac{\pi}{2} - \theta_B$.

At Brewster’s angle,

$$n_1 \sin [\theta_B] = n_2 \sin \left[ \frac{\pi}{2} - \theta_B \right]$$

$$\implies \theta_B = \tan^{-1} \left[ \frac{n_2}{n_1} \right]$$

If $n_1 = 1$ (air) and $n_2 = 1.5$ (glass), then $\theta_B \approx 56.3^\circ$. For incident angles larger than about 56°, the reflected light is plane polarized parallel to the plane of incidence. If the dense medium is water ($n_2 = 1.33$), then $\theta_B \approx 52.4^\circ$. This happens at the interface with any dielectric. The reflection at Brewster’s angle provides a handy means to determine the polarization axis of a linear polarizer – just look through the polarizer at light reflected at a steep angle.

8.5.4 Polarization by Scattering

Light impinging on an air molecule drives the electrons of the molecule in the direction of vibration of the electric field vector. This motion causes light to be reradiated in a dipole pattern; i.e., no light is emitted along the direction of electron vibration. If we look at scattered light (e.g., blue sky) at 90° from the source, the light is completely linearly polarized. Note that if the light is multiply scattered, as in fog, each scattering disturbs the state of polarization and the overall linear state is perturbed into unpolarized radiation.
Many natural crystals and man-made materials interact with the two orthogonal polarizations differently. This is often due to an anisotropy (nonuniformity) in the crystalline structure; such materials are called *dichroic* or *birefringent*. Many crystals (e.g., calcite) divide a nonpolarized light wave into two components with orthogonal polarizations. The two indices of refraction are sometimes denoted $n_f$ and $n_s$ for *fast* and *slow* axes, where $n_f < n_s$. They are also denoted $n_o$ and $n_e$ for *ordinary* and *extraordinary* axes. The ordinary ray obeys Snell’s law; the extraordinary ray does not. One is called the *ordinary ray*, because it obeys Snell’s law for refraction. The second, or *extraordinary ray*, does not obey Snell. By dividing the incoming natural light into two beams in such a crystal, we can select one of the two polarizations.

### 8.6.1 Examples:

Refractive indices along the fast and slow axes at $\lambda = 589.3 \text{nm}$

<table>
<thead>
<tr>
<th>Material</th>
<th>$n_o$</th>
<th>$n_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcite ($CaCO_3$)</td>
<td>1.6584</td>
<td>1.4864</td>
</tr>
<tr>
<td>Crystalline Quartz ($SiO_2$)</td>
<td>1.5534</td>
<td>1.5443</td>
</tr>
<tr>
<td>Ice (crystalline $H_2O$)</td>
<td>1.313</td>
<td>1.309</td>
</tr>
<tr>
<td>Rutile ($TiO_2$)</td>
<td>2.903</td>
<td>2.616</td>
</tr>
<tr>
<td>Sodium Nitrate ($SiNO_3$)</td>
<td>1.5854</td>
<td>1.3369</td>
</tr>
</tbody>
</table>
The wavelength of light in a medium is \( \lambda' = \frac{\lambda}{n} \), so light along the two polarization directions have different wavelengths:

\[
\lambda'_s = \frac{\lambda}{n_s} < \lambda'_f = \frac{\lambda}{n_f}
\]

### 8.6.2 Phase Delays in Birefringent Materials — Wave Plates

Consider light incident on a birefringent material of thickness \( d \). The electric field as a function of distance \( z \) and time \( t \) is:

\[
\mathbf{E}[z,t] = (\hat{x}E_x + \hat{y}E_y) \ e^{i(kz-\omega t)}.
\]

At the input face of the material \( (z = 0) \) and the output face \( (z = d) \), the fields are:

\[
\mathbf{E}[z = 0, t] = (\hat{x}E_x + \hat{y}E_y) \ e^{-i\omega t}
\]

\[
\mathbf{E}[z = d, t] = (\hat{x}E_x + \hat{y}E_y) \ e^{i(kd-\omega t)}
\]

If \( n_x = n_s > n_y = n_f \), then \( \lambda_f > \lambda_s \) and:

\[
k_s = k_x = \frac{2\pi n_s}{\lambda} > k_f = k_y = \frac{2\pi n_f}{\lambda}
\]

The field at the output face \( (z = d) \) is therefore:

\[
\mathbf{E}[d, t] = \left( \hat{x}E_x \exp \left[ +i \left( \frac{2\pi d \cdot n_s}{\lambda} \right) \right] + \hat{y}E_y \exp \left[ +i \left( \frac{2\pi d \cdot n_f}{\lambda} \right) \right] \right) \ e^{-i\omega t}
\]

By defining a constant phase term \( \delta \equiv \frac{2\pi d}{\lambda} (n_f - n_s) \), the electric field at the output face of the birefringent material can be expressed as:

\[
\mathbf{E}[d, t] = (\hat{x}E_x + \hat{y}E_y e^{i\delta}) \ \exp \left[ +i \frac{2\pi d n_f}{\lambda} \right]
\]

On emergence from the material, the \( y \)-component of the polarization has a different phase than the \( x \)-component; the phase difference is \( \delta \).

**Example:**

\( \delta = +\frac{\pi}{2} \Rightarrow (n_f - n_s) d = -\frac{\lambda}{2} \), and there is a phase difference of one quarter wavelength between the polarizations of the \( x \)- and the \( y \)-components of the wave. This is a *quarter-wave plate*. The required thickness \( d \) of the material is:

\[
d = \frac{\lambda}{4(n_s - n_f)}
\]
And the emerging field is:

\[
\mathbf{E}(d, t) = \left[ \hat{x}E_x + \hat{y}E_y e^{+i\phi} \right] \exp \left[ i \left( k \cdot d - \omega t \right) \right]
\]

If \( E_x = E_y \), (i.e., the incident wave is linearly polarized @ 45° to the x-axis), then the emerging wave is circularly polarized. This is the principle of the circular polarizer.

**Example:**

If \( \delta = +\pi \implies d = \frac{\lambda}{2(n_s - n_f)} \), and the relative phase delay is 180°. Such a device is a half-wave plate. If the incident light is linearly polarized along the orientation midway between the fast and slow axes, the plane of polarization of the exiting linearly polarized light is rotated by 90°.

### 8.6.3 Circular Polarizer:

A circular polarizer is a sandwich of a linear polarizer and a \( \frac{\lambda}{4} \) plate, where the polarizing axis is oriented midway between the fast and slow axes of the quarter-wave plate. The LP ensures that equal amplitudes exist along both axes of the quarter-wave plate, which delays one of the components to create circularly polarized light. Light incident from the back side of a circular polarizer is not circularly polarized on exit; rather it is linearly polarized. A circular polarizer can be recognized and properly oriented by placing it on a reflecting object (e.g., a dime). If the image of the coin is dark, the polarizer has the linear polarizer on top. This is because the handedness of the light is changed on reflection; the light emerging from the \( \frac{\lambda}{4} \) plate is now linearly polarized perpendicular to the axis of the LP and no light escapes.

\[ \text{Unpolarized Light} \rightarrow \text{LP} \rightarrow \text{Unpolarized Light} \rightarrow \lambda/4 \rightarrow \text{LHCP} \]

*A circular polarizer is a sandwich of a linear polarizer and a quarter-wave plate.*
8.7 Critical Angle—Total Internal Reflection

We also mentioned this phenomenon during the discussion of the Fresnel equations. From Snell, we have the relation:

\[ n_1 \sin[\theta_1] = n_2 \sin[\theta_2] \]

If \( n_1 > n_2 \) then a specific angle \( \theta_1 \) satisfies the condition:

\[ \frac{n_1}{n_2} \sin[\theta_1] = 1 \implies \sin[\theta_1] = \frac{n_2}{n_1} < 1 \implies \theta_2 = \frac{\pi}{2} \]

which means that the outgoing ray is refracted parallel to the interface ("surface"). The incident angle \( \theta_1 \) that satisfies this condition is the critical angle \( \theta_c \)

\[ \theta_c = \sin^{-1}\left[ \frac{n_2}{n_1} \right] \]

For crown glass with \( n_d = 1.52 \), the critical angle is \( \sin^{-1}\left[ \frac{1}{1.52} \right] \approx 0.718 \) radians \( \cong 41^\circ \). For a common flint glass with \( n_d = 1.70 \), then \( \theta_c \approx 0.629 \) radians \( \cong 36^\circ \). If the incident angle \( \theta_1 > \theta_c \) and \( n_1 > n_2 \) (e.g., the first medium is glass and the second is air), then no real-valued solution for Snell’s law exists, and there is no refracted light. This is the well-known phenomenon of total internal reflection — all of the incident light is reflected at the interface.

This may be analyzed rigorously by applying Maxwell’s equations to show that the refracted angle \( \theta_2 \) is complex valued instead of real valued, so that the electromagnetic field is attenuated exponentially as it crosses the interface. In other words, the electric
field decays so rapidly across the interface that no energy can flow across the boundary, and hence no light escapes. However, we can “frustrate” the total internal reflection by placing another medium (such as another piece of glass) within a few light wavelengths of the interface. If close enough to the boundary, then some electric field can get into the second glass and a refracted wave “escapes”.

\[ \theta_1 - \theta_1 - \tau \approx 1 - 2\lambda \]

*Schematic of “frustrated total internal reflection”:* some energy can “jump” across a small gap between two pieces of glass even though the incident angle exceeds the critical angle. As the width \( \tau \) of the gap increases, then the quantity of energy coupled across the gap decreases very quickly.
Chapter 9

Optical Imaging

9.1 Transition from Wave Optics to Ray Optics

We have mentioned that the rigorous evaluation of light upon interaction with matter (as in diffracting apertures, mirrors, or lenses) involves the solution of the four equations collected by Maxwell subject to the specific conditions of the problem. The exact solution to these equations is often very difficult to obtain. Fortunately, it is often sufficient to find approximate solutions. The very useful approximation relevant to imaging applications is the model of light as rays, which is generally called geometrical optics. This approximation emphasizes the path travelled by light through media to find the locations and sizes of images. The other model of optics as waves, called physical optics, emphasizes the “deviation” or “spreading” of light from the geometrical paths to create interference and/or diffraction, and demonstrates the fundamental limitations on the performance (i.e., the resolution) of optical imaging systems.

Ray: a line in space that maps the direction of energy flow. It is a mathematical construction, not an actual entity.

The geometrical optics approximation is a limiting case of the more general wave optical model in the limit that the wavelength $\lambda$ of light goes to zero. As we shall see later, there is no diffraction in the wave model in this case.

9.1.1 Notational Conventions

One of the more confusing and frustrating aspects of geometrical optics is the existence of multiple notational conventions. These notes use the convention of directed distances, which is also used by Halliday and Resnick, Jenkins and White, Hecht, Nussbaum and Phillips, Crawford, Iizuka, Goodman, and Gaskill. The other common convention is based on a coordinate system with the origin at the first vertex of the optical system, and is used by many authors for lens design, e.g., Born and Wolf, Warren Smith, and Ditchburn.

A powerful advantage of the directed distance convention is the resulting (and pleasing) symmetry between objects and images, and because the nature of the resulting images is obvious.
1. Light travels from left to right (positive direction);

2. Interfaces (i.e., lens or mirror surfaces) are numbered from the first to the last encountered by the ray;

3. Distances are measured from lens vertices, the intersection of the lens surface with the axis of symmetry (optical axis);

4. A horizontal distance is positive if measured from left to right;

5. A vertical distance from the axis is positive if measured “up”;

6. Angles are measured from the optical axis or from the normal to the surface and are positive if in the counterclockwise direction (the normal convention for $\theta$);

7. A radius of curvature is positive if the center is to the right of the vertex;

8. A subscript on a quantity corresponds to the surface with which is it associated;

9. If used, primed quantities (e.g., $n'$) refer to the “outgoing” side of an interface. These are useful when describing a multiple element system where the output (image) space for one element is the input (object) space for the next element.
9.1 TRANSITION FROM WAVE OPTICS TO RAY OPTICS

9.1.2 Fermat’s Principle

Hero of Alexandria hypothesized the model of light propagation that could be called the *principle of least distance*:

*A ray of light traveling between two arbitrary points traverses the shortest possible path in space.*

This statement applies to reflection and transmission through homogeneous media (i.e., the medium is characterized by a single index of refraction). However, Fermat’s principle is not valid if the object and observation points are located in different media (i.e., the normal situation for refraction) or if multiple media are present between the points.

In 1657, Pierre Fermat modified Hero’s statement to formulate the *principle of least time*:

*A light ray travels the path that requires the least time to traverse.*

The laws of reflection and refraction may be easily derived from Fermat’s principle. A moving ray (or car, bullet, or baseball) traveling a distance $s$ at a velocity $v$ requires $t$ seconds:

$$t = \frac{s}{v}$$

If the ray travels at different velocities for different increments of distance, the travel time may be written as:

$$t = \sum_{m=1}^{M} \frac{s_m}{v_m}$$

We know that the velocity of a light ray in a medium of index $n$ is $v = \frac{c}{n}$, so that:

$$t = \sum_{m=1}^{M} \frac{s_m}{n_m^2} = \frac{1}{c} \sum_{m=1}^{M} (n_m s_m) = \frac{\ell}{c}$$

Thus the time traveled to traverse the path through the medium is equal to the time required to travel a longer path $\ell$ in vacuum; the path is longer because $n_m \geq 1$. This longer path $\ell = ns$ is called the *optical path length*. This means that the light requires the least time to traverse the path with the shortest optical path length. The principle of least time may be reworded as:

*A ray traverses the route with the shortest optical path length.*

This result may be derived from Maxwell’s equations.

**Fermat’s Principle for Reflection**

Now consider the path traveled upon reflection that minimizes an easily evaluated optical path length:
As drawn, the angle $\theta_1$ is positive (measured from the normal to the ray) and $\theta_2$ is negative (from the normal to the ray). The ray travels in the same medium of index $n$ both before and after reflection. The components of the optical path length are:

\[
so = \sqrt{h^2 + x^2} \\
op = \sqrt{b^2 + (a - x)^2}
\]

And the expression for the total optical path length $\ell$ is:

\[
\ell = n (so + op) = n \left( \sqrt{h^2 + x^2} + \sqrt{b^2 + (a - x)^2} \right)
\]

which is a function of $x$

By Fermat’s principle, the path length traveled is the minimum of the optical path length $\ell$, so the position of $o$ may be find by setting the derivative of $\ell$ with respect to $x$ to zero:

\[
\frac{d\ell}{dx} = n \left( \frac{2x}{2\sqrt{h^2 + x^2}} + \frac{-2(a - x)}{2\sqrt{b^2 + (a - x)^2}} \right) = 0
\]

\[
\frac{x}{\sqrt{h^2 + x^2}} = \frac{a - x}{\sqrt{b^2 + (a - x)^2}}
\]

\[
\Rightarrow \frac{x}{\sqrt{h^2 + x^2}} = \frac{a - x}{\sqrt{b^2 + (a - x)^2}}
\]
from the drawing, note that
\[
\sin [\theta_1] = \frac{x}{\sqrt{h^2 + x^2}}
\]
\[
\sin [-\theta_2] = \frac{a - x}{\sqrt{b^2 + (a - x)^2}}
\]
\[
\Rightarrow \sin [\theta_1] = \sin [-\theta_2]
\]
\[
\Rightarrow -\theta_1 = \theta_2
\]

In words, the magnitudes of the angles of incidence and reflection are equal (as already derived by evaluating Maxwell’s equations at the boundary). The negative sign is necessary because of the sign convention for the angle; the angle is measured from the normal and increases in the counterclockwise direction.

**Fermat’s Principle for Refraction:**

\[
\begin{align*}
&\text{s} \quad \text{h} \quad \sqrt{h^2 + x^2} \quad \theta_1 \quad \theta_1 \\
&\text{a} \quad \text{O} \quad \text{x} \quad \text{a} - \text{x} \\
&\sqrt{b^2 + (a - x)^2} \quad \theta_2 \quad \theta_2 \\
&\text{p} \quad \text{n}_2 \quad \text{n}_1
\end{align*}
\]

*Schematic for refraction using Fermat’s principle.*

In this drawing, both \(\theta_1\) and \(\theta_2\) are positive (measured from the normal to the ray). The optical path length is:

\[
\ell = n_1 \overrightarrow{SO} + n_2 \overrightarrow{OP}
\]
\[
= n_1 \sqrt{h^2 + x^2} + n_2 \sqrt{b^2 + (a - x)^2}
\]
By Fermat’s principle, the path length traveled is the minimum of \( \ell \), so we again set the derivative of \( \ell \) with respect to \( x \) to zero:

\[
\frac{d\ell}{dx} = n_1 \frac{2x}{2\sqrt{h^2 + x^2}} + n_2 \frac{-2(a-x)}{2\sqrt{b^2 + (a-x)^2}} = 0
\]

\[
\Rightarrow n_1 \frac{x}{\sqrt{h^2 + x^2}} = n_2 \frac{a-x}{\sqrt{b^2 + (a-x)^2}} = 0
\]

\[
\sin[\theta_1] = \frac{x}{\sqrt{h^2 + x^2}}
\]

\[
\sin[\theta_2] = \frac{a-x}{\sqrt{b^2 + (a-x)^2}}
\]

\[
\Rightarrow n_1 \sin[\theta_1] = n_2 \sin[\theta_2]
\]

\[
\Rightarrow \text{Snell’s Law for refraction}
\]

Note that with this sign convention, Snell’s law may be applied to reflection by setting the refractive index of the second medium to be the negative of the first:

\[
n_1 \sin[\theta_1] = n_2 \sin[\theta_2]
\]

\[
\Rightarrow n_1 \sin[\theta_1] = -n_1 \sin[\theta_2]
\]

\[
\Rightarrow -\sin[\theta_1] = \sin[\theta_2]
\]

\[
\Rightarrow \theta_2 = -\theta_1
\]
Chapter 10

Image Formation in the Ray Model

We know that light rays are deviated at interfaces between media with different refractive indices. The goal in this section is to use interfaces of specified shapes to “collect” and “redirect” rays (or to “collect” and “reshape” wavefronts that are normal to the family of rays) in such a way to create “images” of the original source(s).

10.1 Refraction at a Spherical Surface

Consider a point source located at \( o \). The distance to the vertex \( v \) is \( ov \equiv s_1 > 0 \) as drawn. The distance from vertex \( v \) to the point \( p \) is \( vp \equiv s_2 > 0 \). The distance traveled by a ray in medium \( n_1 \) to the surface is \( ov \equiv \ell_1 \) and the distance in medium \( n_2 \) is \( ap \equiv \ell_2 \). The radius of curvature of the surface is \( vc = ac \equiv R > 0 \). For emphasis, we repeat that \( s_1, s_2, \) and \( R \) are all positive in our convention. The ray intersects the surface at the “position angle” \( \varphi \) measured from the center of curvature.
c to a. The optical path length of the ray from o to p that passes through a is
\[ OPL = n_1 \ell_1 + n_2 \ell_2. \]
We will evaluate this length by considering the triangles \( \triangle oac \) and \( \triangle acp \); the hypotenuses \( oc \) and \( ap \) may be evaluated by applying the law of cosines:
\[
\triangle oac \Rightarrow |oa|^2 = |oc|^2 + |ac|^2 - 2 |oc||ac| \cos \varphi
\]
\[
\ell_1^2 = (s_1 + R)^2 + R^2 - 2R (s_1 + R) \cos \varphi
\]
\[
\ell_1 = \sqrt{(s_1 + R)^2 + R^2 - 2R (s_1 + R) \cos \varphi}
\]
\[
\triangle acp \Rightarrow |ap|^2 = |ac|^2 + |cp|^2 - 2 |ac||cp| \cos [\pi - \varphi]
\]
\[
\ell_2^2 = (s_2 - R)^2 + R^2 - 2R (s_2 - R) \cos [\pi - \varphi]
\]
\[
\ell_2 = \sqrt{(s_2 - R)^2 + R^2 - 2R (s_2 - R) \cos \varphi}
\]
Therefore the optical path length is:
\[
OPL = n_1 \ell_1 + n_2 \ell_2
\]
\[
= n_1 \cdot \left( \sqrt{(s_1 + R)^2 + R^2 - 2R (s_1 + R) \cos \varphi} \right)
\]
\[
+ n_2 \cdot \left( \sqrt{(s_2 - R)^2 + R^2 - 2R (s_2 - R) \cos \varphi} \right)
\]
which is obviously a function of the “position angle” $\varphi$. We can now apply Fermat’s principle and evaluate the angle $\varphi$ for which the $OPL$ is a minimum:

$$\frac{d}{d\varphi} (OPL) = 0$$

$$= \frac{n_1 \cdot 2R(R + s_1) \sin [\varphi]}{\sqrt{(s_1 + R)^2 + R^2 - 2R(s_1 + R) \cos [\varphi]}} + \frac{n_2 \cdot 2R(R - s_2) \sin [\varphi]}{\sqrt{(s_2 - R)^2 + R^2 - 2R(R - s_2) \cos [\varphi]}}$$

$$= 2R \sin [\varphi] \frac{n_1 (R + s_1)}{\sqrt{(s_1 + R)^2 + R^2 - 2R(R + s_1) \cos [\varphi]}} + 2R \sin [\varphi] \frac{n_2 (R - s_2)}{\sqrt{(s_2 - R)^2 + R^2 - 2R(R - s_2) \cos [\varphi]}}$$

$$= 2R \sin [\varphi] \left( \frac{n_1 (R + s_1)}{\ell_1} + \frac{n_2 (R - s_2)}{\ell_2} \right)$$

$$\Rightarrow \frac{n_1 (R + s_1)}{\ell_1} + \frac{n_2 (R - s_2)}{\ell_2} = 0$$

$$\Rightarrow \frac{n_1 R}{\ell_1} + \frac{n_2 R}{\ell_2} = \frac{n_2 s_2}{\ell_2} - \frac{n_1 s_1}{\ell_1}$$

$$\Rightarrow \frac{n_1}{\ell_1} + \frac{n_2}{\ell_2} = \frac{1}{R} \left( \frac{n_2 s_2}{\ell_2} - \frac{n_1 s_1}{\ell_1} \right)$$

for one spherical surface

This relation between the physical path lengths $\ell_1$ and $\ell_2$ and the distances $s_1$ and $s_2$ is exact. Now we now identify the ratio of the physical path length $\ell_1$ from $o$ to $a$ to the axial distance $s_1$ from $o$ to the surface vertex $v$:

$$\frac{\ell_1}{s_1} = \frac{s_1}{\sqrt{(s_1 + R)^2 + R^2 - 2R(s_1 + R) \cos [\varphi]}}$$

$$= \left( \frac{(s_1 + R)^2 + R^2 - 2R(s_1 + R) \cos [\varphi]}{s_1^2} \right)^{\frac{1}{2}}$$

$$= \left( \frac{s_1^2 + R^2 + 2Rs_1 + R^2 - 2R^2 \cos [\varphi] - 2Rs_1 \cos [\varphi]}{s_1^2} \right)^{\frac{1}{2}}$$

$$= \left( 1 + \frac{2R^2}{s_1^2} + \frac{2R}{s_1} \left( 1 - \cos [\varphi] \right) \right)^{\frac{1}{2}}$$

This relation also is exact.
10.1.1 Paraxial Approximation

We can expand the cosine of the position angle into its Taylor series:

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

If the angle $\varphi$ is very small (so that the ray remains close to the axis), we can approximate the cosine by truncating the series after the first term:

$$\text{if } \varphi \approx 0 \implies \cos [\varphi] \approx 1 \implies 1 - \cos [\varphi] \approx 0$$

$$\implies \frac{\ell_1}{s_1} \approx 1 \implies \ell_1 \approx s_1, \text{ and } \ell_2 \approx s_2$$

In this, the *paraxial approximation* (the resulting approximate calculations are called *first-order optics* or *Gaussian optics*), we have:

$$\frac{1}{R} \left( \frac{n_2 s_2}{\ell_2} - \frac{n_1 s_1}{\ell_1} \right) \approx \frac{1}{R} (n_2 - n_1)$$

$$\implies \frac{n_1}{s_1} + \frac{n_2}{s_2} \approx \frac{1}{R} (n_2 - n_1)$$

paraxial imaging equation for single spherical surface

Snell’s Law in the Paraxial Approximation

Recall Snell’s law that relates the ray angles before and after refraction:

$$n_1 \sin [\theta_1] = n_2 \sin [\theta_2]$$

In the paraxial approximation where $\theta \approx 0$, the refraction equation simplifies to:

$$n_1 \theta_1 = n_2 \theta_2 \implies \frac{\theta_2}{\theta_1} = \frac{n_1}{n_2} \implies \theta_2 = \frac{n_1}{n_2} \theta_1$$

In words, the ray angle after refraction is proportional to that before refraction.

Power

The value of a lens or lens system is due to its ability to “redirect” rays by changing their direction. This capability is described by the *power* of the lens; a lens system with a large power changes the angles of rays by a large amount. A lens system that does not change the angles of rays has zero power. Most people are more familiar with the concept of “focal length”, which is the reciprocal of the power. The power is measured in terms of the reciprocal length; if measured in meters, the units of power are *diopters*. 
Object- and Image-Space Focal Lengths

Now consider some pairs of object and image distances $s_1$ and $s_2$. If the object is located at $-\infty$, then:

$$\frac{n_1}{\infty} + \frac{n_2}{s_2} = \frac{n_2}{s_2} \approx \frac{1}{R} (n_2 - n_1)$$

$$\Rightarrow s_2 \approx n_2 R \frac{n_2}{n_2 - n_1} \equiv f_2$$

the “image-space focal length” of the single surface

If the image is located at $+\infty$, we have:

$$s_1 \approx \frac{n_1 R}{n_2 - n_1} \equiv f_1$$

the “object-space focal length” of the single surface

Also note that:

$$\frac{f_1}{f_2} = \left( \frac{n_1 R}{n_2 - n_1} \right) \left( \frac{n_2 - n_1}{n_2 R} \right) = \frac{n_1}{n_2}$$

In words, the ratio of the object-space and image-space focal lengths of the single surface between two media equals the ratio of the indices of refraction.

To summarize, the assumptions of paraxial optics reduce the exact trigonometric expressions for ray heights and ray angles to zero. The resulting expressions are accurate ONLY within an infinitesimal region centered on the optical axis of symmetry (the imaginary line through the centers of curvature of the surfaces of the system). Though limited in its descriptive accuracy of the properties of the system, the paraxial approximation results in a set of simple equations that are accurate for locating the axial positions of images. However, because extended objects consist of point sources at various positions in space distant from the optical axis (and thus do not fulfill the requirements of the paraxial approximation), the quality of the image cannot be inferred from this description. “Defects” in the image are due to “aberrations” in the optical system, which are deviations from the paraxial approximation where the output angle is proportional to the input angle.

10.1.2 Nature of Objects and Images:

1. **Real Object**: Rays incident on the lens are diverging from the source; the object distance is positive:
2. **Virtual Object**: Rays converge toward the “source”, which is “behind” the lens; **object distance is negative**:

3. **Real Image**: Rays converge from the lens toward the image; **image distance is positive**:

4. **Virtual Image**: Rays diverge from the lens, so that the “image” is behind the lens; the **image distance is negative**:
10.2 Imaging With Lenses

Normally we do not consider the case of an object in one medium with the image in another – usually both object and image are in air and a lens (a “device” with another index and two usually curved surfaces) forms the image. We can derive the formula for the object and image distances if we know the radii of the lens surfaces and the indices of refraction. We merely cascade the paraxial formulas for a single surface:

At first surface: \[ \frac{n_1}{s_1} + \frac{n_2}{s'_1} = \frac{n_2 - n_1}{R_1} \]

At second surface: \[ \frac{n_2}{s_2} + \frac{n_3}{s'_2} = \frac{n_3 - n_2}{R_2} \]

where \( s_1 \) is the (usually known) object distance, \( s'_1 \) is the image distance for rays refracted by the first surface, \( s_2 \) is the object distance for the second surface, and \( s'_2 \) is the image distance for rays exiting the second surface (and thus from the lens). For the common “convex-convex” lens, the center of curvature of the lens is to the right of the vertex, and thus the radius \( R_1 \) of the first surface is positive. Since the vertex is to the right of the center of curvature of the second surface, \( R_2 < 0 \). If the lens is “thin”, then the ray encounters the second surface immediately after refraction at the first surface, so the magnitude of the image distance for the front surface \( |s'_1| \) is the same as the object distance for the second surface \( |s_2| \). From the directed distance convention, \( s_2 = -s'_1 \). If the lens is “thick”, then \( |s'_1| \neq |s_2| \), so we define the thickness \( t \) to satisfy:

\[ s'_1 + s_2 = t \implies s'_1 = t - s_2 \]

To find the single imaging equation, we add the equations for the two surfaces:

\[ \left( \frac{n_1}{s_1} + \frac{n_2}{s'_1} \right) + \left( \frac{n_2}{s_2} + \frac{n_3}{s'_2} \right) = \left( \frac{n_2 - n_1}{R_1} \right) + \left( \frac{n_3 - n_2}{R_2} \right) \]

\[ = \frac{n_3}{R_2} + n_2 \left( \frac{1}{R_1} - \frac{1}{R_2} \right) - \frac{n_1}{R_1} \]

For a thin lens \( (t = 0) \), substitute \(-s_2 \) for \( s'_1 \)

\[ \left( \frac{n_1}{s_1} + \frac{n_2}{-s_2} \right) + \left( \frac{n_2}{s_2} + \frac{n_3}{s'_2} \right) = \frac{n_1}{s_1} + \frac{n_3}{s'_2} \]

\[ = \frac{n_3}{R_2} + n_2 \left( \frac{1}{R_1} - \frac{1}{R_2} \right) - \frac{n_1}{R_1} \]

where the object is in index \( n_1 \), the lens has index \( n_2 \), and the image is in index \( n_3 \).

In the usual case, the object and image are both in air, so that \( n_3 = n_1 = 1 \). The simplified expression of the power of a thin lens is encapsulated in the so-called
lensmaker’s equation:

\[
\frac{1}{s_1} + \frac{1}{s'_2} = \frac{1}{R_2} + n_2 \left( \frac{1}{R_1} - \frac{1}{R_2} \right) - \frac{1}{R_1}
\]

\[\varphi = \frac{1}{f} = (n_2 - 1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right)\]

which defines the focal length of a lens with known index and surface radii. Note that the object distance \(s_1\) and the image distance \(s'_2\) both “appear” on the left side with the same algebraic sign; this may be interpreted as demonstrating an “equivalence” of the object and image. The reversibility of rays implies that the roles of object and image may be exchanged. The object and image may be considered to be maps of each other, with the mapping function defined by the lens. Corresponding object and image points (or object and image lines or object and image planes) are called conjugate points (or lines or planes).

10.2.1 Examples:

1. Plano-convex lens, curved side forward:

\[R_1 = |R_1| > 0\]

\[R_2 = \pm \infty \text{ (sign has no effect)}\]

\[
\frac{1}{s_1} + \frac{1}{s'_2} = (n_2 - 1) \left( \frac{1}{|R_1|} - \frac{1}{\infty} \right) = \frac{n_2 - 1}{|R_1|} > 0
\]

If \(s_1 = +\infty\), then \(s'_2 = f > 0\), the focal length

\[
\frac{1}{f} = \frac{n_2 - 1}{R_1} = \varphi \text{ the power (measured in m}^{-1} = \text{diopters)}
\]

\[f = \frac{R_1}{n_2 - 1} \approx 2R_1 \text{ (since } n_2 \approx 1.5 \text{ for glass)}
\]

We often use the “power” \(\varphi = f^{-1}\) to describe the lens, since powers add (simpler than adding reciprocals of focal lengths).

2. Plano-convex lens, plane side forward:

\[R_1 = \pm \infty\]

\[R_2 = -|R_2| < 0\]

\[
\frac{1}{s_1} + \frac{1}{s'_2} = -\frac{(n_2 - 1)}{R_2} = +\frac{(n_2 - 1)}{|R_2|} > 0
\]

\[f = \frac{|R_2|}{n_2 - 1} \approx 2 |R_2|
\]
3. Plano-concave, plane side forward:

\[ R_1 = \pm \infty \]
\[ R_2 = + |R_2| > 0 \]
\[
\frac{1}{s_1} + \frac{1}{s_2'} = (n_2 - 1) \left( \frac{1}{\infty} - \frac{1}{|R_2|} \right) = - \frac{(n_2 - 1)}{|R_2|} < 0
\]
\[
f = - \frac{|R_2|}{n_2 - 1} \approx -2 \ |R_2|
\]

4. Double convex lens with equal radii:

\[ R_1 = |R| > 0 \]
\[ R_2 = -R_1 = - |R| \]
\[
\frac{1}{s_1} + \frac{1}{s_2} = (n_2 - 1) \left( \frac{1}{|R|} - \left( -\frac{1}{|R|} \right) \right) = 2 \frac{(n_2 - 1)}{|R|} > 0
\]
\[
\frac{1}{f} = \varphi = \frac{2 \cdot (n_2 - 1)}{|R|}
\]
\[
f = \frac{|R|}{2 \cdot (n_2 - 1)} \approx |R| > 0 \quad \text{(since } n_2 \approx 1.5)\]

\[
\boxed{f \approx |R| \text{ for an equiconvex glass lens}}
\]

### 10.3 Magnifications

The most common use for a lens is to change the apparent size of an object (or image) via the magnifying properties of the lens. The mapping of object space to image space “distorts” the size and shape of the image, i.e., some regions of the image are larger and some are smaller than the original object. We can define two types of magnification: transverse and longitudinal.

#### 10.3.1 Transverse Magnification:

The transverse magnification \( M_T \) is what we usually think of as magnification – it is the ratio of object to image dimension measured transverse to the optical axis:
From the similar triangles $\triangle a_1b_1c$ and $\triangle a_2b_2c$, we see that:

$$\frac{y}{s_1} = \frac{|y_2|}{s_2} = -\frac{y_2}{s_2} \quad \text{because } y_2 < 0$$

$$\Rightarrow \frac{y_2}{y_1} = -\frac{s_2}{s_1} \equiv M_T$$

if $|M_T| > 1$, the image is magnified

if $|M_T| < 1$, the image is mini

$M_T < 0 \implies$ the image is inverted

$M_T > 0 \implies$ the image is upright (or erect)

Consider the case where the object is located at the lens, so that $s_1 = 0$. From the imaging equation, we can find $s_2$:

$$s_2 = \left(\frac{1}{f} - \frac{1}{0}\right)^{-1} \implies s_2 = 0$$

In words, the image distance also is 0. The transverse magnification is not well defined from the equation, but the distances show that object, lens, and image all coincide, which leads to the observation that $M_T = +1$.

### 10.3.2 Longitudinal Magnification:

The longitudinal magnification $M_L$ is the ratio of the length of the image along the optical axis to the corresponding length of the object. Since the inverse distances are related in the paraxial imaging equation ($s_1^{-1}$ and $s_2^{-1}$), the longitudinal magnification varies for different object distances. The longitudinal magnification is the ratio of differential (infinitesimal) elements of length of the image and object:

$$M_L = \frac{\Delta s_2}{\Delta s_1} \rightarrow \frac{ds_2}{ds_1}$$

If evaluated at a single on-axis point (so that $\Delta s_1 \rightarrow 0$), then the infinitesimal quantities are related and the longitudinal magnification is the derivative of the image size relative to the object size:

$$M_L|_{s_1} = \frac{ds_2}{ds_1}$$
The expression may be derived by evaluating the total derivative of the object and image distances:

\[
\frac{1}{s_1} + \frac{1}{s_2} = (n - 1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right)
\]

\[
d\left( \frac{1}{s_1} + \frac{1}{s_2} \right) = d\left( \frac{1}{s_1} \right) + d\left( \frac{1}{s_2} \right) = 0 \text{ (because the terms on the RHS are constants)}
\]

\[
\Rightarrow \left( -\frac{1}{s_1^2} \right) ds_1 + \left( -\frac{1}{s_2^2} \right) ds_2 = 0
\]

\[
\Rightarrow M_L = \frac{ds_2}{ds_1} = -\left( \frac{s_2}{s_1} \right)^2 = - (M_T)^2 < 0
\]

The longitudinal magnification of a positive lens is negative because the image moves away from the lens (increasing \( s_2 \)) as the object moves towards the lens (decreasing \( s_1 \)). The longitudinal magnification also affects the irradiance of the image (i.e., the “flux density” of the rays at the image). If \(|M_L|\) is large, then the light in the vicinity of an on-axis location is “spread out” over a larger region of space at the image, so the irradiance of the image is decreased.
The concept of longitudinal magnification applied to a luminous rod. The section located at \( s_1 = 2f \) is imaged with unit transverse magnification at \( s_2 = 2f \). Sections of the rod with \( s_1 > 2f \) are imaged "closer" to the lens (\( s_2 < 2f \)), and the the energy density is remapped to account for the nonlinear distance relationship 
\[
\frac{1}{s_1} + \frac{1}{s_2} = \frac{1}{f}.
\]

### 10.4 Spherical Mirrors

Consider a mirror with a spherical reflective surface. We can define a power of the surface that is analogous to the power (and thus related to the focal length) of a thin lens. A ray that passes through the center of curvature of the spherical mirror must intersect the surface at right angles, so that the incident angle measured relative to the normal is \( \theta = 0 \). The reflected angle also is \( \theta = 0 \) and thus the reflected ray also passes through the center of curvature. A ray that crosses the optical axis at \( O \) such that \( OC \gg 0 \) makes an angle of \( -\theta \approx 0 \) with the surface normal. It therefore is reflected at an angle \( +\theta \) measured relative to the normal and intersects the axis at \( O' \) such that \( OC' \approx -CO \gg 0 \). A paraxial ray from an object at \( \infty \) makes a small angle \( -\theta \approx 0 \) and reflects at \( +\theta \approx 0 \). The reflected paraxial ray intersects the optical axis at a distance \( FF' \approx CV \cdot \frac{1}{2} = \frac{R}{2} \). From the triangle

Ray redirection by a spherical mirror: (a) an incoming ray through the center of curvature \( C \) produces an exiting ray that also passes through \( C \) because the ray angles are \( 0 \); (b) a ray that just misses \( C \) intersects the mirror at a ray angle of \(-\theta\) and produces a ray at an angle of \( +\theta \) that misses \( C \) to the opposite side; (c) a paraxial ray from an object at infinity intersects the surface at an angle \(-\theta \approx 0\) and
is reflected at $+\theta$ to cross the optical axis at $F'$, which is halfway between $C$ and $V$, which shows that $f = \frac{-R}{2}$ for paraxial rays.

Now consider the case where a nonparaxial ray from an object at $\infty$ reaches the mirror and makes a large angle measured from the surface normal. The ray is reflected at the same large angle $\theta$ and intersects the optical axis at $O'$ as shown.

The distance $\overline{CF'}$ may be evaluated from $\triangle PF'C$ using the law of sines:

$$\frac{R}{\sin(\pi - 2\theta)} = \frac{\overline{CF'}}{\sin[\theta]} \implies \overline{CF'} = R \frac{\sin[\theta]}{\sin(\pi - 2\theta)} = R \frac{\sin[\theta]}{\sin[2\theta]}$$

$$\implies \overline{VF'} = f = \overline{CF'} - R < 0$$

$$f = R \left( \frac{\sin[\theta]}{\sin[2\theta]} - 1 \right)$$

We can evaluate this in the limit $\sin[\theta] \to \theta, \sin[2\theta] \to 2\theta$ to get the focal length of the mirror in the paraxial case:

$$\text{paraxial case:} f \to R \left( \frac{\theta}{2\theta} - 1 \right) = -\frac{R}{2}$$

The paraxial focal length of a spherical mirror is $f = \frac{-2}{R}$, where the negative sign corrects for the observation that a spherical with $R < 0$ makes light converge and
thus has positive power. The focal length of a spherical mirror can be put in the same form as a refracting surface by setting the second index of refraction \( n_2 = -1 \)

\[
\frac{1}{f} = -\frac{2}{R} = \frac{(n_2 - n_1)}{R} \implies n_1 = +1 \text{ and } n_2 = -1
\]

\( R < 0 \implies \text{concave mirror} \implies f > 0 \)
\( R > 0 \implies \text{convex mirror} \implies f < 0 \)

Note that the refractive index of the medium has no effect on the power of a spherical lens because Snell’s law for reflection does not include any contribution by \( n \). In other words, the reflected ray angle is not affected by the index of refraction of the medium “ahead of” the mirror.

Consider an example of a large angle, e.g., \( \theta = \frac{\pi}{6} \). The focal length for this ray is:

\[
f \left[ \theta = \frac{\pi}{6} \right] = R \left( \frac{\sin \left[ \frac{\pi}{6} \right]}{\sin \left[ \frac{\pi}{3} \right]} - 1 \right) \approx -0.423R
\]

In words, the focal point of a spherical lens for paraxial rays is different from the focal point for nonparaxial rays, which means that rays from the same object that are collected at different ray angles do not converge to a sharp focus, thus degrading the quality of the image. This effect is spherical aberration. Probably the most famous example of an imaging mirror with spherical aberration is the Hubble Space Telescope before adding the COSTAR optical corrector system.

Rays from an object at infinity at different ray angles (or equivalently, at different ray heights at the mirror) cross the optical axis at different locations. The paraxial rays intersect the optical axis at \( F' \) such that \( f = -\frac{R}{2} \), but the focal point moves toward \( V \) as the ray height increases.

A graph if the focal length for different ray angles is shown:
10.5 SYSTEMS OF THIN LENSES

The variation in focal length $V^0 = f = CF^0 - R$ with incident ray angle $\theta$ over the interval $0 \leq \theta \leq \frac{\pi}{4}$, showing the decrease in the focal length with increasing angle. This means that the focus for rays making a larger angle is positioned closer to the vertex of the mirror.

Spherical mirrors can create a good image of an object at the center of curvature $C$, though this is not very useful because the image also is located at $C$. As the object distance increases, the spherical aberration of the mirror becomes more apparent, so that the aperture diameter of a spherical mirror typically must be small to obtain good image quality. If the object is located at $\infty$, a spherical mirror cannot give good quality unless the aperture size is quite small compared to the radius of curvature. Mathematically, it is easy to show that the appropriate mirror shape for imaging at infinity is a paraboloid: parallel rays from an object at $\infty$ reflect from a paraboloidal surface and converge to a single image point.

10.5 Systems of Thin Lenses

The images produced by systems of thin lenses may be located by finding the “intermediate” images produced by the individual lenses, which then become the objects for the next lens in the sequence. This type of analysis also is directly applicable to the “thick” lens where the surfaces take the places of the individual thin lenses. The object is labelled by $O$ and the corresponding image by $O'$, the object- and image-space focal points are $F$ and $F'$, and the object- and image-space vertices (first and last surfaces of the system) by $V$ and $V'$. 
Imaging by a system of two thin lenses $L_1$ and $L_2$ separated by the distance $d$. The object and image distances for the first lens are $s_1$ and $s_1'$ and for the second lens are $s_2$ and $s_2'$.

In this particular case, the lenses are separated by the distance $t$ and the object distance for the second lens $s_2 = t - s_1'$. The imaging equation for the first lens determines $s_1'$:

$$\frac{1}{s_1} + \frac{1}{s_1'} = \frac{1}{f_1} \quad \Rightarrow \quad \frac{1}{s_1'} = \frac{1}{f_1} - \frac{1}{s_1} = \frac{s_1 - f_1}{s_1 f_1}$$

$$\Rightarrow s_1' = \frac{s_1 f_1}{s_1 - f_1}$$

So the object distance to the second lens is $s_2$:

$$s_2 = t - s_1' = t - \frac{s_1 f_1}{s_1 - f_1} = \frac{s_1 t - f_1 t - s_1 f_1}{s_1 - f_1} = \frac{s_1 (t - f_1) - f_1 t}{s_1 - f_1}$$
Now apply the imaging equation to the second lens and substitute for $s_2$:

$$\frac{1}{s_2} + \frac{1}{s_2'} = \frac{1}{f_2} \implies \frac{1}{s_2'} = \frac{1}{f_2} - \frac{1}{s_2}$$

$$= \frac{1}{f_2 - s_1 (t - f_1) - f_1 t}$$

$$= \frac{s_1 (t - f_1) - f_1 t - f_2 (s_1 - f_1)}{f_2 (s_1 (t - f_1) - f_1 t)}$$

$$= \frac{t (s_1 - f_1) - s_1 (f_1 - f_2) + f_1 f_2}{f_2 ([t (s_1 - f_1)] - s_1 f_1)}$$

$$= \frac{s_1 (t - f_1)}{f_2 ([t (s_1 - f_1)] - s_1 f_1)}$$

$$s_2' = \frac{f_2 t - \frac{s_1 f_1 f_2}{s_1 - f_1}}{(t - f_2) - \frac{s_1 f_1}{s_1 - f_1}}$$

This complicated expression determines the image distance from the second lens given the focal lengths, the “thickness” (distance between the lenses), and the object distance $s_1$.

### 10.5.1 Back Focal Distance

We would like to collect the results into a single simple equation that is analogous to the imaging equation for the single thin lens. The back focal distance $BFD$ is $s_2'$ for $s_1 = +\infty$. In other words, it is the distance from the image-space vertex to the image-space focal point: $BFD = \nabla F
\n$. Note that it is NOT the focal length of the system. We’ll further analyze the difference between $f_{\text{eff}}$ and $BFD$ shortly.

$$BFD \equiv \nabla F = \lim_{s_1 \to \infty} s_2' = \lim_{s_1 \to \infty} \frac{f_2 t - \frac{s_1 f_1 f_2}{s_1 - f_1}}{(t - f_2) - \frac{s_1 f_1}{s_1 - f_1}}$$

$$= \frac{f_2 t - \left(\frac{s_1}{s_1 - f_1}\right) f_1 f_2}{(t - f_2) - \left(\frac{s_1}{s_1 - f_1}\right) f_1}$$

$$= \frac{f_2 t - f_1 f_2}{(t - f_2) - f_1}$$

$$\lim_{s_1 \to \infty} s_2' = \frac{f_2 (t - f_1)}{t - (f_1 + f_2)} \equiv BFD$$

Note that if $t = f_1 + f_2$ then the $BFD$ is $+\infty$, so the object and image are both an infinite distance from the system. Such a system has an infinite focal length, and thus its power (reciprocal of the focal length) is zero.
10.5.2 Front Focal Distance

Similarly, the front focal distance (FFD) is \( s_1 \) if \( s_2' = \infty \) and is identical to \( FV \). It is calculated by setting the denominator of the expression for \( s_2' \) to zero:

\[
(t - f_2) - \frac{s_1 f_1}{s_1 - f_1} = 0
\]

\[\Rightarrow \frac{s_1 f_1}{s_1 - f_1} = t - f_2\]

\[\Rightarrow \frac{s_1}{s_1 - f_1} = \frac{t - f_2}{f_1}\]

\[\Rightarrow s_1 f_1 = (t - f_2) (s_1 - f_1)\]

\[\Rightarrow s_1 f_1 = ts_1 - t f_1 - s_1 f_2 + f_1 f_2\]

\[\Rightarrow s_1 (f_1 + f_2 - t) = f_1 f_2 - t f_1\]

\[
\lim_{s_2' \to \infty} s_1 = FV = \frac{f_1 (t - f_2)}{t - (f_1 + f_2)} = FFD
\]

Note that this expression has the same form as the front focal distance except that \( f_1 \) and \( f_2 \) are “swapped;” this makes intuitive sense, because the only difference between the two cases is that the two lenses are exchanged.

10.5.3 Thin Lenses in Contact

If the two thin lenses are in contact, then \( t = 0 \) and the focal distances are equal to the focal length of the “equivalent single lens.” We can calculate its value by setting \( t = 0 \) in the equations for \( FFD \) and \( BFD \):

\[FFD|_{t=0} = \frac{f_1 (0 - f_2)}{0 - (f_1 + f_2)} = \frac{f_1 f_2}{f_1 + f_2}\]

\[BFD|_{t=0} = \frac{f_2 (0 - f_1)}{0 - (f_1 + f_2)} = \frac{f_1 f_2}{f_1 + f_2}\]

\[f|_{t=0} = \frac{f_1 f_2}{f_1 + f_2}\]

\[\Rightarrow \frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}, \text{ if } t = 0\]
Two thin positive lenses in contact. The focal length of the system is shorter than
the focal lengths of either, and may be evaluated to see that \( f_{\text{eff}} = \frac{f_1 f_2}{f_1 + f_2} \). The
image-space principal point is the location of the “equivalent thin lens”.

10.5.4 “Effective Focal Length” of a System with Two Lenses

In the general system created from two positive lenses, we need to evaluate the focal
length \( f \) (or, equivalently, the power \( \varphi \)) as the “thickness” \( t \) between the lenses is
increased from zero. We just showed that:

\[
\lim_{t \to 0} \left( \frac{1}{f} \right) = \frac{1}{f_1} + \frac{1}{f_2} > 0
\]

where we are assuming that the two focal lengths are positive. If \( t \) is increased from
0 until the lenses are separated by the sum of the focal lengths, then an incoming ray
parallel to the axis exits parallel to the axis; we have formed a “telescope.” Since the
angles relative to the axis of the incoming and outgoing rays have not changed, then
this system has no power (\( \varphi = 0 \)); its focal length is infinite:

\[
\lim_{t \to f_1 + f_2} \left( \frac{1}{f} \right) = 0 \implies f_{\text{eff}} = \infty
\]

Thus the focal length of the system increased (the system power decreased) as \( t \)
increased from zero. This leads us to suspect that the power of the two-lens system
must have a form like:

\[
\varphi = \frac{1}{f_1} + \frac{1}{f_2} - \alpha t
\]

where \( \alpha \) is some constant with units of (length\(^{-2}\)). Since the only parameters used
in the system are \( f_1 \) and \( f_2 \), which have dimensions of length, we might hypothesize
that the power of the system is:

\[
\varphi = \frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1f_2} = \varphi_1 + \varphi_2 - \varphi_1\varphi_2t
\]

\[
\varphi = \frac{f_1f_2}{(f_1 + f_2) - t}
\]

The effective focal length of the system is the reciprocal of its power:

\[
f_{\text{eff}} = \varphi^{-1} = \frac{f_1f_2}{(f_1 + f_2) - t}
\]

It can be interpreted as the focal length of the single thin lens that generates the same outgoing ray.

The power \(\varphi\) is measured in diopters \([m^{-1}]\) if all distances are measured in m. This expression satisfies the two limiting cases we proposed. If the two lenses have positive power and the separation is just less than the sum of focal lengths, the effective focal length can be very large. This is also the case if one of the two lenses has negative power (so that the numerator is negative) and the separation is just larger than the sum of the focal lengths (so that the denominator is just smaller than zero).

10.5.5 Positive Lenses Separated by \(t < f_1 + f_2\)

If two positive thin lenses are separated by less than the sum of the focal lengths, the image-space focal point \(F'\) is closer to the first lens than it would have been had the second lens been absent. As shown, the effective focal length of the system is \(f_{\text{eff}} < f_1\). We can apply the equation for \(f_{\text{eff}}\) to this case to see that:

\[
f_{\text{eff}} = \frac{f_1f_2}{(f_1 + f_2) - t} > 0
\]

\[
0 < f_{\text{eff}} < f_1 + f_2
\]

As just stated, the effective focal length \(f_{\text{eff}}\) of the system determines the location of the single thin lens that is “equivalent” to the system in image space. A corresponding point exists in object space that will be located next. The equivalent thin lens has the same focal length \(f_{\text{eff}}\) as the two-lens system but is located at the point labeled by \(H'\) in the drawing. \(H'\) is the image-space principal point, and will be discussed in more detail shortly.
A pair of positive thin lenses separated by less than the sum of the focal lengths.

Example:

Consider a specific example with \( f_1 = 100 \text{ mm}, \ f_2 = 50 \text{ mm}, \) and \( t = 75 \text{ mm}. \) The focal length of the equivalent single lens is:

\[
 f_{\text{eff}} = \frac{f_1 f_2}{(f_1 + f_2) - t} = \frac{(100 \text{ mm}) (50 \text{ mm})}{(100 \text{ mm} + 50 \text{ mm}) - 75 \text{ mm}} = \frac{200}{3} \text{ mm} = 66 \frac{2}{3} \text{ mm}
\]

The image from the first lens is formed at its focal point:

\[
 s_1' = \left( \frac{1}{f_1} - \frac{1}{s_1} \right)^{-1} = \left( \frac{1}{100 \text{ mm}} - \frac{1}{\infty} \right)^{-1} = 100 \text{ mm}
\]

The object distance to the second lens is therefore the difference \( t - s_1' \):

\[
 s_2 = t - s_1' = (75 - 100) \text{ mm} = -25 \text{ mm}
\]

The image of an object located at \( s_1 = \infty \) appears at \( s_2' \):

\[
 s_2' = \left( \frac{1}{f_2} - \frac{1}{s_2} \right)^{-1} = \left( \frac{1}{50 \text{ mm}} - \frac{1}{-25 \text{ mm}} \right)^{-1} = \frac{50}{3} \text{ mm} = 16 \frac{2}{3} \text{ mm}
\]

measured from the rear vertex \( V' \) of the system. We already know that the system focal length is \( 66 \frac{2}{3} \text{ mm} \), so the image-space principal point \( H' \) (the position of the equivalent thin lens) is located \( 66 \frac{2}{3} \text{ mm} \) IN FRONT of the system focal point, i.e., 50 mm in front of the second lens and 25 mm behind the first lens. We can locate this point by continuing the object ray “forward” through the system and the image ray “backward” until they intersect; this is the location of the equivalent single thin lens that creates the same object point (this location is \( H' \)). The effective focal length is the distance from \( H' \) to \( F' \), the image of an object at infinity.
Object-Space Principal Point

We have already shown how to find the location of the equivalent single lens on the “output side” by extending the rays entering and exiting the system until they meet. We can locate the equivalent single lens in “object space” by “reversing” the system, as shown in the figure. The “first” lens in the system is now $L_2$ with $f_2 = 50\, \text{mm}$. The “second” lens is $L_1$ with $f_1 = 100\, \text{mm}$ and the separation is $t = 75\, \text{mm}$. The resulting effective focal length remains unchanged at $f_{\text{eff}} = \frac{200}{3}\, \text{mm} = 66\frac{2}{3}\, \text{mm}$. If we bring in a ray from an object at $\infty$, the “intermediate” image formed by $L_2$ is located at the focal point of $L_2$:

$$s'_1 = \left(\frac{1}{f_2} - \frac{1}{s_1}\right)^{-1} = \left(\frac{1}{50} - \frac{1}{\infty}\right)^{-1} = 50\, \text{mm}$$

Thus the image distance to $L_1$ is:

$$s_2 = t - s'_1 = 75 - 50 = +25\, \text{mm}$$

The image of the object at $s_1 = \infty$ produced by the entire system is located at $s'_2$:

$$s'_2 = \left(\frac{1}{f_1} - \frac{1}{s_2}\right)^{-1} = \left(\frac{1}{100} - \frac{1}{+25}\right)^{-1} = -\frac{100}{3} = -33\frac{1}{3}\, \text{mm}$$

measured from the “second” lens $L_1$ (or equivalently from the second vertex). The image is “in front” of the second lens (on the object-space side) thus is virtual. The object-space principal point $H$ is the point such that the distance $HV = f = 66\frac{2}{3}\, \text{mm}$, so $H$ is located $-33\frac{1}{3}\, \text{mm}$ IN FRONT of $L_2$.

![Diagram of the two-lens imaging system showing the object-space principal point](attachment:image.png)

The principal and focal points of the two-lens imaging system in both object and image spaces.

When we “re-reverse” the system to graph the object- and image-space principal points, $H$ is located “behind” the lens $L_2$, as shown in the graphical rendering of the entire system: The object-space principal point is the location of the equivalent thin lens if the imaging system is reversed.
Two-lens system showing the object- and image-space principal points $H$ and $H'$ and focal points $F$ and $F'$.

For a system of two thin lenses in contact, the principal points coincide with the common location of the two lenses, i.e., that $V' = H' = H = V$.

We can now use these locations of the equivalent thin lens in the two spaces to locate the images by applying the thin-lens (Gaussian) imaging equation. HOWEVER, it is VERY important to realize that the distances $s$ and $s'$ are respectively measured from the object $O$ to the object-space principal point $H$ and from the image-space principal point $H'$ to the image point $O'$.

\[
s = \overline{OH} \\
\]
\[
s' = \overline{H'O'}
\]

The process is demonstrated after first locating the images via a direct calculation.

**Imaging Equation for Equivalent Single Lens ("Brute Force" Calculation)**

Now consider the location and magnification of the image created by the original two-lens imaging system (with $L_1$ in front) for an object located 1000 mm in front of the system (so that $OV = 1000$ mm). We can locate the image step by step:

**Intermediate image created by $L_1$:**

\[
s'_1 = \left( \frac{1}{f_1} - \frac{1}{s_1} \right)^{-1} = \left( \frac{1}{100} - \frac{1}{1000} \right)^{-1} = \frac{1000}{9} \text{ mm} \approx 111.11 \text{ mm}
\]
Transverse magnification of first image:

\[(M_T)_1 = -\frac{s'_1}{s_1} = -\frac{\left(\frac{1000}{9}\text{ mm}\right)}{1000\text{ mm}} = -\frac{1}{9}\]

Distance from first image to \(L_2\):

\[s_2 = t - s'_1 = 75\text{ mm} - \frac{1000}{9}\text{ mm} = -\frac{325}{9}\text{ mm} \cong -36.11\text{ mm}\]

Distance from \(L_2\) to final image:

\[s'_2 = \left(\frac{1}{f_2} - \frac{1}{s_2}\right)^{-1} = \left(\frac{1}{50} - \frac{1}{-\frac{325}{9}}\right)^{-1} = +\frac{650}{31} \approx +20.97\text{ mm}\]

Transverse magnification of second image:

\[(M_T)_2 = -\frac{\left(\frac{650}{31}\right)}{-\frac{325}{9}} = +\frac{18}{31}\]

The transverse magnification of the final image is the product of the transverse magnifications of the images created by the two lenses:

\[M_T = (M_T)_1 \cdot (M_T)_2 = \left(-\frac{1}{9}\right) \left(+\frac{18}{31}\right) = -\frac{2}{31} \approx -0.065\]

The transverse magnification indicates that the image is minimized (or demagnified) and inverted.

**Brute-Force Calculation for Object at \(H\)**

Now repeat the calculation for an object is located at \(H\). In this case, the object is located “inside” the system, so the distance from the object to the first lens is:

\[s = \overline{OH} = -\overline{VH} = -100\text{ mm}\]

The object in this case is virtual. The “intermediate” image created by the first lens \(L_1\) is located at:

\[s'_1 = \left(\frac{1}{f_1} - \frac{1}{s_1}\right)^{-1} = \left(\frac{1}{100} - \frac{1}{-100}\right)^{-1} = +50\text{ mm}\]

Note that the transverse magnification of the image created by the first lens is:

\[(M_T)_1 = -\frac{s'_1}{s_1} = -\frac{+50\text{ mm}}{-100\text{ mm}} = +\frac{1}{2}\]
The object distance to the second lens $L_2$ is:

$$s_2 = t - s'_1 = 75 \text{ mm} - 50 \text{ mm} = +25 \text{ mm}$$

so the object for the second lens is real. The distance to the final image is:

$$s'_2 = V'O' = \left( \frac{1}{f_2} - \frac{1}{s_2} \right)^{-1} = \left( \frac{1}{50} - \frac{1}{+25} \right)^{-1} = -50 \text{ mm}$$

The image is located 50 mm “in front” of $L_2$ (again, “inside” the system) and thus the image of the object at $H$ is virtual. The diagram shows that the image coincides with the image-space principal point $H'$, which is conveniently appropriate for our labeling convention. The transverse magnification of the image created by the second lens is:

$$(M_T)_2 = \frac{-50 \text{ mm}}{+25 \text{ mm}} = +2$$

The transverse magnification of the final image relative to the original image is the product of the two individual magnifications:

$$M_T = (M_T)_1 \cdot (M_T)_2 = \left( \frac{1}{2} \right) (+2) = +1$$

In words,

An object located at $H$ creates an image at $H'$ with transverse magnification $+1$.

Ray trace for object located at object-space principal point $H$ showing that the image is located at the image-space principal point $H'$. 
Yet Another Example...

What if the object located so that $OV = \frac{100}{3}$ mm. We can locate the image step by step:

Intermediate image created by $L_1$:

$$s'_1 = \left(\frac{1}{f_1} - \frac{1}{s_1}\right)^{-1} = \left(\frac{1}{100 \text{ mm}} - \frac{1}{\frac{100}{3} \text{ mm}}\right)^{-1} = -50 \text{ mm} \text{ (virtual)}$$

Transverse magnification of first image::

$$(M_T)_1 = -\frac{s'_1}{s_1} = -\frac{-50 \text{ mm}}{+\frac{100}{3} \text{ mm}} = +\frac{3}{2}$$

Distance from first image to $L_2$:

$$s_2 = t - s'_1 = 75 \text{ mm} - (-50 \text{ mm}) = +125 \text{ mm}$$

Distance from $L_2$ to final image:

$$s'_2 = \left(\frac{1}{f_2} - \frac{1}{s_2}\right)^{-1} = \left(\frac{1}{50 \text{ mm}} - \frac{1}{125 \text{ mm}}\right)^{-1} = +\frac{250}{3} \text{ mm}$$

Transverse magnification of second image:

$$(M_T)_2 = -\frac{(+\frac{250}{3} \text{ mm})}{(+125 \text{ mm})} = -\frac{2}{3}$$

The transverse magnification of the final image is the product of the transverse magnifications of the images created by the two lenses:

$$M_T = (M_T)_1 \cdot (M_T)_2 = \left(\frac{3}{2}\right) \left(-\frac{2}{3}\right) = -1$$

Since $M_T = -1$, we know that the object is located $2f$ away and so is the image. This confirms the locations of the principal points.

Imaging Equation for Object- and Image-Space Principal Points

We have just seen that the object- and image-space principal points are the points related by unit magnification. They also are the “reference” locations from which the system focal length is measured;

$$f_{eff} = FH = H F^T$$
(assuming that the object space and image space are the same medium, e.g., air). In exactly the same way, these are the “reference” locations from which the object and image distances are measured for a multi-element system

\[ s = \overline{OH} \]
\[ s' = \overline{H'O'} \]

The ray entering the system can be modeled as traveling from the object \( O \) to the object-space principal point \( H \). The resulting outgoing (image) ray travels from the image-space principal point \( H' \) to the image point \( O' \). This may seem a little “weird”, but actually makes perfect sense if we relate the measurements to the equation for a single thin lens. In that situation, focal lengths are measured from the thin lens to its focal points. In other words, the object- and image-space vertices \( V \) and \( V' \) coincide with the principal points \( H \) and \( H' \). We know that an object located at the lens \((s = 0)\) generates an image at the lens \((s' = 0)\) with magnification of +1; the heights of the object and image at the principal points are identical. In the realistic system where the object- and image-space principal points are at different locations, the image of an object located at \( H \) has an image at \( H' \) and still with unit magnification; an object located at the object-space principal point creates an image at the image-space principal point with the transverse magnification \( M_T = +1 \).

**EMPHASIS:** the principal points \( H \) and \( H' \) are the locations of the object and image related by unit transverse magnification: \( M_T = +1 \)

Contrast this to the situation if the object distance from \( H \) is \( 2f \), so that the image distance is also \( 2f_{eff} \) and the transverse magnification is \( -1 \):

\[ \overline{OH} = s = 2f_{eff} \]
\[ \frac{1}{s} + \frac{1}{s'} = \frac{1}{f_{eff}} \]
\[ s' = \overline{H'O'} = 2f_{eff} \]
\[ M_T = -\frac{2f_{eff}}{2f_{eff}} = -1 \]

The principal points are “crossed” in the imaging system considered thus far, which merely means that the object-space principal point is “behind” the image-space principal point (towards the image-space side of the lens system). Any ray cast into the system from the object point \( O \) to \( H \) creates an image ray that departs from \( H' \) at the same height \((M_T = +1)\) and directed towards the image point \( O' \):
Principal points of an imaging system: The dashed ray from the object at \( O \) reaches the object-space principal point \( H \) with height \( h \). The image ray (solid line) departs from the image-space principal point \( H' \) with the same height \( h \) and goes to the image point \( O' \), so that the distances \( OH = s \) and \( H'V' = s' \) satisfy the imaging equation \( \frac{1}{s} + \frac{1}{s'} = \frac{1}{f_{eff}} \).

Location of Image from System via Principal Points

We can also solve this problem by using the equivalent single thin lens where the distances are measured from the object- and image-space principal points. We have already shown that the system (effective) focal length is:

\[
\frac{1}{f_{eff}} = \frac{200}{3} \text{ mm}
\]

The object and image distances \( s \) and \( s' \) of the single lens equivalent to the two-lens system are respectively measured principal points: \( s = OH \) and \( s' = H'V' \).

The object distance is measured to the object-space principal point, which is 100 mm behind \( L_1 \) (or \( V \)), thus the object distance is the distance from \( O \) to \( L_1 \) plus 100 mm:

\[
s = OV + VH = 1000 \text{ mm} + 100 \text{ mm} = 1100 \text{ mm}
\]
The single-lens imaging equation may be used to find the image distance \( s' \), which now is measured from the image-space principal point \( H' \):

\[
\frac{1}{s'} = \left( \frac{1}{f_{\text{eff}}} - \frac{1}{s} \right)^{-1}
\]

\[
= \left( \frac{1}{\frac{200}{3}} - \frac{1}{1100} \right)^{-1}
\]

\[
= \frac{2200}{31} = \overline{HO'}
\]

The image distance from the vertex is calculated by subtracting the distance from the image-space principal point \( H' \) to the image-space vertex \( V' \):

\[
\overline{VO'} = \overline{HO'} - \overline{HV'}
\]

\[
= \frac{2200}{31} \text{ mm} - 50 \text{ mm} = \frac{650}{31} \approx +20.97 \text{ mm}
\]

The resulting magnification is:

\[
M_T = -\frac{s'}{s} = -\left( \frac{\frac{2200}{31} \text{ mm}}{1100 \text{ mm}} \right) = -\frac{2}{31} \approx -0.065
\]

Note that both the image distance and the transverse magnification match those obtained with the step-by-step calculation performed above.

### 10.5.6 Cardinal Points

The object-space and image-space focal and principal points are four of the six so-called cardinal points that completely determine the paraxial properties of an imaging system. There are three pairs of locations where one of each pair is in object space and the other is in image space. The object- and image-space focal points are \( F \) and \( F' \), while the principal points \( H \) and \( H' \) are the locations on the axis in object and image space that are images of each other with unit magnification. The nodal points \( N \) and \( N' \) are the points in object and image space where the ray angle entering the object-space nodal point and exiting the image-space nodal point are identical. For systems where the object and image spaces are in air (most of the systems we care about), the principal and nodal points coincide.

A table of significant points on the axis of a paraxial system is given below:
### 10.5.7 Positive Thin Lenses Separated by \( t = f_1 + f_2 \)

We’ve already looked at this example, but consider it one more time. If the two lenses are separated by the sum of the focal lengths, then an object at \( \infty \) forms an image at \( \infty \); the system focal length is infinite. Since the focal points are both located at infinity, we say that the system is *afocal*; it has zero power, i.e., the rays exit the system at the same angle that they entered it. If the focal length of the first lens is longer than that of the second, the system is a *telescope*.

![Image of two thin lenses separated by the sum of their focal lengths](attachment:image.png)

*Two thin lenses separated by the sum of their focal lengths. An object located an infinite distance from the first lens forms an “intermediate” image at the image-space focal point \( f'_1 \) of the first lens. The second lens forms an image at infinity. Both object- and image-space focal lengths of the equivalent system are infinite: \( f = f' = \infty \). The system has “no” focal points – it is *afocal*."

The focal length of this system is:

\[
\frac{1}{f_{\text{eff}}} = \frac{1}{\infty} = 0 = \left( \frac{1}{f_1} + \frac{1}{f_2} \right) - \frac{t}{f_1 f_2} = \left( \frac{1}{f_1} + \frac{1}{f_2} \right) - \frac{f_1 + f_2}{f_1 f_2} = \left( \frac{1}{f_1} + \frac{1}{f_2} \right) - \left( \frac{1}{f_1} + \frac{1}{f_2} \right) = 0
\]

where \( t = f_1 + f_2 \) is the separation between the two lenses.
10.5.8 Positive Thin Lenses Separated by $t = f_1$ or $t = f_2$

We now continue the sequence of examples for two positive lenses separated by different distances. If two positive lenses are separated by the focal length of the first lens, then the focal length of the system is:

$$f_{\text{eff}} \text{ (for } t = f_1) = \frac{f_1 f_2}{f_1 + f_2} - f_1 = f_1$$

If separated by the focal length of the second lens, the system focal length is $f_2$.

$$f_{\text{eff}} \text{ (for } t = f_2) = \frac{f_1 f_2}{f_1 + f_2} - f_2 = f_2$$

For the purpose of this example, we analyze the second case because it is the basis for probably the most common application of imaging optics. The extension to the first case is trivial. Since the focal length of the system is identical to the focal length of the second lens if the lenses are separated by $t = f_2$, this suggests the question of the effect of the first lens on the image.

Consider a specific case with $f_2 = 100 \text{ mm}$ and $f_1 = 200 \text{ mm}$. If only $L_2$ is present and the object distance is $s_2 = 1100 \text{ mm}$, the image distance is:

$$s_2' = \left( \frac{1}{f_2} - \frac{1}{s_2} \right)^{-1} = \left( \frac{1}{100} - \frac{1}{1100} \right)^{-1} = 110 \text{ mm}$$

The associated transverse magnification is:

$$M_T = -\frac{s_2'}{s_2} = -\frac{110 \text{ mm}}{1100 \text{ mm}} = -\frac{1}{10}$$
Now add \( L_1 \) at the front focal point of \( L_2 \) and find the associated image. The object distance to \( L_1 \) is \( 1100 \text{ mm} - 100 \text{ mm} = 1000 \text{ mm} \). The lens forms an image at distance:

\[
s'_1 = \left( \frac{1}{f_1} - \frac{1}{s_1} \right)^{-1} = \left( \frac{1}{200 \text{ mm}} - \frac{1}{1000 \text{ mm}} \right)^{-1} = +250 \text{ mm}
\]

The associated transverse magnification is:

\[
(M_T)_1 = -\frac{s'_1}{s_1} = -\frac{250 \text{ mm}}{1000 \text{ mm}} = -\frac{1}{4}
\]

The object distance to the second lens is:

\[
s_2 = t - s'_1 = 100 \text{ mm} - 250 \text{ mm} = -150 \text{ mm}
\]

and the resulting image distance is:

\[
s'_2 = \left( \frac{1}{f_2} - \frac{1}{s_2} \right)^{-1} = \left( \frac{1}{100} - \frac{1}{-150} \right)^{-1} = +60 \text{ mm}
\]

The associated transverse magnification of the image formed by the second lens is:

\[
(M_T)_2 = -\frac{60}{-150} = +\frac{2}{5}
\]

The magnification of the system is the product of the magnifications:

\[
M_T = (M_T)_1 (M_T)_2 = \left( -\frac{1}{4} \right) \left( +\frac{2}{5} \right) = -\frac{1}{10}
\]

This is the same transverse magnification that we obtained from \( L_2 \) alone! The magnification is not changed by the addition of lens \( L_1 \)! However, the position of the image HAS changed (from +110 mm to +60 mm); the image is closer to \( L_2 \) if \( L_1 \) is added.

This system demonstrates the principle of eyeglass lenses, where the corrective lens is placed at the object-space focal point of the eyepiece. The corrective action is to move the image without changing its transverse magnification.

### 10.5.9 Positive Thin Lenses Separated by \( t > f_1 + f_2 \)

If the two positive lenses are separated by more than the sum of the focal lengths, the focal length of the resulting system is **negative**:

\[
f_{\text{eff}} = \frac{f_1 f_2}{(f_1 + f_2) - t} < 0
\]

If the object distance is \( \infty \), the first lens forms an “intermediate” image at its imagespace focal point, i.e., at \( s'_1 = f'_1 \). Since the object distance \( s_2 \) measured from the
second lens is larger than \( f_2 \), a “real” image is formed by the second lens at the system focal point \( F' \). If we extend the exiting ray until it intersects the incoming ray from the object at infinity, we can locate the equivalent single thin lens for the system, i.e., the image-space principal point \( H' \). In this case, this is located farther from the second lens than the focal point. The effective focal length \( f_{eff} = H'H/ < 0 \), so the system has negative power; this system created from two positive lenses is equivalent to a single thin lens with negative power.

The system composed of two thin lenses separated by \( d > f_1 + f_2 \). The image-space focal point \( F' \) of the system is beyond the second lens, but the image-space principal point \( H' \) is located even farther from \( L_2 \). The distance \( H'H/ = f_{eff} < 0 \), so the system has negative power!

### 10.5.10 Systems of Two Positive Thin Lenses with Different Focal Lengths

\[
f_{eff} = H'H/ = FH = \frac{f_1 f_2}{(f_1 + f_2) - t} = \frac{100 \text{ mm} \cdot 25 \text{ mm}}{100 \text{ mm} + 25 \text{ mm} - t}
\]

\[
BFD = V'F = \frac{f_2 (f_1 - t)}{(f_1 + f_2) - t} = \frac{25 \text{ mm} \cdot (100 \text{ mm} - t)}{(100 \text{ mm} + 25 \text{ mm}) - t}
\]

\[
H'V' = H'H/ - VF' = \frac{f_1 f_2}{(f_1 + f_2) - t} - \frac{f_2 (f_1 - t)}{(f_1 + f_2) - t} = \frac{f_2 t}{(f_1 + f_2) - t}
\]

\[
FFD = VF = \frac{f_1 (f_2 - t)}{(f_1 + f_2) - t} = \frac{100 \text{ mm} \cdot (25 \text{ mm} - t)}{(100 \text{ mm} + 25 \text{ mm}) - t}
\]

\[
V'H = FH - VF = \frac{f_1 t}{(f_1 + f_2) - t}
\]

\[
f_1 = +100 \text{ mm} \\

f_2 = +25 \text{ mm}
\]
<table>
<thead>
<tr>
<th></th>
<th>$FFD$</th>
<th>$BFD$</th>
<th>$f_{eff}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 mm</td>
<td>+20 mm</td>
<td>+20 mm</td>
<td>+20 mm</td>
</tr>
<tr>
<td>+25 mm</td>
<td>0 mm</td>
<td>+18.75 mm</td>
<td>+25 mm = $f_2$</td>
</tr>
<tr>
<td>+50 mm</td>
<td>$-33\frac{1}{3}$ mm</td>
<td>+16.2 mm</td>
<td>+33$\frac{1}{3}$ mm</td>
</tr>
<tr>
<td>+75 mm</td>
<td>$-100$ mm</td>
<td>+12.5 mm</td>
<td>+50 mm</td>
</tr>
<tr>
<td>+100 mm</td>
<td>$-300$ mm</td>
<td>0 mm</td>
<td>+100 mm = $f_1$</td>
</tr>
<tr>
<td>+125 mm = $f_1$ + $f_2$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>+150 mm</td>
<td>+500 mm</td>
<td>+50 mm</td>
<td>−100 mm</td>
</tr>
<tr>
<td>+175 mm</td>
<td>+300 mm</td>
<td>+37.5 mm</td>
<td>−50 mm</td>
</tr>
<tr>
<td>+200 mm</td>
<td>+233$\frac{1}{3}$ mm</td>
<td>+33$\frac{1}{3}$ mm</td>
<td>−33$\frac{1}{3}$ mm</td>
</tr>
<tr>
<td>+225 mm</td>
<td>+200 mm</td>
<td>+31.25 mm</td>
<td>−25 mm</td>
</tr>
<tr>
<td>+250 mm</td>
<td>+180 mm</td>
<td>+30 mm</td>
<td>−20 mm</td>
</tr>
</tbody>
</table>

10.5.11 Newtonian Form of Imaging Equation

The familiar Gaussian form of the imaging equation is:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

An equivalent form is obtained by defining the distances $x$ and $x'$ measured from the focal points:

$$s = x + f \implies x = s - f$$
$$s' = x' + f \implies x' = s' - f$$

In the case of the real image $O'$ for the real object $O$ shown in the figure, both $x$ and $x'$ are positive because the distances are measured from left to right:

The definition of the parameters $x, x'$ in the Newtonian form of the imaging equation. For a real image, both $x$ and $x'$ are positive.
By simple substitution into the imaging equation, we obtain:

\[
\frac{1}{f} = \frac{1}{x + f} + \frac{1}{x' + f}
\]

\[
\frac{1}{f} = \frac{(x' + f) + (x + f)}{(x + f)(x' + f)}
\]

\[
f = \frac{xx' + (x + x') f + f^2}{(x + x') + 2f}
\]

\[\Rightarrow |(x + x') + 2f| \cdot f = xx' + (x + x') f + f^2
\]

\[\Rightarrow x \cdot x' = f^2\]

This is the Newtonian form of the imaging equation. The same expression applies for virtual images, but the sign of the distances must be adjusted, as shown:

![Diagram of imaging system with positive and negative lenses](image)

The parameters \(x, x'\) of the Newtonian form for a virtual image.

### 10.5.12 Another System of Two Thin Lenses: the Telephoto Lens

Now consider a system composed of a positive lens and a negative lens separated by slightly more than the sum of the focal lengths. For example, consider \(f_1 = +100\) mm, \(f_2 = -25\) mm, and \(t = +80\) mm \(\approx f_1 + f_2 = 75\) mm. The focal length of the equivalent thin lens is easy to calculate:

\[
\frac{1}{f_{eff}} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1 f_2}
\]

\[= \left(\frac{1}{100\text{ mm}} + \frac{1}{-25\text{ mm}} - \frac{80\text{ mm}}{(+100\text{ mm})(-25\text{ mm})}\right)^{-1}
\]

\[f_{eff} = 500\text{ mm} \gg f_1\]
Now locate the image-space focal point and principal point. For an object located at $\infty$, the $BFD$ is found by substitution into the appropriate equation:

$$BFD = \frac{f_2 (t - f_1)}{t - (f_1 + f_2)} = \frac{(-25 \text{ mm})(80 \text{ mm} - 100 \text{ mm})}{80 \text{ mm} - (100 \text{ mm} + (-25 \text{ mm}))} = +100 \text{ mm}$$

The image of an object at $\infty$ is located 100 mm behind the second lens, and thus 180 mm behind the first lens. The physical length of the system when imaging an object at $\infty$ is 180 mm, which is MUCH less than the equivalent focal length $f_{\text{eff}} = 500 \text{ mm}$. This is an example of an optical system whose physical length is much shorter than the focal length. Such a lens is useful for photographers; it is a “short” (and portable) lens system that has a “long” focal length. Note that astronomers also find such systems to be useful.

The locations of the image-space principal point is determined from the focal distance and the equivalent focal length:

$$H'F' = H'V' + V'F'$$

$$500 \text{ mm} = H'V' + 100 \text{ mm}$$

$$H'V' = +400 \text{ mm}$$

$$H'V = H'V' - V'V' = 400 \text{ mm} - 80 \text{ mm} = +320 \text{ mm}$$

so the principal point is located 320 mm in front of the object-space vertex $V$. A sketch of the system and the image-space cardinal points is shown:

![Image-space focal and principal points of the telephoto system. The equivalent focal length of the system is $f_{\text{eff}} = +500 \text{ mm}$, but the image-space focal point is only $+100 \text{ mm}$ behind the rear vertex $V'$. The image-space principal point is $500 \text{ mm}$ in front of the focal point.]

The object-space focal point is located by applying the expression for the “front focal distance”:

$$FFD = FV = \frac{f_1 (t - f_2)}{t - (f_1 + f_2)} = \frac{(+100 \text{ mm})(80 \text{ mm} - (-25 \text{ mm}))}{80 \text{ mm} - (100 \text{ mm} + (-25 \text{ mm}))} = +2100 \text{ mm}$$
which is far in front of the object-space vertex \( V \). The object-space principal point is found from:

\[
\begin{align*}
\text{FH} &= \text{FV} + \text{VH} \\
500 \text{ mm} &= 2100 \text{ mm} + \text{VH} \\
\text{VH} &= 500 \text{ mm} - 2100 \text{ mm} = -1600 \text{ mm} \\
\Rightarrow \text{HV} &= -\text{VH} = +1600 \text{ mm}
\end{align*}
\]

So the object-space principal point is very far in front of the first vertex.

Object-space focal and principal points of the telephoto system. Both are located way out in front of the front vertex.

Locating the Image from the Telephoto lens

We can locate the image of an object at a finite distance (say, \( OD = 3000 \text{ mm} \)) using any of the three methods: “brute-force” calculation, by applying the Gaussian imaging formula for distances measured from the principal points, and form the Newtonian imaging equation.

Gaussian Formula “step by step” The distance from the object to the first thin lens is 3000 mm, so the intermediate image distance satisfies:

\[
\begin{align*}
\frac{1}{s_1} + \frac{1}{s'_1} &= \frac{1}{f_1} \\
1 = \frac{1}{100 \text{ mm}} - \frac{1}{3000 \text{ mm}} \\
= \frac{29}{3000} \text{ mm} \approx 103.45 \text{ mm}
\end{align*}
\]

The transverse magnification of the image from the first lens is:

\[
(M_T)_1 = -\frac{s'_1}{s_1} = -\frac{1}{29}
\]

The object distance to the second lens is negative:

\[
s_2 = t - s'_1 = 80 \text{ mm} - \frac{3000}{29} \text{ mm} = -\frac{680}{29} \text{ mm} \approx -23.45 \text{ mm}
\]
the object is virtual. The image distance from the second lens is:

\[
\frac{1}{s_2} + \frac{1}{s_2'} = \frac{1}{f_2}
\]

\[
s_1' = \left( -\frac{1}{25 \text{ mm}} - \left( -\frac{29}{680 \text{ mm}} \right) \right)^{-1} = +\frac{3400}{9} \text{ mm} \approx +377.8 \text{ mm}
\]

The corresponding transverse magnification is:

\[
(M_T)_2 = -\frac{s_2'}{s_2} = -\frac{3400}{-680} \text{ mm} \approx -16.1
\]

The system magnification is the product of the component transverse magnifications:

\[
M_T = (M_T)_1 \cdot (M_T)_2 = -\frac{1}{29} \cdot -16.1 = -\frac{5}{9}
\]

**Gaussian Formula from Principal Points** Now evaluate the same image using the Gaussian formula for distances measured from the principal points. The distance from the object to the object-space principal point is:

\[
s_1 = \overline{OH} = \overline{OV} + \overline{VH} = 3000 \text{ mm} + (-1600 \text{ mm}) = +1400 \text{ mm}
\]

The image distance measured from the image-space principal point is found from the Gaussian image formula:

\[
\frac{1}{s'} = \frac{1}{f_{eff}} - \frac{1}{s} \rightarrow s' = \overline{H'O'} = \left( \frac{1}{500 \text{ mm}} - \frac{1}{1400 \text{ mm}} \right)^{-1} = +\frac{7000}{9} \text{ mm} \approx 777.8 \text{ mm}
\]

The distance from the rear vertex to the image is found from the known value for \( \overline{HV'} = +400 \text{ mm} \):

\[
\overline{V'O'} = \overline{H'O'} - \overline{HV'} = +\frac{7000}{9} \text{ mm} - 400 \text{ mm} = \frac{3400}{9} \text{ mm} \approx 377.8 \text{ mm}
\]

thus matching the distance obtained using “brute force”. The transverse magnification of the image created by the system is:

\[
M_T = -\frac{s'}{s} = -\frac{+\frac{7000}{9}}{+1400} \text{ mm} = -\frac{5}{9}
\]

**Newtonian Formula** Now repeat the calculation for the image position using the Newtonian lens formula. The distance from the object to the object-space focal point is:

\[
x = \overline{OF} = \overline{OV} + \overline{VF} = \overline{OV} - \overline{FV} = 3000 \text{ mm} - 2100 \text{ mm} = 900 \text{ mm}
\]
Therefore the distance from the image-space focal point to the image is:

\[ x' = FO' = \frac{f'^2}{x} = \frac{(500 \text{ mm})^2}{900 \text{ mm}} = \frac{2500}{9} \text{ mm} \approx 277.8 \text{ mm} \]

So the distance from the rear (image-space) vertex \( V' \) to the image is:

\[ V'O' = VF' + FO' \]
\[ = 100 \text{ mm} + \frac{2500}{9} \text{ mm} = \frac{3400}{9} \text{ mm} \approx 377.8 \text{ mm} \]

which again agrees with the result obtained by the other two methods.

### 10.6 Stops and Pupils

In any multielement optical system, the beam of light that passes through the system is shaped like a circular “spindle” with different radii at different axial locations. The diameter of a specific element limits the size of this spindle of rays that enters or exits the system. This element is the stop of the system and may be a lens or an aperture with no power (i.e., an iris diaphragm) that is placed specifically to limit the ray cone. Obviously, an imaging system composed of a single lens is also the stop of the system. In a two-element system, the stop must be one of the two lenses; which lens is determined by the relative sizes. The image of the stop seen from the input “side” of the lens is the entrance pupil, which determines the extent of the ray cone from the object that “gets into” the optical system, and thus the “brightness” of the image. The image of the stop seen from the output “side” is the exit pupil.

The locations and sizes of the pupils are determined by applying the ray-optics imaging equation to these objects. To some, the concept of finding the image of a lens may seem confusing, but it is no different from before – just think of the lens as a regular opaque object.
A three-lens imaging system where the second lens is the stop, its image seen through the first lens is the entrance pupil, and its image seen through the last lens is the exit pupil.

Consider the stops and pupils of the Galilean telescope. Which element is the stop depends on the relative sizes of the lenses. In the first case shown below, the first lens (the objective) is small enough that it acts as the stop (and thus also the entrance pupil). The image of the objective lens seen through the eyelens is the exit pupil, and is “between” the two lenses and very small. Because the exit pupil is small, so is the field of view of the Galilean telescope. In the second example, the smaller eyelens is the stop and also the exit pupil, while the image of the eyelens seen through the objective is the entrance pupil and is far behind the eyelens and relatively large.

10.6.1 Stop and Pupils of Galilean and Keplerian Telescopes

Consider the two two-lens telescope designs; the Galilean telescope has a positive-power objective and a negative-power ocular or eyelens, while the Keplerian telescope has a positive objective and a positive eyelens. Assume that the objective is identical in the two cases with \( f_1 = +100 \text{ mm} \) and \( d_1 = 30 \text{ mm} \). The focal lengths of the oculars are \( f = \pm 15 \text{ mm} \) and \( d = +15 \text{ mm} \) (these are the approximate dimensions and focal lengths of the lenses in the OSA Optics Discovery Kit). The lenses of a telescope are separated by \( f_1 + f_2 \), or 85 mm for the Galilean and 115 mm for the Keplerian. We want to locate the stops and pupils. The stop is found by tracing a ray from an object at \( \infty \) through the edge of the first element and finding the ray height at the second lens. If this ray height is small enough to pass through the second lens, then the first lens is the stop; if not, then the second lens is the stop.

Consider the Galilean telescope first. The ray height at the first lens is the “semi-diameter” of the lens: \( \frac{d}{2} = 15 \text{ mm} \). From there, the ray height would decrease to
10.6 STOPS AND PUPILS

0 mm at a distance of \( f_1 = +100 \text{ mm} \), but it encounters the negative lens at a distance \( t = +85 \text{ mm} \). The ray height at this lens is \( \frac{100 \text{ mm} - 85 \text{ mm}}{100 \text{ mm}} \cdot 15 \text{ mm} = 2.25 \text{ mm} \), which is much smaller than the lens semidiameter of \( \frac{d_2}{2} = 7.5 \text{ mm} \). Since the ray “bundle” is constrained by the diameter of the first lens, it is the aperture stop of the system.

The entrance pupil is the image of the stop as seen through all of the elements that come before the stop. In this example, the first lens is also the entrance pupil, so the transverse magnification of the entrance pupil is unity.

The exit pupil is the image of the stop through all elements that come afterwards, i.e., just the negative lens. The distance to the “object” is \( f_1 + f_2 = 85 \text{ mm} \), so the imaging equation is used to locate the exit pupil and determine its magnification:

\[
\frac{1}{85 \text{ mm}} + \frac{1}{s'} = \frac{1}{f_2} = \frac{1}{-15 \text{ mm}}
\]

\[
s' = \left( -\frac{1}{15 \text{ mm}} - \frac{1}{85 \text{ mm}} \right)^{-1} = -\frac{51}{4} \text{ mm} = -12.75 \text{ mm}
\]

\[
M_T = \frac{s'}{s} = -\frac{-12.75 \text{ mm}}{85 \text{ mm}} = 0.15
\]

The exit pupil is upright, but more important, it is virtual and thus is not accessible to the eye (you can’t put your eye at the exit pupil of a Galilean telescope).

Follow the same procedure to determine the stop and locate the pupils and their magnifications for the Keplerian telescope. The ray height at the first lens for an object located at \( \infty \) is again 15 mm. The ray height decreases to 0 mm at the focal point, but then decreases still farther until encountering the ocular lens at a distance of \( f_1 + f_2 = 115 \text{ mm} \). The ray height \( h \) at this lens is determined from similar triangles:

\[
\frac{15 \text{ mm}}{-h} = \frac{100 \text{ mm}}{15 \text{ mm}} \implies h = -2.25 \text{ mm}
\]

So the first lens is the stop and entrance pupil (with unit magnification) in this case too. The distance from the stop to the second lens is \( f_1 + f_2 = 115 \text{ mm} \), so the imaging equation for locating the exit pupil and determining its magnification are:

\[
\frac{1}{115 \text{ mm}} + \frac{1}{s'} = \frac{1}{f_2} = \frac{1}{+15 \text{ mm}}
\]

\[
s' = \left( +\frac{1}{15 \text{ mm}} - \frac{1}{115 \text{ mm}} \right)^{-1} = \frac{69}{4} \text{ mm} = +17.25 \text{ mm}
\]

\[
M_T = \frac{s'}{s} = -\frac{+17.25 \text{ mm}}{85 \text{ mm}} \approx -0.203
\]

The exit pupil of a Keplerian telescope is a real image – we can place our eye at it.
Galilean telescope for object at $s_1 = +\infty$: (a) The objective lens is the stop because it limits the cone of entering rays (it also is the entrance pupil). The image of the stop seen through the eyelens is the exit pupil, and is very small; (b) The eyelens is the stop and the exit pupil. The image of the eyelens seen through the objective is the entrance pupil, and is behind the eyelens because the object distance to the objective is less than a focal length.

10.6.2 System $f$-Number

The “brightness” of a recorded image is determined by the ability of the lens to gather light and by the area of the image. This section considers the “light-gathering power” of an optical system that creates a real image, i.e., one that can be placed on a sensor (sheet of film or CCD). To create a real image, the object distance $s$ must be larger than the focal length $f_{\text{eff}}$. The transverse magnification of the system:

$$M_T = -\frac{h'}{h} = -\frac{s'}{s} = -\left(\frac{f_{\text{eff}}}{s_1} - \frac{1}{s}\right)^{-1} = \frac{f_{\text{eff}}}{f_{\text{eff}} - s} = -\frac{f_{\text{eff}}}{s} \left(\frac{1}{1 - \frac{f_{\text{eff}}}{s}}\right)$$

Since $s > f$ to ensure that the image is real, then $\frac{L}{s} < 1$ and the second term can be expanded using the well-known series:

$$\frac{1}{1 - t} = \sum_{n=0}^{\infty} t^n = 1 + t + t^2 + \cdots \text{ if } |t| < 1$$
Hence
\[ M_T = -\frac{f_{eff}}{s} \sum_{n=0}^{\infty} \left( \frac{f_{eff}}{s} \right)^n = -f_{eff} \left( \frac{1}{s} + \frac{f_{eff}}{s^2} + \frac{f_{eff}^2}{s^3} + \cdots \right) \]

Note that \( M_T < 0 \). If \( s >> f_{eff} \) (as often is the case; the object is quite far from the system), then we can truncate the series after the first term and we can say that:
\[ M_T \cong -\frac{f_{eff}}{s} \propto f_{eff} \text{ if } f_{eff} \ll s \]

In words, the transverse magnification is approximately proportional to the focal length for real images of objects located many focal lengths from the system. Note that the transverse magnification refers to the linear dimension, and that the area of the image therefore is proportional to \( f^2 \).

The “brightness” of a recorded image also is determined by the ability of the lens to gather light, which is determined by the area of the entrance pupil. This is (of course) proportional to the square of the diameter \( D \) of the exit pupil and thus the square of the diameter of the aperture stop. If we combine the contributions from the exit pupil and the transverse magnification, we can see that the “flux density” of light (or the “brightness” of the image) is proportional to
\[ \text{flux density} \propto \left( \frac{D}{f_{eff}} \right)^2 \]

The ratio \( D / f_{eff} \) is sometimes called the relative aperture and its reciprocal is the \( f \)-number or \( f \)-ratio of the system:
\[ f/\# \equiv \frac{f_{eff}}{D} \]

For a fixed focal length, a system with a smaller \( f \)-number collects more light and thus produces a brighter image.

### 10.7 Ray Tracing

The imaging equation(s) become quite complicated in systems with more than a very few lenses. However, we can determine the effect of the optical system by ray tracing, where the action on two (or more) rays is determined. Raytracing may be paraxial or exact. Historically, graphical, matrix, or worksheet ray tracing were commonly used in optical design, but most ray tracing is now implemented in computer software so that exact solutions are more commonly implemented than heretofore.

#### 10.7.1 Marginal and Chief Rays

Many important characteristics of an optical system are determined by two specific rays through the system. The marginal ray travels from the optic axis at the center
of the object, just grazes the edge of the stop, and then travels to the center of the
image. The chief (or principal) ray travels from the edge of the object through the
center of the stop to the edge of the image. An image is created wherever the marginal
ray crosses the axis. The chief ray crosses the axis at the stop and the pupils.

The marginal and chief rays for a two-element imaging system where the second
element is the stop. The marginal ray comes from the center of the object \( O \), grazes
the edge of the stop and through the center of the image \( O' \). The chief ray travels
fromt the edge of the object through the center of the stop to the edge of the image.

## 10.8 Paraxial Ray Tracing Equations

Consider the schematic of a two-element optical system:

\[
\begin{align*}
\text{Power} & \quad \text{Power} \\
\varphi_1 & \quad \varphi_2 \\
\begin{array}{c}
\text{H}_1 \quad \text{H}_1' \quad \text{H}_2 \quad \text{H}_2' \\
n = n_1 \quad n' = n_2 \quad n_2' = n_3
\end{array}
\end{align*}
\]

\[
\begin{align*}
n_1 u_1 &= 0 \\
y_1 & \\
\begin{array}{c}
n_1' u_1' = n_2 u_2 \\
y_2 \\
t_2 \\
n_3 u_3
\end{array}
\end{align*}
\]

Schematic of rays in ray tracing, using the marginal ray as an example. The ray
height at the \( n^{th} \) element is \( y_n \), and the ray angle during transfer between elements
\( n - 1 \) and \( n \) is \( u_n \). The system has two elements, represented by the pairs of
principal planes.

The two elements are represented by their two principal “planes”, which are the
planes of unit magnification. The refractive power of the first element changes the
ray angle of the input ray. In the example shown, the input ray angle \( u_1 = 0 \) radians,
i.e., the ray is parallel to the optical axis. The height of this ray above the axis at the object-space principal plane $H_1$ is $y_1$ units. The ray emerges from the principal plane $H'_1$ at the same height $y_1$ but with a new ray angle $u_2$. The ray “transfers” to the second element through the distance $t_2$ in the index $n_2$ and has ray height $y_2$ at principal plane $H_2$. The ray emerges from the principal plane at the same height but a new angle $u_3$.

### 10.8.1 Paraxial Refraction

Consider refraction of a paraxial ray emitted from the object $O$ at a surface with radius of curvature $R$. For a paraxial ray, the surface may be drawn as “vertical”. The height of the ray at the surface is $y$. The media before and after the surface are respectively $n$ and $n'$.

![Paraxial Refraction Diagram]

Refraction of a paraxial ray at a surface with radius of curvature $R$ between media with refractive indices $n$ and $n'$. The ray height and angle at the surface are $y$ and $u$, respectively. The angle of the ray measured at the center of curvature is $\alpha$. The height and angle immediately after refraction are $y_0$ and $u'_0$. The object and image distances are $s$ and $s'$.

From the drawing, the incoming ray angle $u$ is:

$$u = \tan^{-1} \left( \frac{y}{s} \right) \cong \frac{y}{s} > 0$$

The the outgoing ray angle $u'$ also is positive:

$$u' = \tan^{-1} \left( \frac{y'}{s'} \right) \cong \frac{y'}{s'} > 0$$

and the angle of the refraction measure from the center of curvature is negative:

$$\alpha = -\tan^{-1} \left( \frac{y}{R} \right) \cong -\frac{y}{R} < 0$$
From the drawing, we can see that:

\[ i = |u| + |\alpha| = u - \alpha > 0 \]

Snell’s law also tells us that:

\[
\begin{align*}
    n \sin [i] &= n' \sin [i'] \quad \Rightarrow \quad n \ i \cong n' \ i' \\
    n \sin [i] &= n \sin [u - \alpha] \cong n [u - \alpha] \\
    n' \sin [i'] &= n' \sin [u' - \alpha] \cong n' [u' - \alpha] \\
    \quad \Rightarrow n [u - \alpha] \cong n' [u' - \alpha] \\
    \quad \Rightarrow n'u' \cong nu - n\alpha + n'\alpha = nu + \alpha (n' - n)
\end{align*}
\]

\[ n'u' \cong nu + \alpha (n' - n) \]

\[ \cong nu - y \left(\frac{n' - n}{R}\right) = nu - y\varphi \]

The paraxial refraction equation in terms of the incident angle \( u \), refracted angle \( u' \), ray height \( y \), surface power \( \varphi = \frac{1}{f} \), and indices of refraction \( n \) and \( n' \) is:

\[ n'u' = nu - y\varphi \]

### 10.8.2 Paraxial Transfer

The paraxial transfer equation: the ray traverses the distance \( t \) in the medium with index \( n' \). The initial and final ray heights are \( y \) and \( y' \), respectively. The angle is \( u' = \tan^{-1} \left[ \frac{y'}{t} \right] \Rightarrow y' = y + tu' = y + \frac{1}{n'} (n'u') \)

The transfer equation determines the ray height \( y' \) at the next surface given the initial ray height \( y \), the distance \( t \), and the angle \( u' \). From the drawing, we have:

\[ y' = y + tu' \]

\[ y' = y + \left(\frac{1}{n'}\right) (n'u') \]
where the substitution was made to put the ray angle in the same form \( n'u' \) that appeared in the refraction equation. The distance \( \frac{t}{n'} \leq t \) is called the reduced thickness.

### 10.8.3 Linearity of the Refraction and Transfer Equations

Note that both the refraction and transfer equations are linear in the height and angle, i.e., neither includes any operations involving squares or nonlinear functions (such as sine, logarithm, or tangent). Among other things, this means that they may be scaled by direct multiplication to obtain other “equivalent” rays. For example, the output angle may be scaled by scaling the input ray angle and the height by a constant factor \( \alpha \):

\[
\alpha (nu - y\varphi) = \alpha \cdot (nu) - (\alpha \cdot y) \varphi = \alpha (n'u')
\]

We will take often advantage of this linear scaling property to scale rays to find the exact marginal and chief rays from the provisional counterparts.

### 10.8.4 Paraxial Ray Tracing

To characterize the paraxial properties of a system, two provisional rays are traced:

1. Initial ray height (at first surface) \( y = 1.0 \), initial angle \( nu = 0 \)

2. Initial ray height (at first surface) \( y = 0.0 \), initial angle \( \nu = 1 \)

We have already named these rays; the first is the provisional marginal ray that intersects the optical axis at the object (and thus also at every image of the object). The second ray is called the provisional chief (or principal) ray and travels from the edge of the object to the edge of the field of view through the center of the stop (and thus through the centers of the pupils, which are images of the stop).

The process of ray tracing is perhaps best introduced by example. Consider a two-element three-surface system with three surfaces. The three radii of curvature are \( R_1 = +7.8 \text{ mm}, \) \( R_2 = +10 \text{ mm} \), and \( R_3 = -6 \text{ mm} \). The distance between the first two surfaces (the thickness of the first element) and between the second and third surfaces are both 3.6 mm. The refractive index between the first two surfaces is \( n_2 = 1.336 \) and between the second and third surfaces is \( n_3 = 1.413 \). The index after the last surface is \( n_4 = n_2 = 1.336 \).
The first action of the system is paraxial refraction at the first surface, which changes the ray angle but not the ray height. The new ray angle for the provisional marginal ray is:

\[
(n'u')_1 = (nu)_1 - y_1 \text{[mm]} \cdot \varphi_1 \text{[mm}^{-1}] \\
= 0 - (1.0) (+0.043077) \\
= -0.043077 \text{ radian}
\]

(note the retention of 7 decimal places; after cascading the large number calculations for a complex system, the precision of the final results will be significantly poorer).

The paraxial transfer equation for the provisional marginal ray between the first and second surface changes the height of the ray but not the angle. The height at the second surface is:

\[
y'_1 = y_1 + \left( \frac{t'}{n'} \right)_1 (n'u')_1 \text{[mm]} \\
= 1 + \frac{3.6}{1.336} (-0.043077) = +0.883924 \text{ mm}
\]

Thus the ray exits the first surface at the “reduced angle” \( n'u' \cong -0.04 \) radians and arrives at the second surface at height \( y' \cong +0.88 \) mm. The corresponding equations for the chief ray at the first surface are:

\[
(n'u')_1 = (n\pi)_1 - \varphi_1 \\
= 1 - (0.0) (+0.043077) \\
= 1 \text{ radian}
\]
\[ \bar{y}_1 = \bar{y}_0 + \left( \frac{t'}{n'} \right)_1 (n' \bar{u})_1 = 0 + 3.6 \frac{1}{1.336} (1) = +2.694611 \text{ mm} \]

Since the provisional chief ray went through the center of the first surface \( (\bar{y}_1 = 0) \), the ray angle \( n \bar{u} \) did not change. The height of the chief ray at the second surface \( (\bar{y}_1 = \bar{y}_2) \) is proportional to the initial ray angle \( \bar{y}_1 \).

The equations are evaluated in sequence to compute the rays through the system. These are presented in the table. Each column in the table represents a surface in the system and the “primed” quantities refer to distances and angles following the surface. In words, \( t' \) in the first row are the distances from the surface in the column to the next surface.

<table>
<thead>
<tr>
<th>( R )</th>
<th>( t' )</th>
<th>( n' )</th>
<th>( -\varphi = -\frac{n'-n}{R} )</th>
<th>( \frac{t'}{n'} )</th>
<th>( y )</th>
<th>( n'u' )</th>
<th>( \bar{y} )</th>
<th>( n'\bar{u} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>3.6 mm</td>
<td>1.336</td>
<td>-0.043077 mm(^{-1})</td>
<td>( \frac{3.6 \text{ mm}}{1.336} = 2.694611 \text{ mm} )</td>
<td>1 mm</td>
<td>-0.043077 radian</td>
<td>0 mm</td>
<td>1 radian</td>
</tr>
<tr>
<td>10.0 mm</td>
<td>1.343</td>
<td>1.413</td>
<td>-0.007700 mm(^{-1})</td>
<td>( \frac{3.6 \text{ mm}}{1.413} = 2.54771 \text{ mm} )</td>
<td>0.883924 mm</td>
<td>-0.049883 radian</td>
<td>0.756833 mm</td>
<td>0.979251 radian</td>
</tr>
<tr>
<td>-6.0 mm</td>
<td>1.336</td>
<td>0.012833 mm(^{-1})</td>
<td></td>
<td></td>
<td>0 mm</td>
<td>-0.059596 radian</td>
<td>-0.060 radian</td>
<td></td>
</tr>
</tbody>
</table>

The height and angle of the provisional chief ray at the image location are \( \bar{y} \approx 16.78 \text{ mm} \) and \( n' \bar{u} \approx 0.91 \text{ radians} \), respectively.
This system is often used as a model for the human eye if the lens is relaxed to view objects at $\infty$. The first surface is the cornea of the eye, while the other two surfaces are the front and back of the lens. Note that the power of the cornea ($0.043077 \text{mm}^{-1} = 43$ diopters) is considerably larger than the powers of the lens surfaces (7.7 and 12.8 diopters, respectively); in other words, most of the refraction of the eye system occurs at the cornea.

### 10.8.5 Matrix Formulation of Paraxial Ray Tracing

The same linear paraxial ray tracing equations may be conveniently implemented as matrices acting on ray vectors for the marginal and chief rays whose components are the height and angle. The ray vectors may be defined as:

$$
\begin{bmatrix}
y \\
u
\end{bmatrix}, \quad \begin{bmatrix}
y' \\
u'
\end{bmatrix}
$$

Note that there is nothing magical about the convention for the ordering of $y$ and $nu$; this is the convention used by Roland Shack at the Optical Sciences Center at the University of Arizona, but Willem Brouwer wrote a book on matrix methods in optics that uses the opposite order (which Hecht also uses).

These column vectors may be combined to form a ray matrix $L$, where the columns are the marginal and chief ray vectors:

$$
L \equiv \begin{bmatrix}
y & y' \\
u & \nu'
\end{bmatrix}
$$

which may be evaluated at any point in the system. The determinant of this ray matrix is the so-called Lagrange Invariant (which we will denote by the symbol $\aleph$, because it is the closest character available to the Cyrillic character usually used). As suggested by its name, the Lagrange Invariant is unaffected either by refraction or transfer all of the way through the system.

$$
\det(L) = y \cdot (nu) - (nu) \cdot y' \equiv \aleph
$$

**Refraction Matrix**

Given the ray vectors or the ray matrix, we can now define operators for refraction and transfer. Recall that paraxial refraction of a marginal ray and of a chief ray at a surface with power $\varphi$ is:

$$
n'\nu' = nu - y\varphi \quad \text{for marginal ray}$$

$$
n'\nu' = n\nu - \nu'\varphi \quad \text{for chief ray}
$$
The refraction process for the marginal ray may be written as a matrix $\mathcal{R}$ and the output is the product with the ray vector which will have the same ray height and a different angle:

$$\mathcal{R} \begin{bmatrix} y \\ nu \end{bmatrix} = \begin{bmatrix} y \\ n'u' \end{bmatrix}$$

It is easy to see that the form of the matrix must be:

$$\begin{bmatrix} 1 & 0 \\ -\varphi & 1 \end{bmatrix} \begin{bmatrix} y \\ nu \end{bmatrix} = \begin{bmatrix} y \\ -y\varphi + nu \end{bmatrix} = \begin{bmatrix} y \\ n'u' \end{bmatrix}$$

and its determinant is unity:

$$\det \begin{bmatrix} 1 & 0 \\ -\varphi & 1 \end{bmatrix} = (1)(1) - (-\varphi)(0) = 1$$

**Transfer Matrix**

The transfer of the marginal ray from one surface to the next is $y' = y + \frac{\nu'}{n'} (n'u')$, which also may be written as the product of the matrix $\mathcal{T}$ with the ray vector:

$$\mathcal{T} \begin{bmatrix} y \\ nu \end{bmatrix} = \begin{bmatrix} y + (nu) \left( \frac{\nu'}{n'} \right) \\ nu \end{bmatrix} = \begin{bmatrix} 1 & \frac{\nu'}{n'} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y \\ nu \end{bmatrix} = \begin{bmatrix} y' \\ nu \end{bmatrix}$$

so the determinant of the transfer matrix is also 1:

$$\det \begin{bmatrix} 1 & \frac{\nu'}{n'} \\ 0 & 1 \end{bmatrix} = (1)(1) - (0) \left( \frac{\nu'}{n'} \right) = 1$$

Note that we could operate on the ray matrix instead of individual ray vectors: this allows us to calculate both the marginal and chief rays at the same time:

$$\mathcal{R}\mathcal{L} = \begin{bmatrix} 1 & 0 \\ -\varphi & 1 \end{bmatrix} \begin{bmatrix} y & \overline{y} \\ nu & n\overline{u} \end{bmatrix} = \begin{bmatrix} y & \overline{y} \\ nu - y\varphi & n\overline{u} - \overline{y}\varphi \end{bmatrix}$$

$$\mathcal{T}\mathcal{L} = \begin{bmatrix} 1 & \frac{\nu'}{n'} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y & \overline{y} \\ nu & n\overline{u} \end{bmatrix} = \begin{bmatrix} y + (\frac{\nu'}{n'}) (nu) & \overline{y} + (\frac{\nu'}{n'}) n\overline{u} \\ nu & n\overline{u} \end{bmatrix}$$
The refraction and transfer matrices may be combined in sequence to model a complete system. If we start with the marginal ray vector at the input object, the first operation is transfer to the first surface. The next is refraction by that surface, transfer to the next, and so forth until a final transfer to the output image:

\[
T_n R_n \cdots T_2 R_2 T_1 R_1 T_0 \left( L_{\text{object}} \right) = L_{\text{out}}
\]

If the initial ray matrix is located at the object (as usual), the marginal ray height is zero, so the ray matrix at the object and any images has the form:

\[
\begin{bmatrix}
0 & y_{in} \\
(nu)_{in} & (n\varpi)_{in}
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
0 & y_{out} \\
(nu)_{out} & (n\varpi)_{out}
\end{bmatrix}
\]

so the system from object to image is:

\[
T_n R_n \cdots T_2 R_2 T_1 R_1 T_0
\begin{bmatrix}
0 & y_{in} \\
(nu)_{in} & (n\varpi)_{in}
\end{bmatrix}
= 
\begin{bmatrix}
0 & y_{out} \\
(nu)_{out} & (n\varpi)_{out}
\end{bmatrix}
\]

The matrices appear to be laid out in inverse order, i.e., the last matrix first, but the transfer matrix \( T_0 \) acts on the input ray matrix, so it must appear on the right.

### Ray Matrix for Provisional Marginal and Chief Rays

The system is characterized by using provisional marginal and chief rays located at the object. The linearity of the computations ensure that the rays may be scaled subsequently to satisfy other system constraints, such as the diameter of the stop. The provisional marginal ray at the object has height \( y = 0 \) and ray angle \( nu = \), while the provisional chief ray at the object has height \( \varpi = 1 \) and angle \( n\varpi = 0 \). Thus the provisional ray matrix at the object is:

\[
L_0 = \begin{bmatrix}
0 & 1 \\
1 & 0
\end{bmatrix}
\]

### “Vertex-to-Vertex Matrix” for System

The optical system matrix excludes the input ray matrix, the first transfer matrix, the last transfer matrix, and the output ray matrix. It is called the “vertex-to-vertex matrix” and is labeled \( M_{VV'} \)

\[
M_{VV'} = R_N T_{N-1} \cdots T_2 R_2 T_1 = \begin{bmatrix}
A & B \\
C & D
\end{bmatrix}
\]
where $A,B,C,D$ are real numbers to be determined. Since the determinant of the matrix product is the product of the determinants, we can see that the determinant of the vertex-to-vertex matrix must be 1:

$$\det \mathcal{M}_{VV'} = 1 \implies AD - BC = 1$$

For example, find $\mathcal{M}_{VV'}$ for a two-lens system with powers $\varphi_1 = (f_1)^{-1}$ and $\varphi_2 = (f_2)^{-1}$ separated by $t$:

$$\mathcal{M}_{VV'} = R_2 T_1 R_1$$

$$= \begin{bmatrix} 1 & 0 \\ -\varphi_2 & 1 \end{bmatrix} \begin{bmatrix} t \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\varphi_1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \varphi_1 t & t \\ - (\varphi_1 + \varphi_2 - \varphi_1 \varphi_2 t) & 1 - \varphi_2 t \end{bmatrix}$$

So that $A$ and $D$ are “pure” numbers, while $B$ and $D$ have dimensions of length and reciprocal length, respectively. It is easy to confirm that the determinant of this system matrix is unity. From the values in the last section, we can identify two of the matrix components to be:

$$t = B$$
$$f = -\frac{1}{C}$$

To find the “back focal distance” $\mathbf{VF'}_0$ we need to throw in a ray from an object at infinity and find its height and angle at $\mathbf{V'}$:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} y' \\ n'u' \end{bmatrix} = \begin{bmatrix} A \\ C \end{bmatrix}$$

The ray height is $A$ and the angle is $n'u' = C$. The distance to the point where the ray height is zero is the back focal distance:

$$BFD = \mathbf{VF'}_0 = T \begin{bmatrix} y' \\ n'u' \end{bmatrix} = \begin{bmatrix} 1 & \frac{n'}{n} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y' \\ n'u' \end{bmatrix} = \begin{bmatrix} 0 \\ n'u' \end{bmatrix}$$

$$\implies y' + \frac{t'}{n} \cdot n'u' = 0$$

$$\implies \frac{t'}{1} = -\frac{y'}{n'u'} = -\frac{A}{C} = BFD$$

Note that if the image-space refractive index is $n \neq 1$, the distances have to be scaled by the index of the image-space medium (call it $n_i$); see the table below: Combine
the BFD with the knowledge that $H F' = f_{\text{eff}} = -\frac{1}{C}$ to find $H V'$:

$$\begin{align*}
H F' = H V' + V F' &\implies H V' = H F' - V F' \\
H V' = -\frac{1}{C} - \left(-\frac{A}{C}\right) &\implies \frac{A-1}{C} = H V'
\end{align*}$$

To find the equivalent distances in object space, we can trace the “provisional” marginal ray “backwards” through the system, or trace it through the “reversed” system where the lenses are placed in the opposite order. The “reversed” system matrix is:

$$\mathcal{M}_{VV} = R_1 T_1 R_2 T_2 \cdots T_{N-1} R_N$$

In the simple 2-element case, this is:

$$\mathcal{M}_{VV} = R_1 T_1 R_2$$

$$= \begin{bmatrix}
1 & 0 \\
-\varphi_1 & 1
\end{bmatrix}
\begin{bmatrix}
1 & t \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-\varphi_2 & 1
\end{bmatrix}
= \begin{bmatrix}
1 - \varphi_2 t & t \\
- (\varphi_1 + \varphi_2 - \varphi_1 \varphi_2 t) & 1 - \varphi_1 t
\end{bmatrix}$$

If we substitute the same values for $A, B, C, D$ that were evaluated for $\mathcal{M}_{VV^\prime}$, we obtain the reversed matrix:

$$\mathcal{M}_{V'V} = \begin{bmatrix}
D & B \\
C & A
\end{bmatrix}$$

The “BFD” of the reversed system is the FFD

$$\begin{bmatrix}
D & B \\
C & A
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
= \begin{bmatrix}
y' \\
n' u'
\end{bmatrix}
\implies FFD = VF' = -\frac{D}{C}$$

Because the focal length $f_{\text{eff}}$ is still $-\frac{1}{C}$, we can find the distance from the object-space vertex to the object-space principal point:

$$FH = FV + VH \implies VH = FH - FV = -\frac{1}{C} - \left(-\frac{D}{C}\right) = \frac{D-1}{C}$$

Again, if the object space index is not unity, the distances have to be scaled by the object-space refractive index.

We have four equations in the four unknowns $A, B, C, D$, which may be combined to find useful system metrics in terms of the elements in the vertex-to-vertex matrix $\mathcal{M}_{VV^\prime}$, where the refractive indices of object and image space are $n_o$ and $n_i$. 
10.8 PARAXIAL RAY TRACING EQUATIONS

<table>
<thead>
<tr>
<th>Equation Description</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>effective focal length of system</td>
<td>$f_{\text{eff}} = -\frac{1}{C}$</td>
</tr>
<tr>
<td>front focal distance</td>
<td>$FFD = \frac{FV}{n_o} = -\frac{D}{C}$</td>
</tr>
<tr>
<td>back focal distance</td>
<td>$BFD = \frac{VF'}{n_i} = -\frac{A}{C}$</td>
</tr>
<tr>
<td>distance from front vertex to object-space principal point</td>
<td>$\frac{VH}{n_o} = \frac{D-1}{C}$</td>
</tr>
<tr>
<td>distance from image-space principal point to rear vertex</td>
<td>$\frac{HV'}{n_i} = \frac{A-1}{C}$</td>
</tr>
</tbody>
</table>

**Example:** To illustrate, consider the system of two thin lenses in the last section with $f_1 = 100 \text{ mm}$, $f_2 = 50 \text{ mm}$, and $t = 75 \text{ mm}$, which we showed to have $f_{\text{eff}} = +\frac{200}{3} \text{ mm} \approx 66.7 \text{ mm}$. The system matrix is:

$$M_{VV'} = \begin{bmatrix} 1 & 0 & 175 \text{ mm} \\ -\frac{1}{50 \text{ mm}} & 1 & -\frac{1}{100 \text{ mm}} & 1 \end{bmatrix}$$

The various metrics of the system are:

$$f_{\text{eff}} = -\frac{1}{C} = - \left( -\frac{200}{3} \text{ mm} \right) = 66\frac{2}{3} \text{ mm}$$

$$FFD = \frac{FV}{n_o} = \frac{FV}{n_o} = -\frac{D}{C} = - \left( -\frac{1}{3} \frac{200 \text{ mm}}{200 \text{ mm}} \right) = -\frac{100}{3} \text{ mm}$$

$$BFD = \frac{VF'}{n_i} = \frac{VF'}{n_i} = -\frac{A}{C} = - \left( \frac{1}{3} \frac{200 \text{ mm}}{200 \text{ mm}} \right) = +\frac{50}{3} \text{ mm}$$

$$\frac{VH}{n_o} = VH = \frac{D-1}{C} = -\frac{1}{\frac{3}{200 \text{ mm}}} = +100 \text{ mm}$$

$$\frac{HV'}{n_i} = HV = \frac{A-1}{C} = -\frac{1}{\frac{3}{200 \text{ mm}}} = +50 \text{ mm}$$

**Locating the Image Point** The input ray matrix consists of the provisional marginal and chief ray at the object, which “passes through” the transfer matrix from object to front surface. If the object is located 1000 mm from the first surface, the ray matrix at the front vertex of the system is:

$$T_0 \begin{bmatrix} y \\ nu \end{bmatrix} = T_0 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1000 \text{ mm} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1000 \text{ mm} \\ 1 \end{bmatrix}$$
The height of the provisional marginal ray at the front vertex is 1000 mm and the angle is 1 radian (which is a huge angle, but remember that all equations are linear, so the angle and ray height can be scaled to any value). The emerging provisional marginal ray is:

\[
\begin{bmatrix}
\frac{1}{4} & 75 \text{ mm} \\
-\frac{3}{200 \text{ mm}} & -\frac{1}{2}
\end{bmatrix}
\begin{bmatrix}
1000 \text{ mm} \\
1
\end{bmatrix}
= \begin{bmatrix}
325 \text{ mm} \\
-\frac{31}{2}
\end{bmatrix}
= \begin{bmatrix}
y \\
u
\end{bmatrix}
\]

In words, the marginal ray from an object 1000 mm in front of the lens emerges with height 325 mm and angle of $-\frac{31}{2}$ radians. To find the location of the image, find the distance until the marginal ray height $y = 0$:

\[
\nabla' O' = T \begin{bmatrix}
325 \text{ mm} \\
-\frac{31}{2}
\end{bmatrix} = \begin{bmatrix}
1 & \frac{v'}{n'} \\
0 & 1
\end{bmatrix} \begin{bmatrix}
325 \text{ mm} \\
-\frac{31}{2}
\end{bmatrix} = \begin{bmatrix}
0 \\
-\frac{31}{2}
\end{bmatrix}
\]

\[
\Rightarrow 325 \text{ mm} + \left( -\frac{31}{2} \frac{t'}{n'} \right) = 0
\]

\[
\Rightarrow \frac{t'}{1} = 325 \text{ mm} \cdot 2 + \frac{650}{31} \text{ mm}
\]

\[
\nabla' O' \approx +20.97 \text{ mm}
\]

which agrees with the result obtained earlier. We observed that the magnification of the image in this configuration is

\[
-\frac{s'}{s} = \frac{O'V'}{O'V} = -\frac{2}{31}
\]

so the provisional marginal ray at the image point has the form:

\[
\begin{bmatrix}
y' \\
n' u'
\end{bmatrix}
= \begin{bmatrix}
0 \\
-\frac{31}{2}
\end{bmatrix}
= \begin{bmatrix}
0 \\
\frac{1}{M_T}
\end{bmatrix}
\]

The marginal ray out of the vertex-to-vertex matrix for the object distance $OV = 1000$. 

\[
O
\]

\[
O'
\]

Imaging System

$y = 1000$

$y = 325$

$\frac{31}{2}$

20.97
Matrices that Associate Conjugate Points

We can write a general surface-to-surface matrix $\mathcal{M}_{VV'}$ in the form:

$$
\mathcal{M}_{VV'} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}
$$

where the four coefficients are to be determined. The matrix that relates two image planes $O$ and $O'$ may be obtained by adding transfer matrices for the appropriate distances from the object to the front vertex ($t_1 = \mathbf{OV}$) and from the rear vertex to the image ($t_2 = \mathbf{V'O}$).

$$
\mathcal{M}_{OO'} = \begin{bmatrix} 1 & t_2 \\ 0 & 1 \end{bmatrix} \mathcal{M}_{VV'} \begin{bmatrix} 1 & t_1 \\ 0 & 1 \end{bmatrix}
$$

$$
= \begin{bmatrix} 1 & t_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ 0 & 1 \end{bmatrix}
$$

$$
= \begin{bmatrix} A + t_2C & (A + t_2C)t_1 + B + t_2D \\ C & Ct_1 + D \end{bmatrix}
$$

We know that the marginal ray heights at the object and image are zero, which thus sets some limits on the “conjugate-to-conjugate” matrix:

$$
\begin{bmatrix} 0 & \mathbf{y}_\text{in} \\ (nu)_\text{in} & (n\mathbf{u})_\text{in} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{y}_\text{out} \\ (nu)_\text{out} & (n\mathbf{u})_\text{out} \end{bmatrix}
$$

$$
\Rightarrow \begin{bmatrix} A + t_2C & (A + t_2C)t_1 + B + t_2D \\ C & Ct_1 + D \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{y}_\text{out} \\ (nu)_\text{out} & (n\mathbf{u})_\text{out} \end{bmatrix} \cdot \begin{bmatrix} 0 & \mathbf{y}_\text{in} \\ (nu)_\text{in} & (n\mathbf{u})_\text{in} \end{bmatrix}^{-1}
$$

$$
\Rightarrow \begin{bmatrix} \frac{(n\mathbf{u})_\text{out}}{\mathbf{y}_\text{in}} & \frac{(n\mathbf{u})_\text{out} - (nu)_\text{out}(n\mathbf{u})_\text{in}}{\mathbf{y}_\text{in}(nu)_\text{in}} \\ \frac{(n\mathbf{u})_\text{out}}{\mathbf{y}_\text{in}} & \frac{(nu)_\text{out}}{\mathbf{y}_\text{in}} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{y}_\text{out} \\ (nu)_\text{out} & (n\mathbf{u})_\text{out} \end{bmatrix}
$$

The ratio $\frac{\mathbf{y}_\text{out}}{\mathbf{y}_\text{in}} \equiv M_T$, whereas the ratio $\frac{(nu)_\text{out}}{(nu)_\text{in}} = \frac{1}{M_T}$ may be recast by applying the Lagrangian invariant:

$$
C = \frac{(n\mathbf{u})_\text{out}}{\mathbf{y}_\text{in}} \frac{(nu)_\text{in}}{(n\mathbf{u})_\text{out}} - \frac{(nu)_\text{out}}{(n\mathbf{u})_\text{in}} \frac{(n\mathbf{u})_\text{out}}{\mathbf{y}_\text{in}}
$$

$$
= \frac{(n\mathbf{u})_\text{out} - (nu)_\text{out}(n\mathbf{u})_\text{in}}{(n\mathbf{u})_\text{out}} \frac{(nu)_\text{in}}{\mathbf{y}_\text{in}}
$$
\[ N \equiv y \cdot (n\bar{u}) - (nu) \cdot \bar{y} = y_{\text{out}} \cdot (n\bar{u})_{\text{out}} - (nu)_{\text{out}} \cdot \bar{y}_{\text{out}} = y_{\text{in}} \cdot (n\bar{u})_{\text{in}} - (nu)_{\text{in}} \cdot \bar{y}_{\text{in}} \]

\[ \implies \frac{y_{\text{out}} \cdot (n\bar{u})_{\text{out}}}{y_{\text{in}} \cdot (n\bar{u})_{\text{in}} - (nu)_{\text{in}} \cdot \bar{y}_{\text{in}}} - \frac{(nu)_{\text{out}} \cdot \bar{y}_{\text{out}}}{y_{\text{in}} \cdot (n\bar{u})_{\text{in}} - (nu)_{\text{in}} \cdot \bar{y}_{\text{in}}} = 1 \]

The conjugate-to-conjugate matrix includes the leading and following ray transfers:

\[ M_{\text{oo}} = \begin{bmatrix} A + t_2 C & (A + t_2 C) t_1 + B + t_2 D \\ C & Ct_1 + D \end{bmatrix} \]

\[ = \begin{bmatrix} A + t_2 C & (A + t_2 C) t_1 + B + t_2 D \\ C & Ct_1 + D \end{bmatrix} \]

\[ \implies M_T = A + t_2 C = (Ct_1 + D)^{-1} \]

\[ \varphi = -C \]

\[ 0 = (A + t_2 C) t_1 + B + t_2 D \]

We can now add expressions for the distances from object to front vertex and from rear vertex to image in terms of \(A,B,C,D\): have four equations in the four unknowns \(A,B,C,D\), which may be combined to find useful systems metrics in terms of the elements in the vertex-to-vertex matrix \(M_V\):

<table>
<thead>
<tr>
<th>distance from object to front vertex</th>
<th>( \frac{OV}{n_o} = \frac{t_1}{n_o} = \frac{D - \frac{1}{C}}{A + Ct_2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance from rear vertex to image</td>
<td>( \frac{V'O'}{n_i} = \frac{t_2}{n_i} = \frac{m - A}{C} = \frac{B - At_1}{D - Ct_1} )</td>
</tr>
</tbody>
</table>

The matrix that relates the object and image planes for the two-lens system presented above is:

\[ M_{\text{oo'}} = T_2 M_{V'V} T_1 = \begin{bmatrix} 1 & 650 & 31 \\ 0 & 75 & 1 \\ 0 & 3 \frac{3}{200} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \frac{1}{4} \\ -3 \frac{3}{200} & -\frac{1}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 \end{bmatrix} \]

\[ = \begin{bmatrix} -\frac{2}{31} & 0 \\ -3 \frac{3}{200} & -\frac{31}{2} \end{bmatrix} \]

which has the form of the principal plane matrix except the diagonal elements are not both unity. However, note that they are reciprocals of each other, so that

\[ \det \begin{bmatrix} -\frac{2}{31} & 0 \\ -3 \frac{3}{200} & -\frac{31}{2} \end{bmatrix} = 1 \]
We had evaluated the transverse magnification in this configuration to be $-\frac{2}{31}$, so we note that the upper-left component of the conjugate-to-conjugate matrix is the transverse magnification. The general form of a conjugate-to-conjugate matrix is:

$$
\mathcal{M}_{\text{conjugate}} = \begin{bmatrix}
M_T & 0 \\
-\varphi \frac{1}{M_T}
\end{bmatrix}
$$

For the two-lens system that we have used as an example with the object located 1000 mm in front of the first lens, the conjugate-to-conjugate matrix is

$$
M_{OO'} = \begin{bmatrix}
1 + \frac{650}{31} \text{mm} & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
\frac{1}{4} & 75 \text{mm} \\
-\frac{3}{200 \text{mm}} & -\frac{1}{2}
\end{bmatrix}
\begin{bmatrix}
1 & \frac{1}{n} \\
0 & 1
\end{bmatrix}
= \begin{bmatrix}
1 + \frac{650}{31} \text{mm} & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 1000 \text{mm} \\
0 & 1
\end{bmatrix}
= \begin{bmatrix}
-\frac{2}{31} & 0 \\
-\frac{3}{200 \text{mm}} & -\frac{31}{2}
\end{bmatrix}
$$

"Principal Point-to-Principal Point Matrix" for System

The conjugate-to-conjugate (object-to-image) matrix that relates the principal points of unit magnification is:

$$
M_{HH'} = \begin{bmatrix}
1 & 0 \\
-\varphi & 1
\end{bmatrix}
$$

where $\varphi$ is the power of the system, which is the ability to deviate incoming rays.

10.8.6 Examples of System Matrices:

"Vertex-to-Vertex" Matrix for Telephoto Lens

To illustrate, calculate this matrix for the thin-lens telephoto considered in the last section with $f_1 = 100$ mm, $f_2 = -25$ mm, and $t = 80$ mm. The system matrix is:

$$
\mathcal{M}_{VV'} = \begin{bmatrix}
1 - \varphi_1 t & t \\
-(\varphi_1 + \varphi_2 - \varphi_1 \varphi_2) & 1 - \varphi_2 t
\end{bmatrix}
= \begin{bmatrix}
1 & 0 \\
-\frac{1}{25 \text{mm}} & 1
\end{bmatrix}
\begin{bmatrix}
1 & 80 \text{mm} \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-\frac{1}{100 \text{mm}} & 1
\end{bmatrix}
= \begin{bmatrix}
\frac{1}{5} & 80 \text{mm} \\
-\frac{1}{500 \text{mm}} & \frac{21}{5}
\end{bmatrix}
$$
CHAPTER 10 IMAGE FORMATION IN THE RAY MODEL

\[ f_{\text{eff}} = - \frac{1}{C} = - \left( \frac{1}{\frac{1}{500\text{mm}}} \right) = +500\text{ mm} \]

\[ FFD = \bar{F}V = - \frac{D}{C} = - \frac{\left( \frac{21}{5} \right)}{\left( -\frac{1}{500\text{mm}} \right)} = +2100\text{ mm} \]

\[ BFD = \nabla FV = - \frac{A}{C} = - \frac{\left( \frac{1}{5} \right)}{\left( -\frac{1}{500\text{mm}} \right)} = +100\text{ mm} \]

\[ \nabla H = \frac{D - 1}{C} = \frac{\left( \frac{21}{5} \right) - 1}{\left( -\frac{1}{500\text{mm}} \right)} = -1600\text{ mm} \]

\[ H'V = \frac{A - 1}{C} = \frac{\left( \frac{1}{5} \right) - 1}{\left( -\frac{1}{500\text{mm}} \right)} = +400\text{ mm} \]

These agree with the values obtained by “brute force” previously.

Galilean Telescope made of Thin Lenses

The Galilean telescope consists of an objective lens with positive power and an eyepiece with negative power separated by the sum of the focal lengths. If the focal length of the objective and eyepiece are \( f_1 = +200 \) and \( f_2 = -25 \) units, the separation \( t = (200 - 25) = 175 \) units. The system matrix is:

\[
\begin{bmatrix}
1 & 0 \\
-\frac{1}{(-25)} & 1
\end{bmatrix}
\begin{bmatrix}
1 & 175 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
-\frac{1}{(+200)} & 1
\end{bmatrix}
= \begin{bmatrix}
\frac{1}{8} & 175 \\
0 & 8
\end{bmatrix}
\]

Note that the system power \( \varphi = 0 \implies f_{\text{eff}} = \infty \), which means that the system is “afocal”. The ray from an object at \( \infty \) with unit height generates the outgoing ray:

\[
\begin{bmatrix}
\frac{1}{8} & 175 \\
0 & 8
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix}
= \begin{bmatrix}
y' \\
n'u'
\end{bmatrix}
= \begin{bmatrix}
\frac{1}{8} \\
0
\end{bmatrix}
\]

so the outgoing ray is at height \( \frac{1}{8} \) and the angle is zero. Note that the diagonal elements are positive and the determinant is 1.

The “provisional” chief ray into the system has height 0 and angle 1; the outgoing ray is:

\[
\begin{bmatrix}
\frac{1}{8} & 175 \\
0 & 8
\end{bmatrix}
\begin{bmatrix}
0 \\
1
\end{bmatrix}
= \begin{bmatrix}
y \\
n'
\end{bmatrix}
= \begin{bmatrix}
175 \\
8
\end{bmatrix}
\]

So the outgoing ray angle is 8 times larger.
Keplerian Telescope made of Thin Lenses

The Keplerian telescope with \( f_1 = +200 \) and \( f_2 = +25 \) units with separation \( t = (200 + 25) = 225 \) units. The system matrix is:

\[
\begin{bmatrix}
  1 & 0 & 1 & 225 \\
  -\frac{1}{(25)} & 1 & -\frac{1}{(200)} & 1
\end{bmatrix}
\begin{bmatrix}
  1 & 0 \\
  0 & 1
\end{bmatrix}
\begin{bmatrix}
  y' \\
  0
\end{bmatrix}
= \begin{bmatrix}
  -\frac{1}{8} & 225 \\
  0 & -8
\end{bmatrix}
\begin{bmatrix}
  y' \\
  0
\end{bmatrix}
\]

The diagonal elements are negative, the determinant is 1, and the system power \( \varphi = 0 \implies f_{eff} = \infty \). The ray from an object at \( \infty \) with unit height generates the outgoing ray:

\[
\begin{bmatrix}
  -\frac{1}{8} & 225 \\
  0 & -8
\end{bmatrix}
\begin{bmatrix}
  1 \\
  0
\end{bmatrix}
= \begin{bmatrix}
  y' \\
  n' u'
\end{bmatrix}
= \begin{bmatrix}
  -\frac{1}{8} \\
  0
\end{bmatrix}
\]

so the outgoing ray is at height \(-\frac{1}{8}\) – the image is “inverted” and the angle is zero.

The “provisional” chief ray into the system has height 0 and angle 1; the outgoing ray is:

\[
\begin{bmatrix}
  -\frac{1}{8} & 225 \\
  0 & -8
\end{bmatrix}
\begin{bmatrix}
  0 \\
  1
\end{bmatrix}
= \begin{bmatrix}
  y' \\
  n' u'
\end{bmatrix}
= \begin{bmatrix}
  225 \\
  -8
\end{bmatrix}
\]

So the outgoing ray angle is 8 times larger than the incoming ray but negative.

Thick Lens

Consider the matrix for a thick lens made of glass with index \( n' \) in a medium with index \( n \)

\[
\varphi_1 = \frac{n' - n}{R_1} \\
\varphi_2 = \frac{n - n'}{R_2}
\]

The system matrix of the thick lens is:

\[
\mathcal{M}_{VV'} = \begin{bmatrix}
  1 & 0 & 1 & \frac{n'}{n'} \\
  -\varphi_2 & 1 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  1 & 0 \\
  -\varphi_1 & 1
\end{bmatrix}
\begin{bmatrix}
  1 & 0 \\
  0 & 1
\end{bmatrix}
\begin{bmatrix}
  y' \\
  n' u'
\end{bmatrix}
= \begin{bmatrix}
  1 - \varphi_1 \frac{n'}{n'} & \frac{1}{n'} u' \\
  -\left(\varphi_1 + \varphi_2 - \varphi_1 \varphi_2 \frac{n'}{n'}\right) & 1 - \varphi_2 \frac{n'}{n'}
\end{bmatrix}
\begin{bmatrix}
  y' \\
  n' u'
\end{bmatrix}
\]
We can immediately identify the power of the thick lens, which may be written in the form of the effective focal length:

\[
\phi = \phi_1 + \frac{\phi_2 \cdot t'}{n'}
\]

\[
f_{\text{eff}} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{t'}{n' f_1 f_2}
\]

Note that these equations apply to a thick lens made of glass with index \( n' \) used in air!

Consider an example made of glass with \( n' = 1.5 \) with \( R_1 = +50 \text{ mm} \) and \( R_2 = -100 \text{ mm} \). The thickness of the lens is 10 mm. The powers of the surfaces are:

\[
\phi_1 = \frac{n' - n}{R_1} = \frac{1.5 - 1}{50 \text{ mm}} = +\frac{1}{50 \text{ mm}}
\]

\[
\phi_2 = \frac{n - n'}{R_2} = \frac{1 - 1.5}{-100 \text{ mm}} = +\frac{1}{200 \text{ mm}}
\]

The system matrix is:

\[
\mathcal{M}_{VV'} = \begin{bmatrix}
1 & 0 & \frac{10 \text{ mm}}{1.5} & 1 & 0 \\
-\frac{1}{200 \text{ mm}} & 1 & 0 & 1 & -\frac{2}{100 \text{ mm}} & 1 \\
0.867 & 6.67 \text{ mm} \\
-\frac{1}{41.096 \text{ mm}} & 0.967
\end{bmatrix}
\]

The determinant is 1, as required. The refractive indices of object and image space are assumed to be 1. Substitute into the table of properties to find:

\[
f_{\text{eff}} = -\frac{1}{C} \cong +41.096 \text{ mm}
\]

\[
FFD = -\frac{D}{C} \cong 0.967 \cdot 41.096 \text{ mm} = +39.745 \text{ mm}
\]

\[
BFD = -\frac{A}{C} \cong 0.867 \cdot 41.096 \text{ mm} = +35.635 \text{ mm}
\]

\[
\Psi H = \frac{D - 1}{C} \cong (0.967 - 1) \cdot (-41.096 \text{ mm}) = +1.356 \text{ mm}
\]

\[
\Psi V = \frac{A - 1}{C} \cong (0.867 - 1) \cdot (-41.096 \text{ mm}) = +5.466 \text{ mm}
\]
Chapter 11

Waves and Imaging

We now return to those thrilling days of waves to consider their effects on the performance of imaging systems. We first consider “interference” of two traveling waves that oscillate with the same frequency and then generalize that to the interference of many such waves, which is called “diffraction”.

11.1 Interference of Waves

References: Hecht, Optics §8

Recall the identity that was derived for the sum of two oscillations with different frequencies $\omega_1$ and $\omega_2$:

\[
y_1[t] = A \cos[\omega_1 t] \\
y_2[t] = A \cos[\omega_2 t] \\
y_1[t] + y_2[t] = 2A \cos \left( \frac{\omega_1 + \omega_2}{2} t \right) \cdot \cos \left( \frac{\omega_1 - \omega_2}{2} t \right) \\
\equiv 2A \cos [\omega_{avg} t] \cdot \cos [\omega_{mod} t]
\]

In words, the sum of two oscillations of different frequency is identical to the product of two oscillations: one is the slower varying modulation (at frequency $\omega_{mod}$) and the other is the more rapidly oscillating average sinusoid (or carrier wave) with frequency $\omega_{avg}$. A perhaps familiar example of the modulation results from the excitation of two piano strings that are mistuned. A low-frequency oscillation (the beat) is heard; as one string is tuned to the other, the frequency of the beat decreases, reaching zero when the string frequencies are equal. Acoustic beats may be thought of as interference of the summed oscillations in time.

We also could consider this relationship in a broader sense. If the sinusoids are considered to be functions of the independent variable (coordinate) $t$, the phase angles of the two component functions $\Phi_1(t) = \omega_1 t$ and $\Phi_2(t) = \omega_2 t$ are different at the same coordinate $t$. The components sometimes add (for $t$ such that $\Phi_1[t] \equiv \Phi_2[t] \pm 2n\pi$) and sometimes subtract (if $\Phi_1(t) \equiv \Phi_2(t) \pm (2n + 1)\pi$).
We also derived the analogous effect for two waves traveling along the z-axis:

\[ f_1[z,t] = A \cos[k_1 z - \omega_1 t] \]
\[ f_2[z,t] = A \cos[k_2 z - \omega_2 t] \]
\[ f_1[z,t] + f_2[z,t] = \{2A \cos[k_{mod} z - \omega_{mod} t]\} \cdot \cos[k_{avg} z - \omega_{avg} t] \]

\[ k_{mod} = \frac{k_1 - k_2}{2} \]
\[ \omega_{mod} = \frac{\omega_1 - \omega_2}{2} \]
\[ v_{mod} = \frac{\omega_{mod}}{k_{mod}} = \frac{\omega_1 - \omega_2}{k_1 - k_2} \]
\[ k_{avg} = \frac{k_1 + k_2}{2} \]
\[ \omega_{avg} = \frac{\omega_1 + \omega_2}{2} \]
\[ v_{avg} = \frac{\omega_{avg}}{k_{avg}} = \frac{\omega_1 + \omega_2}{k_1 + k_2} \]

In words, the superposition of two traveling waves with different temporal frequencies (and thus different wavelengths) generates the product of two component traveling waves, one oscillating more slowly in both time and space, i.e., a traveling modulation. Note that both the average and modulation waves move along the z-axis. In this case, \( k_1, k_2, \omega_1, \) and \( \omega_2 \) are all positive, and so \( k_{avg} \) and \( \omega_{avg} \) must be also. However, the modulation wavenumber and frequency may be negative. In fact, the algebraic sign of \( k_{mod} \) may be negative even if \( \omega_{mod} \) is positive. In this case, the modulation wave moves in the opposite direction to the average wave.

Note that if the two 1-D waves traveling in the same direction along the z-axis have the same frequency \( \omega \), they must have the same wavelength \( \lambda_0 \) and the same wavenumber \( k = \frac{2\pi}{\lambda_0} \). The modulation terms \( k_{mod} \) and \( \omega_{mod} \) must be zero, and the summation wave exhibits no modulation. Recall also such waves traveling in opposite directions generate a waveform that moves but does not travel, but is a standing wave:

\[ f_1[z,t] = A \cos[k_1 z - \omega_1 t] \]
\[ f_2[z,t] = A \cos[k_1 z + \omega_1 t] \]
\[ f_1[z,t] + f_2[z,t] = \{2A \cos[k_{mod} z - \omega_{mod} t]\} \cdot \cos[k_{avg} z - \omega_{avg} t] \]
\begin{align*}
    k_{\text{mod}} &= \frac{k_1 - k_1}{2} = 0 \\
    \omega_{\text{mod}} &= \frac{\omega_1 - (-\omega_1)}{2} = \omega_1 \\
    k_{\text{avg}} &= \frac{k_1 + k_1}{2} = k_1 \\
    \omega_{\text{avg}} &= \frac{\omega_1 + (-\omega_1)}{2} = 0
\end{align*}

\begin{align*}
    f_1 [z, t] + f_2 [z, t] &= 2A \cos [k_1 z] \cos [-\omega_1 t] \\
                         &= 2A \cos [k_1 z] \cos [\omega_1 t]
\end{align*}

where the symmetry of \( \cos[\theta] \) was used in the last step.

Traveling waves also may be defined over two or three spatial dimensions; the waves have the form \( f[x, y, t] \) and \( f[x, y, z, t] \), respectively. The direction of propagation of such a wave in a multidimensional space is determined by a vector analogous to \( k \); a 3-D wavevector \( \mathbf{k} \) has components \( [k_x, k_y, k_z] \). The vector may be written:

\[ \mathbf{k} = [k_x \hat{x} + k_y \hat{y} + k_z \hat{z}] \]

The corresponding wave travels in the direction of the wavevector \( \mathbf{k} \) and has wavelength \( \lambda_0 = \frac{2\pi}{k} \). In other words, the length of \( \mathbf{k} \) is the magnitude of the wavevector:

\[ |\mathbf{k}| = \sqrt{k_x^2 + k_y^2 + k_z^2} = \frac{2\pi}{\lambda_0}. \]

The temporal oscillation frequency \( \omega \) is determined from the magnitude of the wavevector through the dispersion relation:

\[ \omega = v_\phi \cdot |\mathbf{k}| \rightarrow \nu = \frac{v_\phi}{\lambda_0} \]

For illustration, consider a simple 2-D analogue of the 1-D traveling plane wave. The wave travels in the direction of the 2-D wavevector \( \mathbf{k} \) which is in the \( x - z \) plane:

\[ \mathbf{k} = [k_x, 0, k_z] \]

The points of constant phase with with phase angle \( \phi = C \) radians is the set of points in the 2-D space \( \mathbf{r} = [x = 0, y, z] = (r, \theta) \) such that the scalar product \( \mathbf{k} \cdot \mathbf{r} = C \):
\[ \mathbf{k} \cdot \mathbf{r} = \mathbf{r} \cdot \mathbf{k} = |\mathbf{k}| \mathbf{r} \cos \theta = k_x x + k_z z = C \text{ for a point of constant phase} \]

Therefore, the equation of a 2-D wave traveling in the direction of \( \mathbf{k} \) with linear wavefronts is:

\[
\begin{align*}
    f[x, y, t] &= A \cos [k_x x + k_z z - \omega t] \\
    &= A \cos [\mathbf{k} \cdot \mathbf{r} - \omega t]
\end{align*}
\]

In three dimensions, the set of points with the same phase lie on a planar surface so that the equation of the traveling wave is:

\[
\begin{align*}
    f[x, y, z, t] &= f[r, t] \\
    &= A \cos [k_x x + k_y y + k_z z - \omega t] \\
    &= A \cos [\mathbf{k} \cdot \mathbf{r} - \omega t]
\end{align*}
\]

This plane wave could have been created by a point source at a large distance to the left and below the \( z \)-axis.

Now, we will apply the equation derived when adding oscillations with different
temporal frequencies. In general, the form of the sum of two traveling waves is:

\[ f_1(x, y, z, t) + f_2(x, y, z, t) = A \cos(k_1 \cdot r - \omega t) + A \cos(k_2 \cdot r - \omega t) \]

\[ = 2A \cos(k_{avg} \cdot r - \omega_{avg} t) \cdot \cos(k_{mod} \cdot r - \omega_{mod} t) \]

where the average and modulation wavevectors are:

\[ k_{avg} = \frac{k_1 + k_2}{2} = \frac{(k_x)_1 + (k_x)_2}{2} \hat{x} + \frac{(k_y)_1 + (k_y)_2}{2} \hat{y} + \frac{(k_z)_1 + (k_z)_2}{2} \hat{z} \]

\[ k_{mod} = \frac{k_1 - k_2}{2} = \frac{(k_x)_1 - (k_x)_2}{2} \hat{x} + \frac{(k_y)_1 - (k_y)_2}{2} \hat{y} + \frac{(k_z)_1 - (k_z)_2}{2} \hat{z} \]

and the average and modulation angular temporal frequencies are:

\[ \omega_{avg} = \frac{\omega_1 + \omega_2}{2} \]

\[ \omega_{mod} = \frac{\omega_1 - \omega_2}{2} \]

Note that the average and modulation wavevectors \( k_{avg} \) and \( k_{mod} \) point in different directions, in general, and thus the corresponding waves move in different directions at velocities determined from:

\[ v_{avg} = \frac{\omega_{avg}}{|k_{avg}|} \]

\[ v_{mod} = \frac{\omega_{mod}}{|k_{mod}|} \]

Because the phase of the multidimensional traveling wave is a function of two parameters (the wavevector \( k \) and the angular temporal frequency \( \omega \)), the phases of two traveling waves usually differ even if the temporal frequencies are equal. Consider the superposition of two such waves:

\[ \omega_1 = \omega_2 \equiv \omega \]

The component waves travel in different directions so the components of the wavevectors differ:

\[ k_1 = [(k_x)_1, (k_y)_1, (k_z)_1] \neq k_2 = [(k_x)_2, (k_y)_2, (k_z)_2] \]

Since the temporal frequencies are equal, so must be the wavelengths:

\[ \lambda_1 = \lambda_2 = \lambda \rightarrow |k_1| = |k_2| \equiv |k| \]

The condition of equal \( \omega \) ensures that the temporal average and modulation frequen-
cies are:

\[ \omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} = \omega_0 \]
\[ \omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2} = 0 \]

The summation of the two traveling waves with identical magnitudes may be expressed as:

\[
\begin{align*}
  f_1[x, y, z, t] + f_2[x, y, z, t] &= A \cos(k_1 \cdot r - \omega_0 t) + A \cos(k_2 \cdot r - \omega_0 t) \\
  &= 2A \cos(k_{\text{avg}} \cdot r - \omega_{\text{avg}} t) \cdot \cos(k_{\text{mod}} \cdot r - 0 \cdot t) \\
  &= 2A \cos(k_{\text{avg}} \cdot r - \omega_{\text{avg}} t) \cdot \cos(k_{\text{mod}} \cdot r)
\end{align*}
\]

Therefore, the superposition of two 2-D wavefronts with the same temporal frequency but traveling in different directions results in two multiplicative components: a traveling wave in the direction of \( k_{\text{avg}} \), and a wave in space along the direction of \( k_{\text{mod}} \) that does not move. This second stationary wave is analogous to the phenomenon of beats, and is called interference in optics.

### 11.1.1 Superposition of Two Plane Waves of the Same Frequency

Consider the superposition of two plane waves:

\[
\begin{align*}
  f_1[x, y, z, t] &= A \cos(k_1 \cdot r - \omega_0 t) \\
  f_2[x, y, z, t] &= A \cos(k_2 \cdot r - \omega_0 t) \\
  k_1 &= [k_x, k_y = 0, k_z] \\
  k_2 &= [-k_x, 0, k_z]
\end{align*}
\]

i.e., the wavevectors differ only in the \( x \)-component, and there only by a sign. Therefore the two wavevectors have the same “length”:

\[
|k_1| = |k_2| = \frac{2\pi}{\lambda} \quad \implies \lambda_1 = \lambda_2 \equiv \lambda.
\]

Also note that:

\[
\begin{align*}
  k_z &= |k| \cos[\theta] = \frac{2\pi}{\lambda} \cos[\theta] \\
  k_x &= \frac{2\pi}{\lambda} \sin[\theta]
\end{align*}
\]
11.1 INTERFERENCE OF WAVES

\[ \mathbf{k}_{\text{avg}} = \frac{\mathbf{k}_1 + \mathbf{k}_2}{2} = \frac{[k_x, 0, k_z] + [k_x, 0, k_z]}{2} = [0, 0, k_z] = \frac{2\pi}{\lambda} \cos[\theta] \]

\[ \mathbf{k}_{\text{avg}} = \frac{\mathbf{k}_1 - \mathbf{k}_2}{2} = \frac{[k_x, 0, k_z] + [-k_x, 0, k_z]}{2} = [k_x, 0, 0] = \frac{2\pi}{\lambda} \sin[\theta] \]

\[ \omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} = \omega_0 \]

\[ \omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2} = 0 \]

The wavevectors of two interfering plane waves with the same wavelength.

These two waves could have been generated at point sources located above and below the z-axis a large distance to the left. This is the classic “Young’s double slit” experiment, where light from a single source is split into to waves (spherical waves in this case) and propagate a large distance to the observation plane:
How two “tilted” plane waves are generated in the Young double-aperture experiment. The two apertures in the opaque screen on the left divide the incoming wave into two expanding spherical waves. After propagating a long distance, the spherical waves approximate plane waves that are tilted relative to the axis by \( \theta = \frac{d}{2L} \).

The “tilts” of the two waves are evaluated from the two distances:

\[
\theta \approx \frac{d/2}{L} = \frac{d}{2L}
\]

If \( L \gg d \), then

\[
\theta \approx \tan[\theta] \approx \sin[\theta] \approx \frac{d}{2L}
\]

The superposition of the two electric fields is:

\[
f[x, y, z, t] = f_1[x, y, z, t] + f_2[x, y, z, t] \\
= 2A \cos\left[\mathbf{k}_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}}t\right] \cdot \cos\left[\mathbf{k}_{\text{mod}} \cdot \mathbf{r}\right] \\
= 2A \cos\left[2\pi \frac{z}{\lambda} \cos[\theta] - \omega_0 t\right] \cos\left[2\pi \frac{x}{\lambda} \sin[\theta]\right]
\]

The first term (with the time dependence) is a traveling wave in the direction defined by \( \mathbf{k} = [0, 0, k_z] \), while the second term (with no dependence on time) is a spatial...
11.1 INTERFERENCE OF WAVES

wave along the $y$ direction. The amplitude variation in the $y$ direction is:

$$2A \cos \left[ 2\pi \frac{y}{\lambda} \sin \theta \right] = 2A \cos \left[ 2\pi \frac{x}{\sin \theta} \right]$$

which has a period of $\frac{\lambda}{\sin \theta}$. The irradiance (the measurable intensity) of the superposition is:

$$|f(x, y, z, t)|^2 = 4A^2 \cos^2 \left[ 2\pi \frac{z}{\lambda} \cos \theta - \omega_0 t \right] \cos^2 \left[ 2\pi \frac{x \sin \theta}{\lambda} \right]$$

The second cosine terms can be rewritten using:

$$\cos^2 \theta = \frac{1}{2} (1 + \cos 2\theta)$$

As before, the first term varies rapidly due to the angular frequency term $\omega_0 \approx 10^{14}$ Hz. Therefore, just the average value is detected:

$$\left\langle |f(x, y, z, t)|^2 \right\rangle = 4A^2 \cos^2 \left[ 2\pi \frac{x \sin \theta}{\lambda} \right] \cdot \frac{1}{2}$$

$$= 2A^2 \left( \frac{1}{2} \left( 1 + \cos \left[ \frac{4\pi x \sin \theta}{\lambda} \right] \right) \right)$$

$$= A^2 \left( 1 + \cos \left[ 2\pi \frac{x}{\sin \theta} \right] \right)$$

This derivation may also be applied to find the irradiance of one of the individual component waves:

$I_1 = \left\langle |f_1(x, y, z, t)|^2 \right\rangle$

$I_2 = \left\langle |f_2(x, y, z, t)|^2 \right\rangle$

$I_0 = \left\langle |f_1(x, y, z, t)|^2 \right\rangle = \left\langle |A \cos (k_x \cdot \vec{r} - \omega_0 t)|^2 \right\rangle = A^2 \left\langle \cos^2 (k_x \cdot \vec{r} - \omega_0 t) \right\rangle$

$$= A^2 \cdot \frac{1}{2}$$
So the irradiance of the sum of the two waves can be rewritten in terms of the irradiance of a single wave:

\[
\langle |f(x, y, z, t)|^2 \rangle = 4I_0 \cos^2 \left[ \frac{2\pi x \sin \theta}{\lambda} \right] \\
= 2I_0 \left( 1 + \cos \left[ \frac{2\pi x \cdot 2 \cdot \sin \theta}{\lambda} \right] \right) \\
= 2I_0 \left( 1 + \cos \left[ 2\pi \frac{x}{\lambda} \cdot \frac{1}{2 \sin \theta} \right] \right)
\]

The irradiance exhibits a sinusoidal modulation of period \( X = \frac{\lambda}{2 \sin \theta} \) and its irradiance oscillates between 0 and \( 2I_0 \cdot (2) = 4I_0 \), so that the average irradiance is \( 2I_0 \). The period varies directly with \( \lambda \) and inversely with \( \sin(\theta) \); for small \( \theta \), the period of the sinusoid is large, for \( \theta = 0 \), there is no modulation of the irradiance. The alternating bright and dark regions of this time-stationary sinusoidal intensity pattern often are called interference fringes. The shape, separation, and orientation of the interference fringes are determined by the incident wavefronts, and thus provide information about them. The argument of the cosine function is the optical phase difference of the two waves. At locations where the optical phase difference is an even multiple of \( \pi \), the cosine evaluates to unity and a maximum of the interference pattern results. This is an example of constructive interference. If the optical phase difference is an odd multiple of \( \pi \), the cosine evaluates to -1 and the irradiance is zero; this is destructive interference.

\[\text{Interference of two "tilted plane waves" with the same wavelength. The two component traveling waves are shown as "snapshots" at one instant of time on the}\]
left (white = 1, black = -1); the sum of the two is shown in the center (white = 2, black = -2), and the squared magnitude on the right (white = 4, black = 0). The modulation in the vertical direction is constant, while that in the horizontal direction is a traveling wave and “averages” out to a constant value of $\frac{1}{2}$.

The amplitude and irradiance observed at one instant of time when the irradiance at the origin ("on axis") is a maximum is shown:

\[ \begin{align*}
\text{(a)} & \quad \text{Interference patterns observed along the } x\text{-axis at one value of } z: \quad \text{(a)} \text{ amplitude fringes, with period equal to } \frac{\lambda}{\sin[\theta]}, \quad \text{irradiance (intensity) fringes, with period equal to } \frac{\lambda}{2\sin[\theta]} . \quad \text{This pattern is averaged over time and scales by a factor of } \frac{1}{2}. \\
\text{(b)} & \quad \text{Again, the traveling wave in the images of the amplitude and intensity of the superposed images moves in the } z\text{-direction (to the right), thus blurring out the oscillations in the } z\text{-direction. The oscillations in the } x\text{-direction are preserved as the interference pattern, which is plotted as a function of } x \text{ below. Note that the spatial frequency of the intensity fringes is twice as large as that of the amplitude fringes.}
\end{align*} \]

\[ \begin{align*}
\text{(a)} & \quad \text{Irradiance patterns observed at the output plane at several instants of time, showing that the spatial variation of the irradiance is preserved but the averaging reduces the maximum value by half.}
\end{align*} \]
11.1.2 Superposition of Two Plane Waves with Different Frequencies

For further illustration, consider the case the two waves travel in the same directions, so that \( \mathbf{k}_1 \neq \mathbf{k}_2 \), but with different temporal frequencies \( \omega_1 \neq \omega_2 \). This means that \( |\mathbf{k}_1| \neq |\mathbf{k}_2| \). The average and modulation wavevectors are found as before, but the modulation wave now travels because both \( \mathbf{k}_{\text{mod}} \neq \mathbf{0} \) and \( \omega_{\text{mod}} \neq 0 \). Consider the example of two component waves: \( f_1[\mathbf{r},t] \) directed at an angle \( \theta_1 = +40^\circ \approx \frac{2}{3} \) radian with \( \lambda_1 = 8 \) units and \( \omega_1 = \frac{1}{8} \) radians/second, and \( f_2[\mathbf{r},t] \) directed at \( \theta_2 = -40^\circ \approx -\frac{2}{3} \) radian with \( \lambda_2 = 12 \) units and \( \omega_2 = \frac{1}{12} \) radians/second. The corresponding average and modulation frequencies are:

\[
\omega_{\text{avg}} = \frac{\omega_1 + \omega_2}{2} = \frac{2\pi}{2} \left( \frac{1}{8} + \frac{1}{12} \right) = \frac{2\pi}{9.6} = \frac{5\pi}{24} \text{ radians/second}
\]

\[
\omega_{\text{mod}} = \frac{\omega_1 - \omega_2}{2} = \frac{2\pi}{2} \left( \frac{1}{8} - \frac{1}{12} \right) = \frac{2\pi}{48} \text{ radians/second}
\]

and the average and modulation wavevectors are:

\[
\mathbf{k}_{\text{avg}} = \frac{\mathbf{k}_1 + \mathbf{k}_2}{2} = 2\pi \cdot \left( \frac{x}{48} \sin [40^\circ] + \frac{5z}{48} \cos [40^\circ] \right) = \frac{2\pi}{48} \left( \frac{x}{74.674} + \frac{z}{12.532} \right)
\]

\[
\mathbf{k}_{\text{mod}} = \frac{\mathbf{k}_1 - \mathbf{k}_2}{2} = 2\pi \cdot \left( 5x \sin [40^\circ] + z \cos [40^\circ] \right) \approx 2\pi \left( \frac{x}{14.935} + \frac{z}{62.0} \right)
\]

The superposition may be written as the product of the average and modulation waves:

\[
f_1[\mathbf{r},t] + f_2[\mathbf{r},t] = 2f_{\text{avg}}[\mathbf{r},t] \cdot f_{\text{mod}}[\mathbf{r},t]
\]

where the full expressions for the average and modulation waves are:

\[
f_{\text{avg}}[\mathbf{r},t] = \cos \left[ \mathbf{k}_{\text{avg}} \cdot \mathbf{r} - \omega_{\text{avg}}t \right]
\]

\[
= \cos \left[ \frac{2\pi}{48} \left( x \sin [40^\circ] + 5z \cos [40^\circ] \right) - \frac{10\pi t}{48} \right]
\]

\[
= \cos \left[ \frac{2\pi}{48} \left( y \sin [40^\circ] + 5z \cos [40^\circ] \right) - 5t \right]
\]

\[
\approx \cos \left[ 2\pi \left( \frac{x}{74.674} + \frac{z}{12.532} \right) - 2\pi \left( \frac{t}{9.6} \right) \right]
\]
11.1 INTERFERENCE OF WAVES

\[ f_{\text{mod}}(\mathbf{x}, t) = \cos(k_{\text{mod}} \cdot \mathbf{x} - \omega_{\text{mod}} t) \]

\[ = \cos \left[ \frac{2\pi}{48} \cdot (5x \sin [40^\circ] + z \cos [40^\circ]) - \frac{2\pi t}{48} \right] \]

\[ = \cos \left[ \frac{2\pi}{48} (5x \sin [40^\circ] + z \cos [40^\circ] - 2t) \right] \]

\[ \cong \cos \left[ 2\pi \left( \frac{x}{74.674} + \frac{z}{12.532} \right) - 2\pi \left( \frac{t}{24} \right) \right] \]

Note that both the average and modulation waves are traveling waves; they are headed in different directions with different frequencies and different velocities. The temporal frequencies are \( \nu_{\text{avg}} = \frac{5}{48} \text{ Hz} \) and \( \nu_{\text{mod}} = \frac{2}{48} \text{ Hz} \). If the intensity (squared-magnitude) of the sum is averaged over time at an observation plane located downstream on the \( z \)-axis, both traveling waves will average out and no stationary fringe pattern will be visible.

\[ |k_1| = 2\pi \lambda_1 \propto 1/12 \]

\[ |k_2| = 2\pi \lambda_2 \propto 1/8 \]

**Sum of two sinusoidal traveling waves where the periods are related by** \( \lambda_2 = \frac{2}{3} \lambda_1 \).

The two waves travel in the directions \( \pm 40^\circ \), respectively. The resulting amplitude sum and power are depicted as “snapshots” at one instant of time. Since the modulation wave now travels too, both waves are averaged to constant values and no fringes are visible.
Intensity patterns observed at the output plane at several instants of time. The velocity of the modulation wave makes this pattern “migrate” towards $-x$, and thus the time-averaged pattern is a constant; no interference is seen.

The same principles just discussed may be used to determine the form of interference fringes from wavefronts with other shapes. Some examples will be considered in the following sections.

### 11.1.3 Fringe Visibility – Coherence

The visibility of a sinusoidal fringe pattern is a quality that corresponds quite closely to modulation, which is a term used by electrical engineers (sparkies). Given a nonnegative sinusoidal irradiance (intensity) distribution with maximum $I_{\text{max}}$ and minimum $I_{\text{min}}$ (so that $I_{\text{min}} \geq 0$), the visibility of the sinusoidal fringe pattern is:

$$V \equiv \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}$$

Note that if $I_{\text{min}} = 0$, then $V = 1$ regardless of the value of $I_{\text{max}}$. The visibility of the fringe pattern is largely determined by the relative irradiances of the individual wavefronts and by the coherence of the light source.

To introduce the concept of coherence, consider first the Young’s two-aperture experiment where the source is composed of equal-amplitude emission at two distinct wavelengths $\lambda_1$ and $\lambda_2$ incident on the observation screen at $\pm \theta$. Possible pairs of wavelengths could be those of the sodium doublet ($\lambda = 589.0\text{ nm}$ and $589.6\text{ nm}$), or the pair of lines emitted by a “greenie” He:Ne laser ($\lambda = 543\text{ nm}$ (green), $594\text{ nm}$ (yellow)). In air or vacuum, the corresponding angular frequencies obviously are $\omega_1 = \frac{2\pi c}{\lambda_1}$ and $\omega_2 = \frac{2\pi c}{\lambda_2}$.

To find the irradiance pattern created by the interference of the four beams, we must compute the superposition of the amplitude of the electromagnetic field, find its squared-magnitude, and then determine the average over time. The sum of the four component terms is straightforward to compute by recognizing that it is the sum of
the amplitude patterns from the pairs of waves with the same wavelength. We have already shown that the sum of the two terms with $\lambda = \lambda_1$ is:

$$f_1[x, z; \lambda_1] + f_2[x, z; \lambda_1] = 2A \cos \left[ \frac{2\pi x}{\lambda_1} \sin \theta \right] \cos \left[ \frac{2\pi z}{\lambda_1} \cos \theta - \omega_1 t \right]$$

$$= 2A \cos \left[ 2\pi \left( \frac{x}{\lambda_1} \right) \cos \left( \frac{2\pi z}{\lambda_1} \cos \theta - \omega_1 t \right) \right]$$

which is the sum of a stationary sinusoid and a traveling wave in the $+z$-direction. If we add a second pair of plane waves with different wavelength $\lambda = \lambda_2$ but the same “tilts,” the amplitude pattern can also be calculated. We have to add the amplitude of four waves, but we can still add them in pairs. The first pair produces the same amplitude pattern that we saw before. The second wave also produces a pattern that differs only in the periods of the sinusoids.

The sum of four tilted plane waves can be calculated by summing the pair due to one wavelength and that due to the other.

The second pair of wavefronts with $\lambda = \lambda_2$ yield a similar result, though the period of the stationary fringes and the temporal frequency of the traveling wave differ. The
expression for the sum of the two pairs is:

\[
\sum_{n=1}^{4} f_n [x, z; \lambda] = f_1 [x, z; \lambda_1] + f_2 [x, z; \lambda_1] + f_1 [x, z; \lambda_2] + f_2 [x, z; \lambda_2]
\]

\[
= 2A \cos \left[ 2\pi \frac{x}{(\lambda_1 \sin[\theta])} \right] \cos \left[ \left( \frac{2}{\lambda_1} \cos [\theta] \right) z - \omega_1 t \right] \\
+ 2A \cos \left[ 2\pi \frac{x}{(\lambda_2 \sin[\theta])} \right] \cos \left[ \left( \frac{2}{\lambda_2} \cos [\theta] \right) z - \omega_2 t \right]
\]

\[
= 2A \cos \left[ 2\pi \frac{x}{(\lambda_1 \sin[\theta])} \right] \cos \left[ (k_{avg})_1 z - \omega_1 t \right] + 2A \cos \left[ 2\pi \frac{x}{(\lambda_2 \sin[\theta])} \right] \cos \left[ (k_{avg})_2 z - \omega_2 t \right]
\]

where:

\[
\left| (k_{avg})_1 \right| = (k_{avg})_1 = \frac{2\pi}{\lambda_1 \cos [\theta]}
\]

\[
\left| (k_{avg})_2 \right| = (k_{avg})_2 = \frac{2\pi}{\lambda_2 \cos [\theta]}
\]

\[
(v_{avg})_1 = \omega_1 \frac{\nu_1}{(k_{avg})_1} = \frac{2\pi \nu_1}{(2\pi \frac{x}{\lambda_1 \cos[\theta]}} = \frac{\nu_1 \lambda_1}{z} \cos [\theta] = \frac{c}{z} \cos [\theta]
\]

\[
v_2 = \omega_2 \frac{\nu_2}{(k_{avg})_2} = \frac{2\pi \nu_2}{(2\pi \frac{x}{\lambda_2 \cos[\theta]}} = \frac{\nu_2 \lambda_2}{z} \cos [\theta] = \frac{c}{z} \cos [\theta] = v_1
\]

\[
\Rightarrow v_2 = v_1
\]

Thus the phase velocities of the two travelling waves are equal.

\[
2A \cos \left[ 2\pi \frac{x}{(\lambda_1 \sin[\theta])} \right] \cos \left[ (k_{avg})_1 z - \omega_1 t \right] + 2A \cos \left[ 2\pi \frac{x}{(\lambda_2 \sin[\theta])} \right] \cos \left[ (k_{avg})_2 z - \omega_2 t \right]
\]

\[
= 2A \cos \left[ 2\pi \frac{x}{(\lambda_1 \sin[\theta])} \right] \cos \left[ (k_{avg})_1 \left( z - \frac{\omega_1}{k_1} t \right) \right] + 2A \cos \left[ 2\pi \frac{x}{(\lambda_2 \sin[\theta])} \right] \cos \left[ (k_{avg})_2 \left( z - \frac{\omega_2}{k_2} t \right) \right]
\]

Because the phase velocities of the two “average” waves are equal, the traveling waves
11.1 INTERFERENCE OF WAVES

differ only in wavenumber and both travel down the z-axis with the same velocity:

\[ \sum_{n=1}^{4} f_n [x, z; \lambda] = 2A \cos \left[ 2\pi \left( \frac{x}{\lambda_1} \sin[\theta] \right) \right] \cos \left[ (k_{avg})_1 (z - v_1 t) \right] + 2A \cos \left[ 2\pi \left( \frac{x}{\lambda_2} \sin[\theta] \right) \right] \cos \left[ (k_{avg})_2 (z - v_2 t) \right] \]

Here we make a simplifying assumption that is not strictly correct: we assume that the two traveling wave parts contribute to the amplitude pattern in the same way and can be factored out:

\[ \sum_{n=1}^{4} f_n [x, z; \lambda] = 2A \cos \left[ \frac{2\pi z}{\lambda_1} \cos[\theta] - \omega_1 t \right] \left( \cos \left[ 2\pi \left( \frac{x}{\lambda_1} \sin[\theta] \right) \right] + \cos \left[ 2\pi \left( \frac{x}{\lambda_2} \sin[\theta] \right) \right] \right) \]

The squared magnitude of the amplitude is:

\[ \left| \sum_{n=1}^{4} f_n [x, z; \lambda] \right|^2 = 4A^2 \left( \cos \left[ 2\pi \left( \frac{x}{\lambda_1} \sin[\theta] \right) \right] + \cos \left[ 2\pi \left( \frac{x}{\lambda_2} \sin[\theta] \right) \right] \right)^2 \left( \cos \left[ \frac{2\pi z}{\lambda_1} \cos[\theta] - \omega_1 t \right] \right)^2 \]

and the time average yields the irradiance:

\[ I [x, z] = \left\langle \left| \sum_{n=1}^{4} f_n [x, z; \lambda] \right|^2 \right\rangle = 4A^2 \left( \cos \left[ 2\pi \left( \frac{x}{\lambda_1} \sin[\theta] \right) \right] + \cos \left[ 2\pi \left( \frac{x}{\lambda_2} \sin[\theta] \right) \right] \right)^2 \left( \cos \left[ \frac{2\pi z}{\lambda_1} \cos[\theta] - \omega_1 t \right] \right)^2 \]

\[ = 2A^2 \left( \cos \left[ 2\pi \left( \frac{x}{\lambda_1} \sin[\theta] \right) \right] + \cos \left[ 2\pi \left( \frac{x}{\lambda_2} \sin[\theta] \right) \right] \right)^2 \]

The sum of the two stationary cosine waves also may be recast as the product of cosines with the average and modulation frequencies:
patterns with identical maxima and zero minima:

\[ \lambda \]

In the case where the two emitted wavelengths are close together such that

\[ 2 \lambda \sim \lambda_{avg} \]

where the respective periods of the two component oscillations may be defined:

\[ \lambda_{avg} \equiv \frac{\lambda_1 + \lambda_2}{2} \quad \text{and} \quad \lambda_{mod} \equiv \frac{\lambda_1 - \lambda_2}{2}. \]

The final expression for the irradiance is the product of two sinusoidal irradiance patterns with identical maxima and zero minima:

\[
I(x, z) = 2A^2 \left[ \cos \left( \frac{2\pi}{\lambda_{avg}} \lambda_{avg} \right) \cdot \cos \left( \frac{2\pi}{\lambda_{mod}} \lambda_{mod} \right) \right]^2
\]

where \( D_{avg} \) and \( D_{mod} \) of the two spatial oscillations are defined:

\[
D_{avg} \equiv \frac{\lambda_1 \lambda_2}{\lambda_{avg} \sin[\theta]} = \frac{1}{\sin[\theta]} \left( \frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right)^{-1} \propto |\lambda_1 + \lambda_2|^{-1} \propto (\lambda_{avg})^{-1}
\]

\[
D_{mod} \equiv \frac{\lambda_1 \lambda_2}{\lambda_{mod} \sin[\theta]} = \frac{1}{\sin[\theta]} \left| \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right|^{-1} \propto |\lambda_1 - \lambda_2|^{-1} \equiv (\Delta \lambda)^{-1}
\]

In the case where the two emitted wavelengths are close together such that \( \lambda_1 \sim \lambda_2 \sim \lambda_{avg} \gg \lambda_{mod} \), the expressions for the periods of the two component oscillations may...
be simplified:

\[ D_{\text{avg}} \approx \frac{L\lambda_{\text{avg}}}{2d} \]
\[ D_{\text{mod}} \approx \frac{L(\lambda_{\text{avg}})^2}{2d \cdot \Delta \lambda} \]

After cancelling the common terms, the relative lengths of the spatial periods of the modulations are:

\[ D_{\text{mod}} = D_{\text{avg}} \cdot \frac{\lambda_{\text{avg}}}{\Delta \lambda} \gg D_{\text{avg}} \text{ if } \Delta \lambda \ll \lambda_{\text{avg}} \]

In words, the period of the modulation due to \( \lambda_{\text{mod}} \) is much longer than that due to \( \lambda_{\text{avg}} \) if the emitted wavelengths are approximately equal. The period \( D_{\text{mod}} \) limits the range of \( x \) over which the short-period fringes can be seen. In fact, the sinusoidal fringes due to \( \lambda_{\text{avg}} \) are visible over a range of \( x \) equal to range of \( x \) between the zeros of \( D_{\text{mod}} \), i.e., half the period of \( D_{\text{mod}} \). The pattern resulting from the example considered above is shown. Note that the amplitude of maxima of the the irradiance fringes decreases away from the center of the observation screen, where the optical path lengths are equal for all four beams, and thus where they will add constructively.

\[ \text{Interference patterns from two wavelengths with same input amplitude: (a) the two amplitude patterns differ in period in proportion to the wavelength; (b) the sum of the two amplitude patterns at one instant of time; (c) the squared magnitude at the same instant, showing that the amplitude of the fringe varies with } x. \]

In this case where \( \Delta \lambda \ll \lambda_{\text{avg}} \), the fringes are visible over a large interval of \( x \). We speak of such a light source as \textit{temporally coherent}; the phase difference of light emitted at \( \lambda_1 \) and at \( \lambda_2 \) changes slowly with time, and thus with the position along the \( x \)-axis. Therefore, fringes are visible over a large range of \( x \). On the other hand, if \( \Delta \lambda \) is of the same order as \( \lambda_{\text{avg}} \), the wavelengths are widely separated. The phase difference of light emitted at the two extreme wavelengths changes rapidly with time, and thus with position along the \( x \)-axis. The fringes are visible only where the phase difference remains approximately constant (for \( x \approx 0 \)) over the averaged time.
interval. Such a source is said to be temporally incoherent. It is difficult (though not impossible) to see fringes generated by an incoherent source.

11.1.4 Coherence Time and Coherence Length

If two wavelengths emitted by the source are separated by $\Delta \lambda$, the corresponding frequency difference often is called the bandwidth of the source:

$$\Delta \nu \equiv |\nu_1 - \nu_2| = \left| \frac{c}{\lambda_1} - \frac{c}{\lambda_2} \right| = c \cdot \left| \frac{\lambda_2 - \lambda_1}{\lambda_1 \cdot \lambda_2} \right| = c \cdot \frac{\Delta \lambda}{\lambda_1 \cdot \lambda_2}$$

If the source includes a third wavelength $\lambda_3$ midway between the extrema $\lambda_1$ and $\lambda_2$ (so that $\lambda_3 = \lambda_{avg}$), the factors $\Delta \lambda$ and $\lambda_{avg}$ are unchanged, but the irradiance pattern must be different in some way. The irradiance pattern generated by this three-line source is more difficult to calculate, but the result can be modeled easily by recognizing that the wavefronts generated by $\lambda_3$ through the two apertures combine in amplitude to create a third pattern of sinusoidal fringes with a spatial period between those due to $\lambda_1$ and $\lambda_2$. The three such patterns may be summed and squared to model the irradiance fringes. Consider first the individual fringe patterns due to the extrema $\lambda_1$ and $\lambda_2$ as shown in (a):

(a) Irradiance pattern resulting from two wavelengths with equal “powers”, showing the long-period fringes due to $\Delta \lambda$ and the short-period fringes due to $\lambda_{avg}$; (b) Irradiance pattern after a third wavelength is added at $\lambda_{avg}$ with the same “power”. The distance between peaks of the fringe pattern has increased; (c) if all wavelengths between the maximum and minimum are used, the amplitudes of the neighboring maxima decrease.

The irradiance pattern generated from the superposition of these fringe patterns exhibits the short-period fringes due to $\lambda_{avg}$ and the long-period fringes due to $\Delta \lambda$.

If we add a third fringe pattern due to $\lambda_3 = \lambda_{avg}$, the resulting irradiance fringe pattern is shown in (b). Note that the region of visible fringes covers approximately the same extent of the $x$-axis, but the distance between such regions has increased. By extension, if all wavelengths are included in the source between $\lambda_1$ and $\lambda_2$, visible fringes will exist only in one region centered about $x = 0$.
11.1 INTERFERENCE OF WAVES

Because the region of fringes created by the three-line source is similar in size to that from the two-line source, but (infinitely) much smaller than the region of interference from the single-line source, we say that light from the first two are equally coherent, but less coherent than light emitted by the single-line source. The coherence may be quantified based on the temporal bandwidth. For a source whose range emitted wavelengths is:

\[ \Delta \lambda = \lambda_{\text{max}} - \lambda_{\text{min}}, \]

the corresponding temporal bandwidth is:

\[ \Delta \nu = c \cdot \frac{\Delta \lambda}{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}. \]

Note that the dimensions of \( \Delta \nu \) are \((\text{time})^{-1}\). The time delay over which the phase difference of light emitted from one source point is predictable (and thus over which fringes may be generated) is the inverse of this bandwidth:

\[ \Delta \tau = \frac{1}{\Delta \nu} = \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{c \cdot \Delta \lambda}, \]

which is called the coherence time. Obviously, if \( \Delta \lambda \) is large, then so is \( \Delta \nu \) and the corresponding coherence time is small. The coherence length is the distance traveled by an electromagnetic wave during the coherence time:

\[ c \cdot \Delta \tau \equiv \Delta \ell = \frac{c}{\Delta \nu} = \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{\Delta \lambda} \]

and is a measure of the length of the electromagnetic wave packet over which the phase difference is predictable. Recall that for interference of waves from a source with wavelength range \( \Delta \lambda \), the range of coordinate \( x \) over which fringes are visible is half the period \( D_{\text{mod}} \):

\[ \frac{D_{\text{mod}}}{2} = \frac{1}{2 \cdot \sin \theta} \left| \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \right|^{-1} = \frac{1}{2 \cdot \sin \theta} \cdot \frac{\lambda_{\text{max}} \cdot \lambda_{\text{min}}}{\Delta \lambda} \]

Thus the range of \( x \) over which fringes are visible is proportional to the coherence length; if the bandwidth is “wide”, then the coherence length is “short” and the fringes are visible over only a small region.

Lasers are the best available coherent sources; to a very good approximation, most lasers emit a single wavelength \( \lambda_0 \) so that \( \Delta \lambda = 0 \). The period of the modulating sinusoid \( D_{\text{mod}} = \infty \); fringes are visible at all \( x \). A coherent source should be employed when an optical interference pattern is used to measure a parameter of the system, such as the optical image quality (as was used to test the Hubble space telescope). Thus the range of \( x \) over which fringes are visible is determined by the coherence length (and thus the bandwidth) of the source. Therefore, the visibility of the interference fringes may be used as a measure the source coherence.
11.1.5 Effect of Polarization of Electric Field on Fringe Visibility

Up to this point, we have ignored the effect on the orientation of the electric field vectors on the sum of the fields. In fact, this is an essential consideration; two orthogonal electric field vectors cannot add to generate a time-invariant modulation in the irradiance. Consider the sum of two electric field vectors \( \mathbf{E}_1 \) and \( \mathbf{E}_2 \) to generate a field \( \mathbf{E} \). The resulting irradiance is:

\[
I = \mathbf{E} \cdot \mathbf{E} = |\mathbf{E}|^2 = |\mathbf{E}_1 + \mathbf{E}_2|^2 \\
= (\mathbf{E}_1 + \mathbf{E}_2) \cdot (\mathbf{E}_1 + \mathbf{E}_2) \\
= (\mathbf{E}_1 \cdot \mathbf{E}_1) + (\mathbf{E}_2 \cdot \mathbf{E}_2) + (\mathbf{E}_1 \cdot \mathbf{E}_2) + (\mathbf{E}_2 \cdot \mathbf{E}_1) \\
= (I_1 + I_2) + 2(\mathbf{E}_1 \cdot \mathbf{E}_2)
\]

where \( I_1 \) and \( I_2 \) are the irradiances due to \( \mathbf{E}_1 \) and \( \mathbf{E}_2 \), respectively.

Consider the irradiance in the case where the incident fields are plane waves traveling in directions \( \hat{\mathbf{e}}_1 \) and \( \hat{\mathbf{e}}_2 \), respectively:

\[
\mathbf{E}_1 = \hat{\mathbf{e}}_1 E_1 \cos \left( \mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t \right) \\
\mathbf{E}_2 = \hat{\mathbf{e}}_2 E_2 \cos \left( \mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t \right)
\]

\[
I = I_1 + I_2 + 2 \left( \hat{\mathbf{e}}_1 E_1 \cos \left( \mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t \right) \cdot \hat{\mathbf{e}}_2 E_2 \cos \left( \mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t \right) \right)
\]

\[
= I_1 + I_2 + 2E_1 E_2 (\hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_2) \left( \cos \left( \mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t \right) \cos \left( \mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t \right) \right)
\]

\[
= I_1 + I_2 + 2E_1 E_2 (\hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_2) \left( \cos \left( \mathbf{k}_1 \cdot \mathbf{r} - \omega_1 t \right) \cos \left( \mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t \right) \right)
\]

In the case where the two components are polarized orthogonally so that \( \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_2 = 0 \), then the irradiance is the sum of the component irradiances and no interference is seen.

In the case where \( \omega_1 = \omega_2 \) so that \( \omega_{\text{mod}} = 0 \), and \( I_1 = I_2 \), then the output irradiance is:

\[
I = 2I_1 \left( 1 + \cos \left( 2 \mathbf{k}_{\text{mod}} \cdot \mathbf{r} \right) \right)
\]

which again says that the irradiance includes a stationary sinusoidal fringe pattern with spatial period \( \lambda_{\text{mod}} = \frac{2\pi}{|\mathbf{k}_{\text{mod}}|} \).

11.2 Interferometers

We have seen several times now that optical interference results when two (or more) waves are superposed in such a way to produce a time-stationary spatial modulation of the superposed electric field, which may be observed by eye or a photosensitive detector. Interferometers use this result to measure different parameters of the light (e.g., wavelength \( \lambda \), bandwidth \( \Delta \lambda \) or coherence length \( \Delta \ell \), angle and sphericity of the wavefront, etc.), or of the system (path length, traveling distance, index of refraction, etc.).
etc.). Interferometers are generally divided into two classes that specify the method of separating a single wavefront into two (or more) wavefronts that may be recombined. The classes are division of wavefront and division of amplitude. The former type has been considered; it divides wavefronts emitted by the source into two pieces and redirects one or both of them down different paths. They are recombined in a fashion such that \( k_1 \neq k_2 \), even though \( |k_1| = |k_2| \). The interference pattern is generated from the \( k \mod \) portion of the sum of the wavefronts. Division-of-amplitude interferometers use a partially reflecting mirror — the beamsplitter — to divide the wavefront into two beams which travel down different paths and are recombined by the original or another beamsplitter. The optical interference is generated by the phase difference between the recombined wavefronts.

### 11.2.1 Division-of-Amplitude Interferometers

This class of amplifiers are distinguished from the just-considered division-of-wavefront interferometers by the presence of a beamsplitter, which divides the incident radiation into two parts by partial reflection/partial transmission. The two wavefronts are directed down different paths before recombining to create interference. For example, consider the Michelson interferometer shown below.

![Michelson Interferometer Diagram](image)

Note that the beamsplitter reflects part of the wave and transmits the rest. From the definition of the amplitude reflection coefficient:

\[
    r = \frac{n_1 - n_2}{n_1 + n_2} \quad \text{at normal incidence}
\]

we see that the amplitude is multiplied by a negative number if \( n_1 < n_2 \), meaning that the phase is changed by \( \pi \) radians if reflected at a rare-to-dense interface (second surface has larger \( n \)). The reflection at a dense-to-rare interface exhibits no phase shift.

If the beamsplitter both reflects and transmits 50% of the irradiance (NOT 50% of the amplitude), then equal portions of the energy are directed toward the mirrors.
M_{1} \text{ and } M_{2}. \text{ The amplitude of the electric field of the reflected beam is:}

\[ E_{1} = \sqrt{I_{1}} = \sqrt{\frac{I_{0}}{2}} = \sqrt{\frac{E_{0}^2}{2}} \]

\[ = E_{0} \cdot \frac{1}{\sqrt{2}} = 0.707 \cdots E_{0} \]

and the amplitude of the transmitted beam \( E_{2} = E_{1} \).

Because each beam is reflected once and transmitted once before being recombined, the amplitude of each component when recombined is:

\[ (E_{1})_{\text{out}} = (E_{2})_{\text{out}} = E_{0} \cdot \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2}} = \frac{1}{2} E_{0} \]

Each beam experiences a phase delay proportional to the optical distance traveled in its arm of the interferometer:

\[ \Phi_{1} = \frac{2\pi}{\lambda_{1}} \cdot d_{1} = \frac{2\pi}{\lambda_{1}} \cdot n_{1}d_{1} = k \cdot n_{1}d_{1} \]

where \( d_{1} \) is the distance traveled by beam \#1 and \( n_{1} \) is the refractive index in that path (\( n = 1 \) in vacuum or air). The beam directed at mirror \( M_{1} \) travels distance \( L_{1} \) from the beamsplitter to the mirror and again on the return, so the total physical path length \( d_{1} = 2L_{1} \). Similarly the physical length of the second path is \( d_{2} = 2L_{2} \) and the optical path is \( n_{2}d_{2} = 2n_{2}L_{2} \). After recombination, the relative phase delay is:

\[ \Delta \Phi = \Phi_{1} - \Phi_{2} = k (n_{1}d_{1} - n_{2}d_{2}) \]

\[ = 2k (n_{1}L_{1} - n_{2}L_{2}) \]

\[ = \frac{4\pi}{\lambda_{1}} (n_{1}L_{1} - n_{2}L_{2}) \]

Note that the phase delay is proportional to \( \frac{1}{\lambda} \), i.e., longer wavelengths (red light) experience smaller phase delays than shorter wavelengths (blue light).

The amplitude at the detector is the sum of the amplitudes:

\[ E \left[ t \right] = \frac{E_{0}}{2} \cos \left[ \frac{2\pi}{\lambda_{1}} \cdot n_{1} \cdot 2L_{1} - \omega_{1}t \right] + \frac{E_{0}}{2} \cos \left[ \frac{2\pi}{\lambda} \cdot n_{2} \cdot 2L_{2} - \omega_{1}t \right] \]

\[ = \frac{E_{0}}{2} \left( \cos \left[ \frac{2\pi}{\lambda_{1}} \cdot n_{1} \cdot 2L_{1} - \omega_{1}t \right] + \cos \left[ \frac{2\pi}{\lambda_{1}} \cdot n_{2} \cdot 2L_{2} - \omega_{1}t \right] \right) \]
which has the form $\cos[A] + \cos[B]$, and thus may be rewritten as:

$$E(t) = \frac{2E_0}{2} \cos \left[ \frac{2\pi (2n_1L_1 + 2n_2L_2)}{2\lambda_1} - 2\pi\nu t \right] \cdot \cos \left[ \frac{2\pi (2n_1L_1 - 2n_2L_2)}{2\lambda_1} \right]$$

$$= E_0 \cos \left[ \frac{2\pi (n_1L_1 + n_2L_2)}{\lambda_1} - 2\pi\nu t \right] \cdot \cos \left[ \frac{2\pi (n_1L_1 - n_2L_2)}{\lambda_1} \right]$$

If the indices of refraction in the two paths are equal (usually $n_1 = n_2 \approx 1$), then the expression is simplified:

$$E(t) = E_0 \cos \left[ \frac{2\pi n (L_1 + L_2)}{\lambda_1} - 2\pi\nu t \right] \cdot \cos \left[ \frac{2\pi n (L_1 - L_2)}{\lambda_1} \right]$$

$$= E_0 \cos \left[ \frac{2\pi (L_1 + L_2)}{\lambda_1} - 2\pi\nu t \right] \cdot \cos \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right] \text{ for } n = 1$$

One of the multiplicative terms is a rapidly oscillating function of time; the other term is stationary in time. The time average of the squared magnitude is the irradiance:

$$I = \langle |E(t)|^2 \rangle = E_0^2 \left\langle \cos^2 \left[ \frac{2\pi (L_1 + L_2)}{\lambda_1} - 2\pi\nu t \right] \cdot \cos^2 \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right] \right\rangle$$

$$= E_0^2 \left\langle \cos^2 \left[ \frac{2\pi (L_1 + L_2)}{\lambda_1} - 2\pi\nu t \right] \right\rangle \cdot \cos^2 \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right]$$

$$= E_0^2 \cdot \frac{1}{2} \cos^2 \left[ \frac{2\pi (L_1 - L_2)}{\lambda_1} \right]$$

The identity $\cos^2[\theta] = \frac{1}{2} (1 + \cos[2\theta])$ may be used to recast the expression:

$$I = E_0^2 \cdot \frac{1}{2} \cdot \frac{1}{2} \left( 1 + \cos \left[ \frac{2\pi}{\lambda_1} \cdot 2(L_1 - L_2) \right] \right)$$

$$= E_0^2 \cdot \frac{1}{4} \left( 1 + \cos \left[ \frac{2\pi}{\lambda_1} \cdot 2(L_1 - L_2) \right] \right)$$

Note that this is not a function of time or position, but only of the lengths of the arms of the interferometer and of $\lambda$. The Michelson interferometer with monochromatic plane-wave inputs generates a uniform irradiance related to the path difference. If input wavefronts with other shapes are used, then fringes are generated whose period is a function of the local optical path difference, which is directly related to the shape of the wavefronts. If the incident light is a tilted plane wave, then the resulting pattern is analogous to that from a division-of-wavefront interferometer (Young’s double-slit). If a point source emitting spherical waves is used, then the interferometer may be modeled as:
Images of the point source are formed at $S_1$ (due to mirror $M_1$) and at $S_2$ (due to $M_2$). The wavefronts superpose and form interference fringes. The positions of the fringes may be determined from the optical path difference (OPD).

The OPD is the excess distance that one wavefront has to travel relative to the other before being recombined. For a ray oriented at angle $\theta$ measured relative to the axis of symmetry, the ray reflected from mirror $M_2$ travels an extra distance:

$$OPD = 2 (L_1 - L_2) \cdot \cos [\theta]$$

The symmetry about the central axis ensures that the fringes are circular. The phase difference of the waves is the extra number of radians of phase that the wave must
travel in that OPD:

\[ \Delta \Phi = k \cdot OPD = \frac{2\pi}{\lambda_1} \cdot 2[L_1 - L_2] \cdot \cos[\theta] = \frac{4\pi \cdot (L_1 - L_2) \cdot \cos[\theta]}{\lambda_1} \text{ [radians]} \]

If the phase difference is a multiple of \(2\pi\) radians, then the waves recombine in phase and a maximum of the amplitude results. If the phase difference is an odd multiple of \(\pi\) radians, then the waves recombine out of phase and a minimum of the irradiance results due to destructive interference. The locations of constructive interference (irradiance maxima) may be specified by:

\[ \Delta \Phi = 2\pi m = \frac{4\pi \cdot (L_1 - L_2) \cdot \cos[\theta]}{\lambda_1} \rightarrow m\lambda_1 = 2(L_1 - L_2) \cdot \cos[\theta] \]

The corresponding angles \(\theta\) are specified by:

\[ \theta = \cos^{-1}\left[\frac{m\lambda_1}{2 \cdot (L_1 - L_2)}\right] \]

As the physical path difference \(L_1 - L_2\) decreases, then \(\frac{m\lambda}{2(L_1 - L_2)}\) increases and \(\theta\) decreases. In other words, if the physical path difference is decreased, the angular size of a circular fringe decreases and the fringes disappear into the center of the pattern. Since a particular fringe occurs at the same angle relative to the optical axis, these are called fringes of equal inclination.

If one mirror is tilted relative to the other, then the output beams travel in different directions when recombined. The fringes thus obtained are straight and have a constant spacing just like those from a Young’s double-slit experiment.

### 11.2.2 Applications of the Michelson Interferometer

1. **Measure refractive index \(n\):** Insert a plane-parallel plate of known thickness \(t\) and unknown index \(n\) into one arm of a Michelson interferometer illuminated with light of wavelength \(\lambda\). Count the number of fringes due to the plate:

   \[ OPD = (n - 1) \cdot t = \Delta m \cdot \lambda_1 \]

2. **Measure the wavelength of an unknown source or \(\Delta \lambda\) between two spectral lines of a single source.** Count the number of fringes that pass a single point as one mirror is moved a known distance.

3. **Measure lengths of objects:** The standard \(m\) now is defined in terms of a par-
ticular spectral line:

\[ 1 \text{ m} = 1,650,763.73 \text{ wavelengths of emission at } \lambda = 605.8 \text{ nm} \]

from Kr-86 measured in vacuum

This standard has the advantage of being transportable.

4. Measure deflections of objects: The optical phase difference will be significant for very small physical path differences. Place one mirror on an object and count fringes to measure the deflection.

5. Measure the velocity of light (Michelson-Morley experiment)

11.2.3 Other Types of Division-of-Amplitude Interferometers

(2) Mach-Zehnder

The M-Z interferometer is very similar to the Michelson, except that a second beamsplitter is used to recombine the beams so that the light does not traverse the same path twice. Therefore there is no factor of 2 in the OPD for the M-Z. Mach-Zehnder interferometers often are used to measure the refractive index of liquids or gases. The container \( C_1 \) (or \( C_2 \)) is filled with a gas while examining the fringe pattern. As the container fills, the refractive index \( n \) increases and so does the optical path length in that arm. The optical path difference is:

\[ OPD = (n - 1) \cdot \delta \]

The fringes move and each new cycle of the pattern corresponds to an increase in the OPD of \( \lambda \). After \( m \) fringes are counted, the index of refraction is found via:

\[
(n_1 - 1) \cdot \delta \cdot \frac{2\pi}{\lambda_1} = m \cdot 2\pi \rightarrow n_1 = 1 + \frac{m\lambda_1}{\delta}
\]
(3) Sagnac Interferometer

The Sagnac interferometer is a single-beam splitter version of the M-Z; the output beamsplitter is exchanged for a mirror which is reversed to create a loop path. Light travels around the loop in both directions so that the optical path difference is zero for a stable configuration. However, if the interferometer (including the illuminator) is rotated as on a turntable, then light in one path will experience a Doppler shift with increasing frequency (blue-shift), while light in the reverse direction will experience a red shift. The phase of the two beams will change in proportion to the frequency shift, and the superposed light will exhibit a sinusoidal variation in the detected signal over time:

$$\cos[\omega_1t] + \cos[\omega_2t] = 2\cos[\omega_{avg}t] \cdot \cos[\omega_{mod}t]$$

The slower-varying modulation frequency is detectable and linearly proportional to the rotation rate. This device may be used as a gyroscope with no moving parts, and in fact may be constructed from a single optical fiber that forms a loop with counterrotating beams.
(4) Fizeau Interferometer

The Fizeau interferometer uses a single beamsplitter and may be used to measure the difference in shape between a test optical surface and a reference surface. In the drawing, the physical length difference between the path reflected from the bottom of the test optic and from the top of the reference surface is $d$.

Part of the incident beam is reflected from the glass-air interface of the test object. This dense-to-rare reflection has no phase shift. The reflection from the glass reference surface is rare-to-dense, and the phase of the light is changed by $\pi$ radians. The two waves are recombined when they emerge from the top of the test surface, and detected. Because the beams traverse the same path in each direction, the optical path difference is doubled, so an increment in the physical path of $\frac{\lambda}{2}$ changes the optical path by $\lambda$ and one fringe cycle is seen. If the test optic is spherical, then the physical path difference $d$ may be expressed in terms of the radius of curvature $R$ and the radial distance $r$:
Pythagoras says that:

\[ R^2 = (R - d)^2 + r^2 \Rightarrow r^2 = 2Rd - d^2 \approx 2Rd \text{ for } d \ll R \]

\[ d \approx \frac{r^2}{2R} \text{ for } d \ll R \]

If the interstice between the optical elements in filled with air, and if \( m \) fringes are counted between two points at radial distances \( r_1 \) and \( r_2 \), then the corresponding thickness change is:

\[ m \frac{\lambda}{2} = \text{OPD} = nd \rightarrow d = m \frac{\lambda}{2} \text{ in air} \]

### 11.3 Diffraction

In geometrical (ray) optics, light is assumed to propagate in straight lines from the source (rectilinear propagation). However, Grimaldi observed in the 1600s that this model does not conform to reality after light interacts with an obstruction. Grimaldi observed that light deviates from straight-line propagation into the shadow region. He named this phenomenon *diffraction*. This spreading of a bundle of rays affects the sharpness of shadows cast by opaque objects; the edges become fuzzy because light propagates into the geometrical shadow region.
Diffraction really is the same phenomenon as interference. In both, the wave character of light creates stationary regions of constructive and destructive interference that may be observed as bright and dark regions. In the simplest case of two sources of infinitesimal size, the superposition wave may be determined by summing spherical wave contributions from the sources; the effect is considered to be interference. If the apertures are large (compared to the wavelength $\lambda$), then the spherical-wave contributions from a large number of subsources are summed (by integrating over the area of the aperture) to determine the total electric field. The superposition electric field vector (magnitude and phase) is the vector sum of the fields due to these spherical-wave subsources. The mathematical model for diffraction is straightforward to develop, though computations may be tedious.

Recall the form of a spherical wave emitted by a source located at the origin of coordinates; energy conservation requires that the energy density of the electric field of a spherical wave decrease as the square of the distance from the source. Correspondingly, the electric field decreases as the distance from the source. The electric field observed at location $[x_1, y_1, z_1]$ due to a spherical wave emitted from the origin is:

$$s [x_1, y_1, z_1, t] = \frac{E_0}{\sqrt{x_1^2 + y_1^2 + z_1^2}} \cos [k_x \cdot x + k_y \cdot y + k_z \cdot z - \omega_0 t]$$

This observation that light from a point source generates a spherical wave is the first step towards *Huygen’s principle*, which states that every point on a wavefront may be modeled as a “secondary source” of spherical waves. The summation of the waves from the secondary sources (sometimes called “wavelets”) produces a new wavefront that is “farther downstream” in the optical path.
11.3 DIFFRACTION

In the more general case of a spherical wave emitted from a source located at coordinates \([x_0, y_0, z_0]\) and observed at \([x_1, y_1, z_1]\) has the form:

\[
s[r, t] = \frac{E_0}{r} \cos (k_0 \cdot r - \omega_0 t) \rightarrow \frac{E_0}{r} \exp (k_0 \cdot r - \omega_0 t)
\]

where \(|r| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}\) and \(|k_0| = \frac{2\pi}{\lambda_0}\).

For large \(|r|\), the spherical wave may be approximated accurately as a paraboloidal wave, and for VERY large \(|r|\) the sphere becomes a plane wave. The region where the first approximation is acceptable defines Fresnel diffraction, while the more distant region where the second approximation is valid is the Fraunhofer diffraction region.

11.3.1 Diffraction Integrals

Consider the electric field emitted from a point source located at \([x_0, y_0, z_0 = 0]\). The wave propagates in all directions. The electric field of that wave is observed on an observation plane centered about coordinate \(z_1\). The location in the observation plane is described by the two coordinates \([x_1, y_1]\). The electric field at \([x_1, y_1]\) at this distance \(z_1\) from a source located in the plane \([x_0, y_0]\) centered about \(z = z_0 = 0\) is:

\[
E[x_1, y_1; z_1, x_0, y_0, 0] = \frac{E_0}{r} \cos (k_0 \cdot r - \omega_0 t), \text{ where } |r| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}
\]

Though this may LOOK complicated, it is just an expression of the electric field propagated as a spherical wave from a source in one plane to the observation point in another plane; the amplitude decreases as the reciprocal of the distance and the phase is proportional to the distance and time. Diffraction calculations based on the superposition of spherical waves is Rayleigh-Sommerfeld diffraction.

Now, observe the electric field at that same location \([x_1, y_1, z_1]\) that is generated from many point sources located in the \(x - y\) plane located at \(z = 0\). The summation of the fields is computed as an integral of the electric fields due to each point source. The integral is over the area of the source plane. If all sources emit the same amplitude \(E_0\), then the integral is simplified somewhat:

\[
E_{total}[x_1, y_1; z_1] = \int_{-\infty}^{+\infty} E[x_1, y_1; z_1; x_0, y_0, 0] \, dx_0 \, dy_0
= \int_{\text{aperture}} \frac{E_0}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}} \times \cos \left\{ \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} - \omega_0 t \right\} \, dx_0 \, dy_0
\]

This integral may be recast into a different form by defining the shape of the aperture.
to be a 2-D function $f[x_0, y_0]$ in the source plane:

$$E_{total} [x_1, y_1; z_1] = E_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{f[x_0, y_0]}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}} \times \cos \left[ \frac{2\pi}{\lambda_0} \cdot \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2} - \omega_0 t \right] dx_0 dy_0$$

This expression is the *diffraction integral*. Again, this expression LOOKS complicated, but really represents just the summation of the electric fields due to all point sources. Virtually the entire study of optical diffraction is the application of various schemes to simplify and apply this equation. We will simplify it for two cases:

1. Observation plane $z_1$ located near to the source plane $z_0 = 0$; this is “near-field,” or “Fresnel diffraction” (the $s$ in “Fresnel” is silent — the name is pronounced *Fre’nel*).

2. Observation plane $z_1$ located far enough from the source plane (at $z_0 = 0$) so that $z_1 \approx \infty$. This is called “far-field,” or “Fraunhofer diffraction,” which is particularly interesting (and easier to compute results) because the diffraction integral is proportional to the Fourier transform of the object distribution (the shape of the planar aperture).

A schematic of the diffraction regions is shown in the figure.
11.3 DIFFRACTION

Schematic of the diffraction regions for spherical waves emitted by a point source. Rayleigh-Sommerfeld diffraction is based on the spherical waves emitted by the source. Fresnel diffraction is an approximation based on the assumption that the wavefronts are parabolic and with unit amplitude to $\infty$. The “width” of the quadratic phase is indicated by $\alpha_n$; this is the off-axis distance from the origin where the phase change is $\pi$ radians. Fraunhofer diffraction assumes that the spherical wave has traveled a large distance and the wavefronts may be approximated by planes.

11.3.2 Fresnel Diffraction

Consider the first case of the diffraction integral where the observation plane is near to the source plane, where the concept of near must be defined. Note that the distance $|\mathbf{r}|$ appears twice in the expression for the electric field due to a point source – once in the denominator and once in the phase of the cosine. The first term affects the size (magnitude) of the electric field, and the scalar product of the second with the wavevector $\mathbf{k}$ is computed to determine the rapidly changing phase angle of the sinusoid. The optical phase changes very quickly with time (because $\omega_0$ is very large, $\omega_0 \approx 10^{15}$ radians/second) and with distance (because $|\mathbf{k}|$ is very small, $|\mathbf{k}| \approx 10^{-7}$ m), so the phase difference of light observed at one point in the observation plane but generated from two points in the source plane may differ by MANY radians. Simply put, small changes in the propagation distance $|\mathbf{r}|$ are very significant in the computation of the phase, but much less so when computing the amplitude of the electric field. Therefore, the distance may be approximated more crudely in the denominator than in the phase.

Now consider the approximation of the distance $|\mathbf{r}|$. The complete expression is:

$$|\mathbf{r}| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + z_1^2}$$

$$= \sqrt{z_1^2 \cdot \left( 1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2} \right)}$$

$$= z_1 \cdot \sqrt{1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2}}$$

$$= z_1 \cdot \left( 1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2} \right)^{\frac{1}{2}}$$

This is an EXACT expression that may be expanded into a power series by applying the binomial theorem. The general binomial expansion is:

$$(1 + \alpha)^n = 1 + n\alpha + \frac{n \cdot (n - 1)}{2!} \alpha^2 + \frac{n \cdot (n - 1) \cdot (n - 2)}{3!} \alpha^3 + \cdots + \frac{n!}{(n - r)! r!} \alpha^r + \cdots$$

This series converges to the correct value if $\alpha^2 < 1$. For the case $n = \frac{1}{2}$ (square root), the result is:

$$(1 + \alpha)^{\frac{1}{2}} = 1 + \frac{\alpha}{2} - \frac{1}{8} \alpha^2 + \frac{1}{16} \alpha^3 - \cdots$$
which leads to an expression for the distance $|\mathbf{r}|$:

$$
|\mathbf{r}| = z_1 \cdot \left( 1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2} \right)^{\frac{1}{2}} 
= z_1 \left( 1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2} - \frac{1}{8} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^4} + \ldots \right)
$$

If $z_1$ is sufficiently large, terms of second and larger order may be assumed to be sufficiently close to zero that they may be ignored, leaving the approximation:

$$
|\mathbf{r}| \cong z_1 \left( 1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{z_1^2} \right) = z_1 + \frac{1}{2} \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{2z_1}
$$

This may be simplified further by recasting the electric field expression into complex notation:

$$
E[x_1, y_1; z_1, x_0, y_0, 0] \cong \frac{E_0}{z_1} \text{Re} \left\{ \exp \left[ + \frac{2\pi i}{\lambda_0} \left( z_1 + \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{2z_1} \right) - 2\pi \nu_0 t \right] \right\}
= \frac{E_0}{z_1} \text{Re} \left\{ \exp \left[ + \frac{2\pi i}{\lambda_0 z_1} \right] \cdot \exp \left[ -2\pi \nu_0 t \right] \cdot \exp \left[ + \frac{i\pi}{\lambda_0 z_1} \right] \cdot \exp \left[ + \frac{i\pi}{\lambda_0 z_1} \right] \cdot \exp \left[ \frac{(x_1 - x_0)^2 + (y_1 - y_0)^2}{\lambda_0^2 z_1^2} \right] \right\}
$$

The phase of this approximation of the spherical wave includes a constant phase $\frac{2\pi z_1}{\lambda_0}$, a time-varying phase $-2\pi \nu_0 t$, and the last term whose phase is proportional to the square of the distance off-axis from the source point from the observation point. In the approximation, the wavefront emitted by a point source is not a sphere, but rather a paraboloid.

Note the unreasonable part of the assumption of Fresnel diffraction; the wavefront is assumed to have constant squared magnitude regardless of the location $[x_1, y_1]$ where the field is measured. In other words, the paraboloidal wave in Fresnel diffraction has the same “brightness” regardless of how far off axis it is measured.

For larger values of $z_1$ (observation plane farther from the source), the radius of curvature of the approximate paraboloidal waves increases, so the change in phase measured for nearby points in the observation plane decreases. As $z_1$ approaches $\infty$, the paraboloid approaches a plane wave.

This electric field is substituted into the diffraction integral to obtain the approximate expression in the near-field:
11.3 DIFFRACTION

\[ E_{\text{total}}[x_1, y_1; z_1] = \int_{-\infty}^{+\infty} E[x_1, y_1; z_1; x_0, y_0, 0] \, dx_0 \, dy_0 \]

\[ \approx \frac{1}{z_1} \exp \left[ \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ \frac{i\pi}{\lambda_0 z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \, dx_0 \, dy_0 \]

\[ = \frac{1}{z_1} \exp \left[ \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \times \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ \frac{i\pi}{\lambda_0 z_1} (x_1 - x_0)^2 \right] \exp \left[ \frac{i\pi}{\lambda_0 z_1} (x_1 - x_0)^2 \right] \, dx_0 \, dy_0 \]

Again, this LOOKS complicated, but really is just a collection of the few parts that we have considered already. In words, the integral says that the electric field downstream but near to the source function is the summation of paraboloidal fields from the individual sources. The paraboloidal approximation significantly simplifies the computation of the diffracted light.

11.3.3 Fresnel Diffraction Integral as a Convolution

Consider the Fresnel diffraction integral:

\[ F[x_1, y_1] = \int_{-\infty}^{+\infty} f[x_0, y_0] \exp \left[ + \frac{i\pi}{\lambda_0 z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \, dx_0 \, dy_0 \]

Define the exponential to be a function \( h \) that depends on the four variables in a particular way:

\[ h[x_1 - x_0, y_1 - y_0] = \exp \left[ \frac{i\pi}{\lambda_0 z_1} ((x_1 - x_0)^2 + (y_1 - y_0)^2) \right] \]

In other words, the Fresnel diffraction integral may be written as:

\[ F[x_1, y_1] = \int_{-\infty}^{+\infty} f[x_0, y_0] \, h[x_1 - x_0, y_1 - y_0] \, dx_0 \, dy_0 \]

Integral equations of this form abound in all areas of physical science, and particularly in imaging; they are called \textit{convolution integrals}. The function \( h \) is the shape of the integral function and is often called the \textit{impulse response} of the integral operator. In imaging, and particularly in optics, the impulse response often is called the \textit{point-spread function}. In other areas of physics, it has other names (e.g., \textit{Green's function}). The integral operator often is given a shorthand notation, such as the asterisk “*.”
The variables of integration also often are renamed as dummy variables, such as $\alpha, \beta$:

$$F [x, y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f [\alpha, \beta] \ h [x - \alpha, y - \beta] \ d\alpha \ d\beta$$

$$\equiv f [x, y] * h [x, y]$$

where the form of the impulse response for Fresnel diffraction is:

$$h [x, y] = \frac{1}{z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i\frac{\pi (x^2 + y^2)}{\lambda_0 z_1} \right]$$

This impulse response is called a “chirp” function – the real and imaginary parts are both sinusoids with varying spatial frequency and that also differs with distance $z_1$. The parameters of the chirp often are combined into $\sqrt{\lambda_0 z_1} \equiv \alpha_0$

$$h [x, y] = \frac{1}{z_1} \exp \left[ 2\pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \exp \left[ +i\frac{\pi}{\alpha_0^2} \left( \frac{x^2 + y^2}{\lambda_0 z_1} \right) \right]$$

so that the phase of the chirp function is $\pi$ where $\sqrt{x^2 + y^2} = \alpha_0$. Again note that the magnitude of the impulse response is the unit constant:

$$|h [x, y]| = 1 [x, y]$$

which indicates that the assumed illumination from a point source in the Fresnel diffraction region is constant off axis; there is no “inverse square law.” This obviously unphysical assumption limits the usefulness of calculated diffraction patterns to the immediate vicinity of the optical axis of symmetry.

Profiles of the impulse response along a radial axis are shown for $\alpha_0 = 1$ and $\alpha_0 = 2$. The source distance is larger by a factor of four in the second case.
11.3 DIFFRACTION

1-D profiles of the impulse response of Fresnel diffraction for (a) $\sqrt{\lambda z_1} = 1$ and (b) $\sqrt{\lambda z_1} = 2$, so that $z_1$ is four times larger in (b). Note that the phase increases less rapidly with $x$ for increasing distances from the source.

1-D profiles of the impulse response of Fresnel diffraction for $\lambda_0 z_1 = 1$. Note that the magnitude of the impulse response is 1 and the phase is a quadratically increasing function of $x$.

The convolution integral is straightforward to implement.

**Computed Examples of Fresnel Diffraction**

Below are computed simulations of the profiles of diffraction patterns that would be generated from a knife edge at the same distances from the source as shown above. Note the “ringing” at the edges and that the fringes are farther apart when observed farther from the source. Compare these images to actual Fresnel diffraction patterns.
1-D profiles of the irradiance (squared magnitude) of diffraction patterns from a sharp "knife edge" (modeled as the STEP function shown) for the same distances from the origin: (a) $\sqrt{\lambda z_1} = 1$; (b) $\sqrt{\lambda z_1} = 2$ \Rightarrow $z$ is four times larger. Note that the "period" of the oscillation has increased with increasing distance from the source and that the irradiance at the origin is not zero but rather $\frac{1}{4}$.

Since convolution is linear and shift invariant, the "images" of rectangular apertures may be calculated at these two distance by replicating the impulse responses, reversing one, and adding the amplitudes before computing the irradiance.

1-D profiles of the irradiance (squared magnitude) of diffraction patterns from rectangle functions for different distances from the origin by replicating the impulse responses, reversing one, and adding the amplitudes before computing the irradiance: (a) $\sqrt{\lambda z_1} = 1$; (b) $\sqrt{\lambda z_1} = 2$ \Rightarrow $z$ is four times larger.

Characteristics of Fresnel Diffraction

The parabolic approximation to the spherical impulse response of light propagation produces "images" of the original object that have "fuzzy edges" and oscillating
amplitude on the bright side of an edge. At a fixed distance from the object, the width of the diffraction pattern is proportional to the width of the original object; if the object becomes wider, so does the “image” in the Fresnel diffraction pattern.

The Fresnel diffraction patterns of at the same distance from the origin for two rectangles with different widths, showing that the “width” of the Fresnel pattern is proportional to the width of the object.

11.3.4 Diffraction Integral Valid Far from Source

The diffraction integral may be further simplified for the case where the distance from the source to the observation plane is sufficiently large to allow the electric field from an individual source to be approximated by a plane wave. The process may be considered for one of the paraboloidal waves:

\[
\exp \left[ \frac{i\pi}{\lambda_0 z_1} \left( (x_1 - x_0)^2 + (y_1 - y_0)^2 \right) \right] = \exp \left[ \frac{i\pi}{\lambda_0 z_1} \left( x_0^2 + y_0^2 \right) \right] \cdot \exp \left[ \frac{i\pi}{\lambda_0 z_1} \left( x_1^2 + y_1^2 \right) \right] \cdot \exp \left[ -\frac{i2\pi}{\lambda_0 z_1} (x_0x_1 + y_0y_1) \right]
\]

If the source is restricted to be near to the optical axis so that \( x_0, y_0 \approx 0 \) (or, more rigorously, if \( x_0^2 + y_0^2 \ll \lambda_0 z_1 \)), then

\[
\frac{x_0^2 + y_0^2}{\lambda_0 z_1} \approx 0
\]

so that:

\[
\exp \left[ \frac{i\pi(x_0^2 + y_0^2)}{\lambda_0 z_1} \right] \approx 1.
\]
This resulting diffraction integral simplifies to:

\[
F [x_1, y_1] \propto \int_{-\infty}^{+\infty} f [x_0, y_0] \exp \left[ + \frac{i \pi}{\lambda_0 z_1} \left( (x_1 - x_0)^2 + (y_1 - y_0)^2 \right) \right] \, dx_0 \, dy_0
\]

\[
\cong \exp \left[ \frac{i \pi (x_1^2 + y_1^2)}{\lambda_0 z_1} \right] \cdot \int_{-\infty}^{+\infty} f [x_0, y_0] \cdot \exp \left[ - \frac{2 \pi i}{\lambda_0 z_1} (x_0 x_1 + y_0 y_1) \right] \, dx_0 \, dy_0
\]

Similarly, if the observation point is near to the optic axis so that \( x_1^2 + y_1^2 \ll \lambda_0 z_1 \), then:

\[
\exp \left[ \frac{i \pi (x_1^2 + y_1^2)}{\lambda_0 z_1} \right] \cong 1.
\]

Though \( x_0 \) and \( x_1 \) are sufficiently small for these approximations, the third exponential term is retained because their difference may be larger:

\[
\exp \left[ \frac{i \pi (x_1^2 + y_1^2)}{\lambda_0 z_1} \right] \cong \exp \left[ - \frac{2 \pi i (x_0 x_1 + y_0 y_1)}{\lambda_0 z_1} \right]
\]

Considered in the observation plane as functions of \( x_1 \) and \( y_1 \), the phase of the wavefront is proportional to the source variables \( [x_0, y_0] \); the wavefront is a plane. The corresponding approximation for the diffraction integral is:

\[
E_{\text{total}} [x_1, y_1; z_1] = \frac{1}{z_1} \exp \left[ 2 \pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right]
\]

\[
\cdot \int_{-\infty}^{+\infty} f [x_0, y_0] \exp \left[ - \frac{2 \pi i (x_0 x_1 + y_0 y_1)}{\lambda_0 z_1} \right] \, dx_0 \, dy_0
\]

The diffracted light far from the source is a summation of the plane waves generated by each source point. This is called the Fraunhofer diffraction formula, and the resulting patterns are VERY different from Fresnel diffraction from the same aperture. In fact, the formula can be interpreted as a Fourier transform where the frequency coordinates are mapped back to the space domain via \( \xi = \frac{x_1}{\lambda_0 z_1}, \eta = \frac{y_1}{\lambda_0 z_1} \):

\[
E_{\text{total}} [x_1, y_1; z_1] = \left( \frac{1}{z_1} \exp \left[ 2 \pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right)
\]

\[
\cdot \int_{-\infty}^{+\infty} f [x_0, y_0] \exp \left[ - \frac{2 \pi i (x_0 [\frac{x_1}{\lambda_0 z_1}] + y_0 [\frac{y_1}{\lambda_0 z_1}])} \right] \, dx_0 \, dy_0
\]

\[
= \left( \frac{1}{z_1} \exp \left[ 2 \pi i \left( \frac{z_1}{\lambda_0} - \nu_0 t \right) \right] \right) \cdot \mathcal{F}_2 \{ f [x, y] \}_{\xi = \frac{x}{\lambda_0 z_1}, \eta = \frac{y}{\lambda_0 z_1}}
\]
The resulting irradiance patterns are the time-averaged squared magnitudes of the fields.

\[ |E_{total}[x_1, y_1; z_1]|^2 \propto \frac{1}{z_1^2} \left| \mathcal{F}_2 \{ f [x, y] \}_{\xi = \frac{x_1}{\lambda_0 z_1}, \eta = \frac{y_1}{\lambda_0 z_1}} \right|^2 \]

\[ = \frac{1}{z_1^2} \left| \mathcal{F} \left[ \frac{x_1}{\lambda_0 z_1}, \frac{y_1}{\lambda_0 z_1} \right] \right|^2 \]

All of the Fourier transform relationships apply to Fraunhofer diffraction patterns. The patterns (the “images”) scale in inverse proportion to the original functions; larger input functions \( f [x, y] \) produce smaller (and brighter) diffraction patterns. Movement of an object off axis produces a linear phase is the pattern that is not visible in the irradiance, but has an effect if two such patterns are added.

**Computed Examples of Fraunhofer Diffraction**

Below are shown the profiles of square apertures and computed simulations of the resulting amplitude diffraction pattern in the Fraunhofer diffraction region (the irradiance is the squared magnitude of the plotted amplitude). In both cases, the object is a point source located at \( z_1 \equiv \infty \) so that the light “fills” both apertures with zero phase. For a fixed (large) distance from the object, the “images” of the diffracted light get “narrower” and “taller” as the aperture width increases.

*A useful measure of the “width” of the Fraunhofer pattern is labeled for both cases; this is the distance from the center of symmetry to the first zero. This “width” provides a measure of the ability of the system to resolve fine detail, because the individual diffraction patterns generated by two point sources may overlap and become...*
difficult to distinguish as individual sources as the angular separation of the point sources decreases.

Illustrations of resolution in Fraunhofer diffraction: (a) individual images of two point sources in the Fraunhofer domain; (b) sum of the two images, showing that they may be distinguished easily; (c) images of two sources that are closer together; (d) sum showing that it is much more difficult to distinguish the sources.

Fraunhofer Diffraction in Optical Imaging Systems

A monochromatic point object located a long distance away from an imaging system produces a set of wavefronts that are “regularly” spaced (separated by the wavelength) and are approximately planar. The entrance pupil of the optical system (the image of the aperture stop in object space) collects a section of the plane wavefront and the optical elements convert it to a spherical wave that converges to an image “point.” We can use the concept of Fraunhofer diffraction to define the “angular resolution” of the imaging system.

As an introduction, consider an optical system that consists of only the entrance pupil (which coincides with the aperture stop because no other optics are involved).
The sections of the object wavefronts collected by the stop would continue to propagate “downstream.” If observed a long distance from the stop, the irradiance would be the Fraunhofer diffraction pattern of the stop; the smaller the stop, the larger the diffraction pattern and vice versa. Of course, the observed irradiance is the time average of the squared magnitude of the amplitude, and thus is nonnegative.

If the object for this imaging system consisted instead of two monochromatic point sources displaced by a small angle \( \theta \), the diffracted amplitude would be the sum of two replicas of the Fraunhofer diffracted amplitude that are displaced by the same angle \( \theta \). The observed irradiance again is the time average of the squared magnitude of this amplitude. If the aperture stop is “wide” in some sense, then the diffraction patterns will be “narrow” because of the reciprocal relation between the widths of the aperture and its Fraunhofer diffraction pattern. If the diffraction patterns are “narrow”, the fact that the object consisted of two points sources may be apparent. Conversely, if the stop is “narrow” and the diffraction patterns are “wide,” the patterns from the two sources may overlap and be difficult to distinguish. Thus the ability of the system to resolve two point sources separated by the angle \( \theta \) depends on the size of the entrance pupil. Lord Rayleigh derived an equation for the angular resolution of an imaging system for sources with wavelength \( \lambda_0 \) and separated by a distance \( d \). These are located a distance \( L \) from the system whose entrance pupil diameter is \( D \). The Rayleigh criterion for resolution is that:

\[
D \propto \frac{L \lambda_0}{d}
\]

Note the similarity to the formula for the separation between interference fringes in Young’s experiment. The angular radius of the image spot is a good measure of the angular resolution:

\[
\Delta \theta \simeq \frac{d}{L} \geq \frac{\lambda_0}{D}
\]

(note the similarity between this equation and that for Young’s two-slit experiment). For systems with imaging elements that have circular cross-sections, the angular radius of the image spot is somewhat larger:

\[
\Delta \theta \simeq 1.22 \cdot \frac{\lambda_0}{D}
\]

so the diameter of the image spot is:

\[
2 \cdot \Delta \theta \simeq 2.44 \cdot \frac{\lambda_0}{D}
\]

For the 200in-diameter (\( D = 5.08 \) m) Hale telescope on Palomar Mountain, the theoretical minimum angular separation is \( \Delta \theta \simeq 0.24 \cdot \lambda_0 \), for \( \lambda_0 \) measured in meters. In green light (\( \lambda_0 = 550 \text{ nm} \)), the angular separation is:

\[
\Delta \theta \simeq 1.32 \cdot 10^{-7} \text{ radians} = 0.132 \text{ \mu radians} \simeq 0.03 \text{ arc-seconds}
\]
The diameter of the primary mirror of the Hubble space telescope is approximately 1/2 that of the Hale telescope \((D \approx 2.4 \text{m})\), so the angular resolution of the optics at 550 nm is approximately twice as large \((0.26 \mu \text{rad} \approx 0.6 \text{ arc-seconds})\). Of course, there is no atmosphere to mess up the Hubble images.

\[
\text{Fraunhofer Diffraction}
\]

\[
\text{Pattern of } S
\]

\[
\text{Fraunhofer diffraction of stop: the monochromatic point object is located a large distance to the left of the stop and the diffracted light forms a Fraunhofer irradiance pattern on the observation screen a large distance from the stop; if the object consists of two point sources, the diffracted amplitude is the sum of the translated individual amplitudes.}
\]

Of course, real optical imaging systems consist of more than a stop; they also include lenses and/or mirrors that change the curvature of the plane wavefront to create a section of an (approximately) spherical wave that converges to the real image point on the image plane (often called the “focal plane” of the sensor because the object is generally assumed to be a long distance away from the optical system). We can interpret the action of the optics as “bringing infinity close,” i.e., the lens creates a scaled replica of the light pattern that would have been generated at a large distance from the stop. This pattern has been moved to a sensor that is located a short distance from the stop. In other words, the optics move the Fraunhofer diffraction pattern of the stop from its original location \((at \infty)\) to the image plane. The image of a point object is created by the imaging system is a scaled replica of the Fraunhofer diffraction pattern of the entrance pupil (and thus of the aperture stop); the scale factor is the ratio of the original propagation distance \(z_1\) to the effective focal
11.3 DIFFRACTION

length $f_{eff}$. This scaled replica of the Fraunhofer diffraction pattern of the stop is the impulse response of the optical imaging system in monochromatic (coherent) light. Note that the lens also adds a phase factor to the wavefront due to the curvature of the converging wave. This phase factor may be approximated by an additional quadratic-phase factor, but this effect is not generally significant in imaging because the calculation of the squared magnitude eliminates this factor. After accounting for the phase factor, the irradiance image of a single point source in monochromatic light is proportional to the squared magnitude of the Fourier transform of the entrance pupil:

$$|E_{total} [x_1, y_1; f_{eff}]|^2 = h [x_1, y_1; f_{eff}] \propto |F \left[ \frac{x_1}{\lambda_0 \cdot f_{eff}}, \frac{y_1}{\lambda_0 \cdot f_{eff}} \right]|^2$$

So, again, a large-diameter entrance pupil produces a “small” impulse response and “better” angular resolution, if the performance of the system is limited by diffraction.

A circular entrance pupil of diameter $D$ may be written using Gaskill’s notation:

$$f [x, y] = CYL \left( \frac{r}{D} \right)$$

Its Fourier transform is:

$$F [\xi, \eta] = \pi \left( \frac{D}{2} \right)^2 SOMB (D \rho)$$

$$= \pi D^2 \cdot \frac{2 J_1 (\pi D \rho)}{\pi D \rho} = \frac{D}{2 \rho} J_1 (\pi D \rho)$$

We make the substitution:

$$\rho = \sqrt{\xi^2 + \eta^2} \rightarrow \frac{\sqrt{x_1^2 + y_1^2}}{\lambda_0 \cdot f_{eff}} = \frac{r}{\lambda_0 \cdot f_{eff}}$$

$$\Rightarrow D \rho \rightarrow \frac{Dr}{\lambda_0 \cdot f_{eff}}$$

We know that the first zero of $J_1 (\pi u)$ occurs at $u_0 \approx 1.22$, which occurs in this system at:

$$\frac{Dr_0}{\lambda_0 \cdot f_{eff}} \approx 1.22$$

$$\Rightarrow r_0 \approx 1.22 \cdot \lambda_0 \left( \frac{f_{eff}}{D} \right)$$

$$r_0 \approx 1.22 \cdot \lambda_0 \cdot f/\#$$

This is the dimension of the diffraction spot on the image plane (focal plane) and is an easily remembered equation for the spatial resolution of an optical system. That the resolution is directly proportional to the wavelength means that a system with a fixed focal length and entrance pupil diameter can resolve smaller separations if the
wavelength is reduced. Also, systems with smaller focal ratios have better resolution (“faster systems resolve better”).

An optical system imaging a point source creates a scaled replica the Fraunhofer diffraction pattern of the stop on the image plane. In other words, the impulse response of the imaging system is the Fraunhofer diffraction pattern of the aperture stop.

In real life, it is very difficult (meaning very expensive) to fabricate optical imaging systems with large entrance pupils whose performance is constrained by this diffraction limit. The performance of the system is much more often constrained by the limitations of the optical design and fabrication that results in optical aberrations (e.g., spherical aberration, astigmatism, distortion). Generally speaking, the aberrations become more difficult and expensive to correct for larger entrance pupils.

Of course, the ultimate resolution of the Hale telescope actually is limited by atmospheric turbulence, which creates random variations in the air temperature and thus in the refractive index. These variations are often decomposed into the aberrations introduced into the wavefront by the phase errors. The constant phase (“piston”) error has no effect on the irradiance, the squared magnitude of the amplitude. Linear phase errors (“tip-tilt”) move the image from side to side and up-down. Quadratic phase errors (“defocus”) act like additional lenses that move the image plane backwards or forwards along the optical axis. In general, the tip-tilt error is the most significant, which means that correcting this aberration significantly improves the image quality. The field of correcting atmospheric aberrations is called “adaptive optics,” and is an active research area.
Chapter 12

Basic Principles of Digital Image Processing

During the last decade, inexpensive yet powerful digital computers have become widely available and have been applied to a multitude of tasks. By hitching computers with imaging detectors and displays, very capable systems for creating and analyzing imagery have been constructed and are being applied in many arenas. For example, they now are used to reconstruct x-ray and magnetic resonance images in medicine, to analyze multispectral aerial and satellite images for environmental and military uses, to read Universal Product Codes that specify products and prices in retail stores, just to name a few.

This part of the course will investigate the basic principles and introductory applications of digital imaging systems, and includes many simple examples to illustrate the concepts. Most of the images used to illustrate the concepts are rather “crude”, consisting of only 4096 individual picture elements (pixels) in a $64 \times 64$ array. Each pixel has one of up to 64 different brightnesses (gray values or digital counts). The crudeness of the examples is intentional because it allows the effects due to processing on individual pixels to be apparent. In no way do these examples represent the capabilities of most modern digital imaging systems; indeed, it is usually essential that individual pixels not be visible so that the image appears to be a continuously varying function.

*IMAGE: A reproduction or imitation of form of a person or thing.*

*The optical counterpart of an object produced by a lens, mirror, etc.*

.........................Noah Webster

We normally think of an image in the sense of a picture, i.e., a planar representation of the brightness, , i.e., the amount of light reflected or transmitted by an object.

An image is usually a function of two spatial variables, e.g., $f[x, y]$, which represents the brightness $f$ at the Cartesian location $[x, y]$. It may therefore be graphed in three dimensions, with brightness on the $z$-axis.
Function of Two Spatial Coordinates $f[x, y]$

An image may have more than two coordinate dimensions, e.g.,

<table>
<thead>
<tr>
<th>$f[x, y, t_n]$</th>
<th>monochrome “movie”</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f[x, y, \lambda]$</td>
<td>color image</td>
</tr>
<tr>
<td>$f[x, y, \lambda_m]$</td>
<td>discrete set of wavelengths, multispectral image</td>
</tr>
<tr>
<td>$f[x, y, t]$</td>
<td>time-varying monochrome image</td>
</tr>
<tr>
<td>$f[x, y, z]$</td>
<td>3-D image (e.g., hologram)</td>
</tr>
<tr>
<td>$f[x, y, t_n, \lambda_m]$</td>
<td>image discretely sampled in time and wavelength, e.g., color movie</td>
</tr>
<tr>
<td>$f[x, y, z, t_n, \lambda_m]$</td>
<td>reality</td>
</tr>
</tbody>
</table>

Note that 2-D slices can be “cut” from multidimensional images, and the resulting image needn’t be “pictorial,” e.g., consider the 2-D slices “cut” from the 3-D function $f[x, y, t_n]$; $f[x, y]$ is pictorial, $f[x, t_n]$ is not. But the dimensionality of the axes has no effect on the computations; it is perfectly feasible for computers to process and display $f[x, t_n]$ as well as $f[x, y]$.

After converting image information into an array of integers, the image can be manipulated, processed, and displayed by computer. Computer processing is used for image enhancement, restoration, segmentation, description, recognition, coding, reconstruction, transformation.
12.1 Digital Processing

The general digital image processing system may be divided into three components: the input device (or digitizer), the digital processor, and the output device (image display).

1. The digitizer converts a continuous-tone and spatially continuous brightness distribution \( f[x, y] \) to an discrete array (the digital image) \( f_q[n, m] \), where \( n, m \), and \( f_q \) are integers.

2. The digital processor operates on the digital image \( f_q[n, m] \) to generate a new digital image \( g_q[k, \ell] \), where \( k, \ell \), and \( g_q \) are integers. The output image may be represented in a different coordinate system, hence the use of different indices \( k \) and \( \ell \).

3. The image display converts the digital output image \( g_q[k, \ell] \) back into a continuous-tone and spatially continuous image \( g[x, y] \) for viewing. It should be noted that some systems may not require a display (e.g., in machine vision and artificial intelligence applications); the output may be a piece of information. For example, a digital imaging system that was designed to answer the question, Is there evidence of a cancerous tumor in this x-ray image?, ideally would have two possible outputs (YES or NO), i.e., a single bit of information.

We shall first consider the mathematical description of image digitizing and display devices, and follow that by a long discussion of useful processing operations. Some aspects of this material are covered in more depth in the linear mathematics sequence, SIMG-716,717.
Chapter 13

Review of Sampling

13.1 Digitization

Digitization is the conversion of a continuous-tone and spatially continuous brightness distribution \( f[x, y] \) to an discrete array of integers \( f[q][n, m] \) by two operations which will be discussed in turn:

(A) SAMPLING – a function of continuous coordinates \( f[x, y] \) is evaluated on a discrete matrix of samples indexed by \([n, m]\).

(B) QUANTIZATION – the continuously varying brightness \( f \) at each sample is converted to a one of set of integers \( f[q] \) by some nonlinear thresholding process.

The digital image is a matrix of picture elements, or pixels if your ancestors are computers. Video descendents (and imaging science undergraduates) often speak of pels (often misspelled pelz). Each matrix element is an integer which encodes the brightness at that pixel. The integer value is called the gray value or digital count of the pixel.

Computers store integers as BInary digiTS, or bits (0,1)

2 bits can represent: \( 00\Delta = 0.,\ 01\Delta = 1,\ 10\Delta = 2.,\ 11\Delta = 3.; \)a total of \( 2^2 = 4 \) numbers.

(The symbol “\( \Delta \)” denotes the binary analogue to the decimal point “.”, that is, the binary point divides the ordered bits with positive and negative powers of 2).

\( m \) BITS can represent \( 2^m \) numbers \( \implies 8 \) BITS = 1 BYTE \( \implies 256 \) decimal numbers, \([0, 255]\)

Note that digitized image contains a finite amount of information: the number of bits required to store the data. This will usually be less than the quantity of information in the original image. In other words, digitization creates errors. We will discuss digitizing and reconstruction error after describing the image display process.
13.2 Sampling

This operation derives a discrete set of data points at (usually) uniform spacing. In its simplest form, sampling is expressed mathematically as multiplication of the original image by a function that measures the image brightness at discrete locations:

\[ f_s [n \cdot \Delta x] = f [x] \cdot s [x; n \cdot \Delta x] \]

where:
\( f [x] \) = brightness distribution of input image
\( s [x; n \cdot \Delta x] \) = sampling function
\( f_s [n \cdot \Delta x] \) = sampled input image defined at coordinates \( n \cdot \Delta x \)

The ideal sampling function for functions of continuous variables is generated from the so-called “Dirac delta function” \( \delta [x] \), which is defined by many authors, including Gaskill. For the (somewhat less rigorous) purpose here, we may consider the sampling function to be the sum of uniformly spaced “discrete” Dirac delta functions, which Gaskill calls the \textit{COMB} and Bracewell calls it the \textit{SHAH}:

\[ s [x; n \cdot \Delta x] \equiv \begin{cases} 
1 & \text{if } x = n \cdot \Delta x (n = 0, \pm 1, \pm 2, \ldots) \\
0 & \text{otherwise}
\end{cases} \]

The \textit{COMB} function defined by Gaskill (called the \textit{SHAH} function by Bracewell).

13.2.1 Ideal Sampling

Multiplication of the input \( f [x] \) by a \textit{COMB} function merely evaluates \( f [x] \) on the uniform grid of points located at \( n \cdot \Delta x \), where \( n \) is an integer. Because it measures the value of the input at an infinitesimal point, this is a mathematical idealization that cannot be implemented in practice. Even so, the discussion of ideal sampling usefully introduces some essential concepts.

Consider ideal sampling of a sinusoidal input function with spatial period \( X_0 \) that
is ideally sampled at intervals separated by $\Delta x$:

$$f[x] = \frac{1}{2} \left[ 1 + \cos \left( \frac{2\pi x}{X_0} + \phi \right) \right]$$

$$\implies f_s[n \cdot \Delta x] = \frac{1}{2} \left[ 1 + \cos \left( \frac{2\pi x}{X_0} + \phi \right) \right] \cdot COMB \left[ \frac{x}{\Delta x} \right]$$

The amplitude of the function at the sample indexed by $n$ is:

$$f_s[n \cdot \Delta x] = \frac{1}{2} \left( 1 + \cos \left[ \frac{2\pi x}{X_0} + \phi \right] \right) \cdot \delta [x - n \cdot \Delta x]$$

$$= f_s[n \cdot \Delta x] = \frac{1}{2} \cdot \left( 1 + \cos \left[ \frac{\Delta x}{X_0} \cdot \phi \right] \right)$$

The dimensionless parameter $\frac{\Delta x}{X_0}$ is the ratio of the sampling interval to the spatial period (wavelength) of the sinusoid and is a measurement of the fidelity of the sampled image. Mathematical expressions for the sampled function $f_s$ obtained for several values of $\frac{\Delta x}{X_0}$ are:

Case I: $\frac{\Delta x}{X_0} = \frac{1}{12}, \phi = 0 \implies f_s[n] = \frac{1}{2} \cdot \left( 1 + \cos \left[ \frac{\pi n}{6} \right] \right)$

Case II: $\frac{\Delta x}{X_0} = \frac{1}{2}, \phi = 0 \implies f_s[n] = \frac{1}{2} \cdot (1 + \cos [\pi n]) = \frac{1}{2} [1 + (-1)^n]$

Case III: $\frac{\Delta x}{X_0} = \frac{1}{2}, \phi = -\frac{\pi}{2} \implies f_s[n] = \frac{1}{2} \cdot (1 + \sin [\pi n]) = \frac{1}{2}$

Case IV: $\frac{\Delta x}{X_0} = \frac{3}{4}, \phi = 0 \implies f_s[n] = \frac{1}{2} \cdot \left( 1 + \cos \left[ \frac{3\pi n}{2} \right] \right)$

Case V: $\frac{\Delta x}{X_0} = \frac{5}{4}, \phi = 0 \implies f_s[n] = \frac{1}{2} \cdot \left( 1 + \cos \left[ \frac{5\pi n}{2} \right] \right)$
CHAPTER 13 REVIEW OF SAMPLING

Illustration of sampling of a biased sinusoid, showing aliasing if the signal oscillates with a period smaller than $2 \cdot \Delta x$.

The output evaluated for $\frac{\Delta x}{X_0} = \frac{1}{2}$ depends on the phase of the sinusoid; if sampled at the extrema, then the sampled signal has the same dynamic range as $f[x]$ (i.e., it is fully modulated), show no modulation, or any intermediate value. The interval $\Delta x = \frac{X_0}{2}$ defines the Nyquist sampling limit. If $\frac{\Delta x}{X_0} > \frac{1}{2}$ sample per period, then the same set of samples could have been obtained from a sinusoid with a longer period and a different sampling interval $\Delta x$. For example, if $\frac{\Delta x}{X_0} = \frac{3}{4}$, then the reconstructed function appears as though obtained from a sinusoid with period $X_0' = 3X_0$ if sampled with $\frac{\Delta x'}{X_0'} = \frac{1}{4}$. In other words, the data set of samples is ambiguous; the same samples...
could be obtained from more than one input, and thus we cannot distinguish among the possible inputs based only on knowledge of the samples.

13.3 Aliasing – Whittaker-Shannon Sampling Theorem

As just demonstrated, the sample values obtained from a sinusoid which has been sampled fewer than two times per period will be identical to those from a sinusoid with a longer period. This ambiguity is called aliasing in sampling, but similar effects show up whenever periodic functions are multiplied or added. In other disciplines, these go by different names such as beats, moiré fringes, and heterodyning. To illustrate, consider the product of two sinusoidal functions with the different periods $X_1$ and $X_2$ (and thus spatial frequencies $\xi_1 = \frac{1}{X_1}, \xi_2 = \frac{1}{X_2}$).

$$\cos [2\pi \xi_1 x] \cdot \cos [2\pi \xi_2 x] = \frac{1}{2} \cos [2\pi (\xi_1 + \xi_2) x] + \frac{1}{2} \cos [2\pi (\xi_1 - \xi_2) x]$$

The second term oscillates slowly and is the analog of the aliased signal.

Though the proof is beyond our mathematical scope at this time, we state that a sinusoidal signal that has been sampled without aliasing can be perfectly reconstructed from its ideal samples. This will be demonstrated in the section on image displays. Also without proof, we make the following claim:

*Any function can be expressed as a unique sum of sinusoidal components with (generally) different amplitudes, frequencies, and phases.*

If the sinusoidal representation of $f[x]$ has a component with a maximum spatial frequency $\xi_{\text{max}}$, and if we sample $f[x]$ so that this component is sampled without aliasing, then all sinusoidal components of $f[x]$ will be adequately sampled and $f[x]$ can be perfectly reconstructed from its samples. Such a function is band-limited and $\xi_{\text{max}}$ is the cutoff frequency of $f[x]$. The corresponding minimum spatial period is $X_{\text{min}} = \frac{1}{\xi_{\text{max}}}$. Thus the sampling interval $\Delta x$ can be found from:

$$\frac{\Delta x}{X_{\text{min}}} < \frac{1}{2} \implies \Delta x < \frac{X_{\text{min}}}{2} \implies \Delta x < \frac{1}{2\xi_{\text{max}}}$$

This is the Whittaker-Shannon sampling theorem. The limiting value of the sampling interval $\Delta x = \frac{1}{2\xi_{\text{max}}}$ defines the Nyquist sampling limit. Sampling more or less frequently than the Nyquist limit is oversampling or undersampling, respectively.

$$\Delta x > \frac{1}{2\xi_{\text{max}}} \implies \text{undersampling}$$

$$\Delta x < \frac{1}{2\xi_{\text{max}}} \implies \text{oversampling}$$
The Whittaker-Shannon Sampling Theorem is valid for all types of sampled signals. An increasingly familiar example is digital recording of audio signals (e.g., for compact discs or digital audio tape). The sampling interval is determined by the maximum audible frequency of the human ear, which is generally accepted to be approximately 20kHz. The sampling frequency of digital audio recorders is 44,000 samples per second, which translates to a sampling interval of $\frac{1}{44,000} = 22.7 \mu s$. At this sampling rate, sounds with periods greater than $2 \cdot 22.7 \mu s = 45.4 \mu s$ (or frequencies less than $(45.4 \mu s)^{-1} = 22 kHz$) can theoretically be reconstructed perfectly, assuming that $f [t]$ is sampled perfectly (i.e., at a point). Note that if the input signal frequency is greater than the Nyquist frequency of 22kHz, the signal will be aliased and will appear as a lower-frequency signal in the audible range. Thus the reconstructed signal will be wrong. This is prevented by ensuring that no signals with frequencies above the Nyquist limit is allowed to reach the sampler; higher frequencies are filtered out before sampling.

13.4 Realistic Sampling – Averaging by the Detector

Signals cannot really be sampled at infinitesimal points by multiplication by a COMB; this would mean that the signal would be measured by a detector that has infinitesimal area; such a measurement would have infinitesimal magnitude. In realistic sampling, the continuous input is measured at uniformly spaced samples by using a detector with finite spatial (or temporal) size. The measured signal is an average of the input the detector area, and the image structure is blurred by the averaging process:

Realistic sampling averages the signal over a finite area and blurs information about fine structure that existed in the original continuous image.

The discrete samples are obtained by averaging the input at the sample coordinates. This is mathematically equivalent to averaging the continuous input $f [x]$ with the detector weighting function $h_1 [x]$ and sampling the result by multiplication with a COMB function. Spatial averaging may be expressed as the integral of the product of the input function and the averaging (weighting) function $h_1 [x]$, and is called a convolution. The averaging process is sometimes called prefiltering, or antialiasing:

$$\int_{-\infty}^{+\infty} f [x - x_0] \cdot h_1 [x] \ dx \equiv (f [x] * h_1 [x]) |_{x=x_0}$$

The sampled signal is obtained by multiplying the averaged signal by the COMB:
13.4 REALISTIC SAMPLING – AVERAGING BY THE DETECTOR

\[ f_s[n \cdot \Delta x] = (f[x] * h_1[x]) \cdot \text{COMB} \left[ \frac{x}{\Delta x} \right] \]

where:
\[ f[x] = \text{brightness distribution of input image} \]
\[ h_1[x] = \text{antialiasing prefilter} \]
\[ f_s[n \cdot \Delta x] = \text{sampled input image defined at coordinates } n \cdot \Delta x \]

realistic sampling is composed of two cascaded operations:

1. averaging (convolution, prefiltering) over a detector function, and
2. multiplication by an ideal sampling function \( \text{COMB} \left[ \frac{x}{\Delta x} \right] \)

The nature of the antialiasing prefilter determines the effect of realistic sampling on the output. This is typically characterized by measuring the effect on the modulation of a sinusoidal wave \( f[x] = \frac{1}{2} (1 + \cos(2\pi \xi_0 x)) \). The modulation of a sinusoid is defined as:

\[ m = \frac{f_{\text{max}} - f_{\text{min}}}{f_{\text{max}} + f_{\text{min}}} \text{ for } 0 \leq m \leq 1 \]

Note that modulation is defined for nonnegative (i.e., biased) sinusoids ONLY. The analogous quantity for a nonnegative square wave is called contrast. For example, consider a sinusoid with unit modulation that is sampled by an array with elements of width \( d \) spaced at intervals of width \( \Delta x \) as shown:

The signal is averaged over the detector area, e.g., the sampled value at \( n = 0 \) is:
For $f \left[ x \right]$ as defined above, the set of samples is derived by integrating $f \left[ x \right]$ over the area of width $d$ centered at coordinates that are integer multiples of $\Delta x$:

\[
\frac{1}{d} \int_{n \cdot \Delta x - \frac{d}{2}}^{n \cdot \Delta x + \frac{d}{2}} \left( 1 + \cos \left[ \frac{2\pi x}{X_0} + \phi \right] \right) dx = \frac{1}{2d} \left( \int_{n \cdot \Delta x - \frac{d}{2}}^{n \cdot \Delta x + \frac{d}{2}} dx + \int_{n \cdot \Delta x - \frac{d}{2}}^{n \cdot \Delta x - \frac{d}{2}} \cos \left[ \frac{2\pi x}{X_0} + \phi \right] dx \right)
\]

\[
= \frac{1}{2d} \left[ \left( n \cdot \Delta x + \frac{d}{2} \right) - \left( n \cdot \Delta x - \frac{d}{2} \right) \right] + \frac{1}{2d} \sin \left[ \frac{2\pi n \cdot \frac{\Delta x}{X_0} + \frac{\pi d}{X_0} + \phi \right] \sin \left[ \frac{2\pi n \cdot \frac{\Delta x}{X_0} - \frac{\pi d}{X_0} + \phi \right]_{x=n \cdot \Delta x - \frac{d}{2}}^{x=n \cdot \Delta x + \frac{d}{2}}
\]

\[
= \frac{1}{2} + \frac{1}{2d} \frac{\sin \left[ 2\pi n \cdot \frac{\Delta x}{X_0} + \frac{\pi d}{X_0} + \phi \right] - \sin \left[ 2\pi n \cdot \frac{\Delta x}{X_0} - \frac{\pi d}{X_0} + \phi \right]}{\left( \frac{2\pi}{X_0} \right)}
\]

By defining $\alpha = 2\pi n \cdot \frac{\Delta x}{X_0} + \phi$ and $\beta = \frac{\pi d}{X_0}$, and by using the trigonometric identity:

\[
\sin \left[ \alpha + \beta \right] - \sin \left[ \alpha - \beta \right] = 2 \cos \alpha \sin \beta,
\]

we find an expression for the integral over the detector area:

\[
f_s \left[ n \right] = \frac{1}{2} + \frac{1}{2d} \left( 2 \cos \left[ 2\pi n \cdot \frac{\Delta x}{X_0} + \phi \right] \sin \left[ \frac{\pi d}{X_0} \right] \right)
\]

\[
\equiv \frac{1}{2} + \frac{1}{2} \cos \left[ 2\pi n \cdot \frac{\Delta x}{X_0} + \phi \right] SINC \left[ \frac{d}{X_0} \right]
\]

where $SINC \left[ \alpha \right] \equiv \frac{\sin \left[ \pi \alpha \right]}{\pi \alpha}$. 
13.4 REALISTIC SAMPLING — AVERAGING BY THE DETECTOR

Graph of $SINC \equiv \frac{\sin(\pi x)}{\pi x}$

Note that for constant functions $X_0 = \infty$ and $SINC \left(\frac{d}{X_0}\right) \to 1$; uniform weighted averaging has no effect on constant inputs. The samples of cosine of period $X_0$ obtained with sampling interval $\Delta x$ in the two cases are:

Realistic: $f_s[n] = \frac{1}{2} \cdot \left(1 + \frac{1}{2} SINC \left[\frac{d}{X_0}\right] \cdot \cos \left[2\pi n \cdot \frac{\Delta x}{X_0} + \phi\right]\right)$

Ideal: $f_s[n] = \frac{1}{2} \cdot \left(1 + \cos \left[2n \left(\frac{\Delta x}{X_0}\right) + \phi\right]\right)$

where $d$ is the width of the detector. The amplitude of the realistic case is multiplied by a factor of $SINC \left[\frac{d}{X_0}\right]$, which is less than unity everywhere except at the origin, i.e., where $d = 0$ or $X_0 = \infty$. As the detector size increases relative to the spatial period of the cosine (i.e., as $\frac{d}{X_0}$ increases), then $SINC \left[\frac{d}{X_0}\right] \to 0$ and the modulation of the sinusoid decreases.

The modulation of the image of a sine-wave of period $X_0$, or spatial frequency $\xi = \frac{1}{X_0}$, is reduced by a factor $SINC \left[\frac{d}{X_0}\right] = SINC \left[d\xi_0\right]$.

Example of Reduced Modulation due to Prefiltering

The input function $f[x]$ has a period of 128 units with two periods plotted. It is the sum of six sinusoidal components plus a constant:

$$f[x] = \frac{1}{2} + \frac{1}{2} \sum_{n=1}^{6} \frac{(-1)^{n-1}}{n} \sin \left[\frac{2\pi (2n-1) x}{256}\right].$$
The periods of the component sinusoids are:

\[
X_1 = \frac{128}{1} \text{ units} \Rightarrow \xi_1 = \frac{1}{128} \text{ cycles/unit} \approx 0.0078 \text{ cycles/unit}
\]

\[
X_2 = \frac{128}{3} \text{ units} \approx 42.7 \text{ units} \Rightarrow \xi_2 = \frac{3}{128} \text{ cycles/unit} \approx 0.023 \text{ cycles/unit}
\]

\[
X_3 = \frac{128}{5} \text{ units} = 25.6 \text{ units} \Rightarrow \xi_3 = \frac{5}{128} \text{ cycles/unit} \approx 0.039 \text{ cycles/unit}
\]

\[
X_4 = \frac{128}{7} \text{ units} \approx 18.3 \text{ units} \Rightarrow \xi_4 = \frac{7}{128} \text{ cycles/unit} \approx 0.055 \text{ cycles/unit}
\]

\[
X_5 = \frac{128}{9} \text{ units} \approx 14.2 \text{ units} \Rightarrow \xi_5 = \frac{9}{128} \text{ cycles/unit} \approx 0.070 \text{ cycles/unit}
\]

\[
X_6 = \frac{128}{11} \text{ units} \approx 11.7 \text{ units} \Rightarrow \xi_6 = \frac{11}{128} \text{ cycles/unit} \approx 0.086 \text{ cycles/unit}
\]

The constant bias of 0.5 ensures that the function is positive. The first sinusoidal component \((X_0 = 128\text{ units})\) is the fundamental and carries most of the modulation of the image; the other components (the higher harmonics) have less amplitude. The spatial frequency of each component is much less than the Nyquist limit of 0.5.

\[
SINC \left[ \frac{d}{X_{01}} \right] = SINC [d\xi_1] = SINC \left[ 8 \cdot \frac{1}{128} \right] \approx 0.994
\]

\[
SINC \left[ \frac{d}{X_{02}} \right] = SINC [d\xi_2] = SINC \left[ 8 \cdot \frac{3}{128} \right] \approx 0.943
\]

\[
SINC \left[ \frac{d}{X_{03}} \right] = SINC [d\xi_3] = SINC \left[ 8 \cdot \frac{5}{128} \right] \approx 0.847
\]

\[
SINC \left[ \frac{d}{X_{04}} \right] = SINC [d\xi_4] = SINC \left[ 8 \cdot \frac{7}{128} \right] \approx 0.714
\]

\[
SINC \left[ \frac{d}{X_{05}} \right] = SINC [d\xi_5] = SINC \left[ 8 \cdot \frac{9}{128} \right] \approx 0.555
\]

\[
SINC \left[ \frac{d}{X_{06}} \right] = SINC [d\xi_6] = SINC \left[ 8 \cdot \frac{11}{128} \right] \approx 0.385
\]

Note that the modulation of sinusoidal components with shorter periods (higher frequencies) are diminished more severely by the averaging. A set of prefiltered images for several different averaging widths is shown on a following page. If the detector width is 32 units, the resulting modulations are:
13.4 REALISTIC SAMPLING — AVERAGING BY THE DETECTOR

\[
SINC [d\xi_1] = SINC \left[ 32 \cdot \frac{1}{128} \right] \approx 0.900
\]

\[
SINC [d\xi_2] = SINC \left[ 32 \cdot \frac{3}{128} \right] \approx 0.300
\]

\[
SINC [d\xi_3] \approx -0.180
\]

\[
SINC [d\xi_4] \approx -0.129
\]

\[
SINC [d\xi_5] \approx -0.100
\]

\[
SINC [d\xi_6] \approx +0.082
\]

Note that the components with periods \(X_{04}\) and \(X_{05}\) have negative modulation, i.e., \(f_{\text{max}} < f_{\text{min}}\). The contrast of those components is reversed. As shown, the sampled image looks like a sawtooth with a period of 128 units.

If the detector size is 128, each component is averaged over an integral number of periods and the result is just the constant bias; the modulation of the output is zero:

\[
SINC [d\xi_1] = SINC \left[ \frac{128}{128} \right] = SINC [1] = 0
\]

\[
SINC [d\xi_2] = SINC \left[ \frac{128 \cdot 7}{42} \right] = SINC [3] = 0
\]

For a detector width of 170 units, the modulations are:

\[
SINC [d\xi_1] = SINC \left[ 170 \cdot \frac{1}{128} \right] \approx -0.206
\]

\[
SINC [d\xi_2] = SINC \left[ 170 \cdot \frac{3}{128} \right] \approx -0.004
\]

\[
SINC [d\xi_3] = SINC \left[ 170 \cdot \frac{5}{128} \right] \approx +0.043
\]

\[
SINC [d\xi_4] = SINC \left[ 170 \cdot \frac{7}{128} \right] \approx -0.028
\]

\[
SINC [d\xi_5] \approx -0.004
\]

\[
SINC [d\xi_6] \approx +0.021
\]

Because the first (largest amplitude) sinusoidal component has negative modulation, so does the resulting image. The overall image contrast is reversed; darker areas of the input become brighter in the image.
Illustration of the reduction in modulation due to “prefiltering”: (a) input function \( f[n] \); (b) result of prefiltering with uniform averagers of width \( d = 0 \), \( d = \frac{X_0}{16} \), and \( d = \frac{X_0}{8} \); (c) magnified view of (b), showing the change in the signal; (d) result of filtering with uniform averagers of width \( d = \frac{X_0}{2} \), \( d = X_0 \), and \( d = \frac{X_0}{0.75} \), showing the “contrast reversal” in the last case.
Chapter 14

Review of Quantization

14.1 Tone-Transfer Curve

The second operation of the digitization process converts the continuously valued irradiance of each sample at the detector (i.e., the brightness) to an integer, i.e., the sampled image is quantized. The entire process of measuring and quantizing the brightnesses is significantly affected by detector characteristics such as dynamic range and linearity. The dynamic range of a detector image is the range of brightness (irradiance) over which a change in the input signal produces a detectable change in the output. The input and output quantities need not be identical; the input may be measured in $\frac{W}{\text{mm}^2}$ and the output in optical density. The effect of the detector on the measurement may be described by a transfer characteristic or tone-transfer curve (TTC), i.e., a plot of the output vs. input for the detector. The shape of the transfer characteristic may be used as a figure of merit for the measurement process. A detector is linear if the TTC is a straight line, i.e., if an incremental change in input from any level produces a fixed incremental change in the output. Of course, all real detectors have a limited dynamic range, i.e., they will not respond at all to light intensity below some minimum value and their response will not change for intensities above some maximum. All realistic detectors are therefore nonlinear, but there may be some regions over which they are more-or-less linear, with nonlinear regions at either end. A common such example is photographic film; the TTC is the $H-D$ curve which plots recorded optical density of the emulsion vs. the logarithm of the input irradiance $[\frac{W}{\text{mm}^2}]$. Another very important example in digital imaging is the video camera, whose TTC maps input light intensity to output voltage. The transfer characteristic of a video camera is approximately a power law:

$$V_{out} = c_1 B^{\gamma}_{in} + V_0$$

where $V_0$ is the threshold voltage for a dark input and $\gamma$ (gamma) is the exponent of the power law. The value of $\gamma$ depends on the specific detector: typical values are $\gamma \approx 1.7$ for a vidicon camera and $\gamma \approx 1$ for an image orthicon.
14.2 Quantization

Quantization converts continuously valued measured irradiance at a sample to a member of a discrete set of gray levels or digital counts, e.g., the sample \( f[x, y] \) e.g., \( f[0, 0] = 1.234567890 \cdots \frac{W}{mm^2} \), is converted to an integer between 0 and some maximum value (e.g., 255) by an analog-to-digital conversion (A/D converter or ADC). The number of levels is determined by number of bits available for quantization in the ADC. A quantizer with \( m \) bits defines \( M = 2^m \) levels. The most common quantizers have \( m = 8 \) bits (one byte); such systems can specify 256 different gray levels (usually numbered from \([0, 255]\), where 0 is usually assigned to “black” and 255 to “white”. Images digitized to 12 or even 16 bits are becoming more common, and have 4096 and 65536 levels, respectively.

The resolution, or step size \( b \), of the quantizer is the difference in brightness between adjacent gray levels. It makes little sense to quantize with a resolution \( b \) which is less than the uncertainty in gray level due to noise in the detector system. Thus the effective number of levels is often less than the maximum possible.

Conversion from a continuous range to discrete levels requires a thresholding operation (e.g., truncation or rounding). Some range of input brightnesses will map to a single output level, e.g., all measured irradiances between 0.76 and 0.77 \( \frac{W}{mm^2} \) might map to gray level 59. Threshold conversion is a nonlinear operation, i.e., the threshold of a sum of two inputs is not necessarily the sum of the thresholded outputs. The concept of linear operators will be discussed extensively later, but we should say at this point that the nonlinearity due to quantization makes it inappropriate to analyze the complete digital imaging system (digitizer, processor, and display) by common linear methods. This problem is usually ignored, as is appropriate for large numbers of quantized levels that are closely spaced so that the digitized image appears continuous. Because the brightness resolution of the eye-brain is limited, quantizing to
14.2 QUANTIZATION

only 50 levels is satisfactory for many images; in other words, 6 bits of data is often sufficient for images to be viewed by humans.

The quantization operation is performed by digital comparators or sample-and-hold circuits. The simplest quantizer converts an analog input voltage to a 1-bit digital output and can be constructed from an ideal differential amplifier, where the output voltage $V_{out}$ is proportional to the difference of two voltages $V_{in}$ and $V_{ref}$:

$$V_{out} = \alpha(V_{in} - V_{ref})$$

$V_{ref}$ is a reference voltage provided by a known source. If $\alpha$ is large enough to approximate $\infty$, then the output voltage will be $+\infty$ if $V_{in} > V_{ref}$ and $-\infty$ if $V_{in} < V_{ref}$. We assign the digital value “1” to a positive output and “0” to a negative output. A quantizer with better resolution can be constructed by cascading several such digital comparators with equally spaced reference voltages. A digital translator converts the comparator signals to the binary code. A 2-bit ADC is shown in the figure:

Comparator and 2-Bit ADC. The comparator is a “thresholder;” its output is “high” if $V_{in} > V_{ref}$ and “low” otherwise. The ADC consists of 4 comparators whose reference voltages are set at different values by the resistor-ladder voltage divider.

The translator converts the 4 thresholded levels to a binary-coded signal.

In most systems, the step size between adjacent quantized levels is fixed (“uniform quantization”):

$$b = \frac{f_{max} - f_{min}}{2^m - 1}$$

where $f_{max}$ and $f_{min}$ are the extrema of the measured irradiances of the image samples and $m$ is the number of bits of the quantizer.

If the darkest and brightest samples of a continuous-tone image have measured irradiances $f_{min}$ and $f_{max}$ respectively, and the image is to be quantized using $m$ bits ($2^m$ gray levels), then we may define a set of uniformly spaced levels $f_q$ that span the
dynamic range via:

\[ f_q [x, y] = Q \left\{ \frac{f [x, y] - f_{\min}}{b} \right\} = Q \left\{ \frac{f [x, y] - f_{\min}}{f_{\max} - f_{\min}} \cdot 2^m - 1 \right\} \]

where \( Q \{ \} \) represents the nonlinear truncation or rounding operation, e.g., \( Q \{3.657\} = 3 \) if \( Q \) is truncation or 4 if \( Q \) is rounding. The form of \( Q \) determines the location of the decision levels where the quantizer jumps from one level to the next. The image irradiances are reconstructed by assigning all pixels with a particular gray level \( f_q \) to the same irradiance value \( E [x, y] \), which might be defined by “inverting” the quantization relation. The reconstruction level is often placed between the decision levels by adding a factor \( \frac{b}{2} \):

\[ \hat{E} [x, y] = \left( f_q [x, y] \cdot \frac{E_{\max} - E_{\min}}{2^m - 1} \right) + E_{\min} + \frac{b}{2} \]

Usually (of course), \( \hat{E} [x, y] \neq E [x, y] \) due to the quantization, i.e., there will be quantization error. The goal of optimum quantization is to adjust the quantization scheme to reconstruct the set of image irradiances which most closely approximates the ensemble of original values. The criterion which defines the goodness of fit and the statistics of the original irradiances will determine the parameters of the quantizer, e.g., the set of thresholds between the levels.

The quantizer just described is memoryless, i.e., the quantization level for a pixel is computed independently that for any other pixel. The schematic of a memoryless quantizer is shown below. As will be discussed, a quantizer with memory may have significant advantages.

### 14.3 Quantization Error (“Noise”)

The gray value of the quantized image is an integer value which is related to the input irradiance at that sample. For uniform quantization, where the steps between adjacent levels are the same size, the constant of proportionality is the difference in irradiance between adjacent quantized levels. The difference between the true input irradiance (or brightness) and the corresponding irradiance of the digital level is the quantization error at that pixel:

\[ \epsilon [n \cdot \Delta x, m \cdot \Delta y] \equiv f [n \cdot \Delta x, m \cdot \Delta y] - f_q [n \cdot \Delta x, m \cdot \Delta y]. \]

Note that the quantization error is bipolar in general, i.e., it may take on positive or negative values. It often is useful to describe the statistical properties of the quantization error, which will be a function of both the type of quantizer and the input image. However, if the difference between quantization steps (i.e., the width of a quantization level) is \( b \), is constant, the quantization error for most images may be approximated as a uniform distribution with mean value \( \langle \epsilon [n] \rangle = 0 \) and variance \( \langle (\epsilon [n])^2 \rangle = \frac{b^2}{12} \). The error distribution will be demonstrated for two 1-D 256-sample
14.3 QUANTIZATION ERROR (“NOISE”)  

images. The first is a section of a cosine sampled at 256 points and quantized to 64 levels separated by $b = 1$:

Illustration of the statistics of quantization noise: (a) $f[n] = 63 \cos \left( \frac{2\pi n}{256} \right)$ for $0 \leq n \leq 255$; (b) after quantization by rounding to nearest integer; (c) quantization error $\varepsilon[n] \equiv f[n] - f_q[n]$, showing that $-\frac{1}{2} \leq \varepsilon \leq +\frac{1}{2}$; (d) histogram of 256 samples of quantization error, showing that the statistics are approximately uniform.

The histogram of the error $\epsilon_1[n] = f_1[n] - Q\{f_1[n]\}$ is approximately uniform over the interval $-\frac{1}{2} \leq \epsilon_1 < +\frac{1}{2}$. The computed statistics of the error are $\langle \epsilon_1[n] \rangle = -5.1 \cdot 10^{-4} \cong 0$ and variance is $\sigma^2 = \langle \epsilon_1^2[n] \rangle \cong 0.08 \cong \frac{1}{12}$.

The second image is comprised of 256 samples of Gaussian-distributed random noise in the interval [0, 63] that again is quantized to 64 levels. The histogram of the error $\epsilon_2[n]$ again is approximately uniformly distributed in the interval $[-0.5, +0.5]$ with mean $4.09 \cdot 10^{-2} \cong 0$ and variance $\sigma^2 = \langle \epsilon_2^2[n] \rangle \cong 0.09 \cong \frac{1}{12}$. 
Illustration of the statistics of quantization noise: (a) $f[n]$ is Gaussian noise with measured $\mu = 27.7$, $\sigma = 10.9$ for $0 \leq n \leq 255$; (b) after quantization by rounding to nearest integer; (c) quantization error $\varepsilon[n] \equiv f[n] - f_q[n]$, showing that $-\frac{1}{2} \leq \varepsilon \leq +\frac{1}{2}$; (d) histogram of 256 samples of quantization error, showing that the statistics are STILL approximately uniform.

The total quantization error is the sum of the quantization error over all pixels in the image:

$$\varepsilon = \sum_i \sum_j \varepsilon[n \cdot \Delta x, m \cdot \Delta y].$$

An image with large bipolar error values thus may have a small total error. The mean-squared error (average of the squared error) is a better descriptor of the fidelity of the quantization:

$$\varepsilon^2 = \frac{1}{N} \sum_i \sum_j (\varepsilon^2[n \cdot \Delta x, m \cdot \Delta y]).$$
where \( N \) is the number pixels in the image. If the irradiance is measured in \( \frac{W}{mm^2} \), \( \epsilon^2 \) will have units of \( \left( \frac{W}{mm^2} \right)^2 \). The root-mean-squared (RMS) error has the same dimensions as the error:

\[
\text{RMS Error} \equiv \sqrt{\epsilon^2} = \sqrt{\frac{1}{N} \sum_i \sum_j \epsilon^2 [n \cdot \Delta x, m \cdot \Delta y]}
\]

It should be obvious that the RMS error for one image is a function of the quantizer used, and that the RMS error from one quantizer will differ for different images. It should also be obvious that it is desirable to minimize the RMS error in an image. The brute-force method for minimizing quantization error is to add more bits to the ADC, which increases the cost of the quantizer and the memory required to store the image.

We now extend the discussion to consider the concepts of signal bandwidth and digital data rate, which in turn require an understanding of signal-to-noise ratio (SNR) and its relationship to quantization. Recall that the variance \( \sigma^2 \) of a signal is a measure of the spread of its amplitude about the mean value.

\[
\sigma_f^2 = \int_{-\infty}^{+\infty} \left[ f[x] - \langle f[x] \rangle \right]^2 dx \]

\[
\implies \frac{1}{X_0^2} \int_{-X_0}^{+X_0} \left[ f[x] - \langle f[x] \rangle \right]^2 dx
\]

The signal-to-noise power ratio of an analog signal is most rigorously defined as the dimensionless ratio of the variances of the signal and noise:

\[
\text{SNR} \equiv \frac{\sigma_f^2}{\sigma_n^2}
\]

Thus a large \( \text{SNR} \) means that there is a larger variation of the signal amplitude than of the noise amplitude. This definition of \( \text{SNR} \) as the ratio of variances may vary over a large range – easily several orders of magnitude – so that the numerical values may become unwieldy. The range of \( \text{SNR} \) may be compressed by expressing it on a logarithmic scale with dimensionless units of \( \text{bels} \):

\[
\text{SNR} = \log_{10} \left[ \frac{\sigma_f^2}{\sigma_n^2} \right] = 2 \log_{10} \left[ \frac{\sigma_f}{\sigma_n} \right] \ [\text{bels}]
\]

This definition of \( \text{SNR} \) is even more commonly expressed in units of tenths of a bel so that the integer value is more precise. The resulting metric is in terms of \( \text{decibels} \):

\[
\text{SNR} = 10 \log_{10} \left[ \frac{\sigma_f^2}{\sigma_n^2} \right] = 20 \log_{10} \left[ \frac{\sigma_f}{\sigma_n} \right] \ [\text{decibels}]
\]

Under this definition, \( \text{SNR} = 10 \ dB \) if the signal variance is ten times larger than the noise variance and 20 \( \ dB \) if the standard deviation is ten times larger than that of
the noise.

The variances obviously depend on the statistics (the histograms) of the signal and noise. The variances depend only on the range of gray values and not on their “arrangement” (i.e., numerical “order” or “pictorial” appearance in the image. Since the noise often is determined by the measurement equipment, a single measurement of the noise variance often is used for many signal amplitudes. However, the signal variance must be measured each time. Consider the variances of some common 1-D signals.

14.3.1 Example: Variance of a Sinusoid

The variance of a sinusoid with amplitude \( A_0 \) is easily computed by direct integration:

\[
\sigma_f^2 = \frac{1}{X_0} \int_{-\frac{X_0}{2}}^{\frac{X_0}{2}} (f[x] - \langle f[x] \rangle)^2 \, dx = \frac{1}{X_0} \int_{-\frac{X_0}{2}}^{\frac{X_0}{2}} \left( A_0 \cos \left( \frac{2\pi x}{X_0} \right) \right)^2 \, dx
\]

\[
= \frac{A_0^2}{X_0} \int_{-\frac{X_0}{2}}^{\frac{X_0}{2}} \frac{1}{2} \left[ 1 + \cos \left( 4\pi \frac{x}{X_0} \right) \right] \, dx = \frac{A_0^2}{2X_0} (X_0 + 0)
\]

\[
= \sigma_f^2 = \frac{A_0^2}{2} \text{ for sinusoid with amplitude } A_0
\]

Note that the variance does not depend on the period (i.e., on the spatial frequency) or on the initial phase – it is a function of the histogram of the values in a period and not of the “ordered” values. It also does not depend on any “bias” (additive constant) in the signal. The standard deviation of the sinusoid is just the square root of the variance:

\[
\sigma_f = \frac{A_0}{\sqrt{2}} \approx 0.707 \, A_0 \text{ for sinusoid with amplitude } A_0
\]

14.3.2 Example: Variance of a Square Wave:

The variance of a square wave with the same amplitude also is easily evaluated by integration of the thresholded sinusoid:

\[
f[x] = A_0 \text{ SGN} \left[ \cos \left( \frac{2\pi x}{X_0} \right) \right]
\]

\[
\sigma_f^2 = \frac{1}{X_0} \int_{-\frac{X_0}{2}}^{\frac{X_0}{2}} (f[x] - \langle f[x] \rangle)^2 \, dx = \frac{1}{X_0} \left( \int_{-\frac{X_0}{2}}^{\frac{X_0}{2}} [-A_0]^2 \, dx + \int_{-\frac{X_0}{2}}^{\frac{X_0}{2}} [+A_0]^2 \, dx \right)
\]

\[
= \frac{1}{X_0} \left( A_0^2 \frac{X_0}{2} + A_0^2 \frac{X_0}{2} \right) = A_0^2
\]


\[ \sigma_f^2 = A_0^2 \text{ for square wave with amplitude } A_0 \]

\[ \sigma_f = A_0 \text{ for square wave with amplitude } A_0 \]

Note that the variance of the square wave is larger than that of the sine wave with the same amplitude:

\[ \sigma_f \text{ for square wave with amplitude } A_0 > \sigma_f \text{ for sinusoid with amplitude } A_0 \]

which makes intuitive sense, because the amplitude of the square wave is more often “distant” from its mean than the sinusoid is.

### 14.3.3 Variance of “Noise” from a Gaussian Distribution

A set of amplitudes selected at random from a Gaussian probability distribution is called (conveniently enough) “Gaussian noise.” The most common definition of the statistical distribution is:

\[
p[n] = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]
\]

This probability distribution function has unit area, as required. The Gaussian distribution is specified by the two parameters \( \mu \), the mean value of the distribution, and \( \sigma^2 \), its variance. The standard deviation \( \sigma \) is a measure of the “width” of the distribution and so influences the range of output amplitudes.
14.3.4 Approximations to SNR

Since the variance depends on the statistics of the signal, it is common (though less rigorous) to approximate the variance by the square of the dynamic range, which is the “peak-to-peak signal amplitude” \( f_{\text{max}} - f_{\text{min}} \equiv \Delta f \). In most cases, \((\Delta f)^2\) is larger (and often much larger) than \(\sigma^2\). In the examples of the sinusoid and the square wave already considered, the approximations are:

- **Sinusoid with amplitude** \( A_0 \implies \sigma_f^2 = \frac{A_0^2}{2} \), \((\Delta f)^2 = (2A_0)^2 = 4A_0^2 = 8 \sigma_f^2\)
- **Square wave with amplitude** \( A_0 \implies \sigma_f^2 = A_0^2 \), \((\Delta f)^2 = (2A_0)^2 = 4A_0^2 = 4 \sigma_f^2\)

For the example of Gaussian noise with variance \( \sigma^2 = 1 \) and mean \( \mu \), the dynamic range \( \Delta f \) of the noise technically is infinite, but its extrema often be approximated based on the observation that few amplitudes exist outside of four standard deviations, so that \( f_{\text{max}} \approx \mu + 4\sigma \), \( f_{\text{min}} \approx \mu - 4\sigma \), leading to \( \Delta f \approx 8\sigma \). The estimate of the variance of the signal is then \((\Delta f)^2 \approx 64\sigma_f^2\), which is (obviously) 64 times larger than the actual variance. Because this estimate of the signal variance is too large, the estimates of the SNR thus obtained will be too optimistic.

Often, the signal and noise of images are measured by photoelectric detectors as differences in electrical potential in volts; the signal dynamic range is \( V_f = V_{\text{max}} - V_{\text{min}} \), the average noise voltage is \( V_n \), and the signal-to-noise ratio is:

\[
\text{SNR} = 10 \log_{10} \left( \frac{V_f^2}{V_n^2} \right) = 20 \log_{10} \left( \frac{V_f}{V_n} \right) \quad [\text{dB}]
\]

As an aside, we mention that the signal amplitude (or level) of analog electrical signals often is described in terms of \( \text{dB} \) measured relative to some fixed reference. If the reference level is 1 Volt, the signal level is measured in units of \( \text{dBV} \):

\[
\text{level} = 10 \log_{10} \left( \frac{(V_f V)^2}{V_n^2} \right) \quad \text{dBV} = 20 \log_{10} [V_f] \quad \text{dBV}
\]

The level is measured relative to 1 mV is in units of \( \text{dBm} \):

\[
\text{level} = 10 \log_{10} \left( \frac{(V_f)^2}{(10^{-3} V)^2} \right) \quad \text{dBV} = 10 \log_{10} (V_f^2) \quad \text{dBm}
\]

14.3.5 SNR of Quantization

We can use these definitions to evaluate the signal-to-noise ratio of the quantization process. Though the input signal and the type of quantizer determine the probability density function of the quantization error in a strict sense, the quantization error for the two examples of quantized sinusoidal and Gaussian-distributed signals both exhibited quantization errors that were approximately uniformly distributed. We will continue this assumption that the probability density function is a rectangle. In the case of an \( m \)-bit uniform quantizer (\( 2^m \) gray levels) where the levels are spaced by
intervals of width $b$ over the full analog dynamic range of the signal, the error due to quantization will be (approximately) uniformly distributed over this interval $b$. If the nonlinearity of the quantizer is *rounding*, the mean value of the error is 0; if *truncation* to the next lower integer, the mean value is $-\frac{b}{2}$. It is quite easy to evaluate the variance of uniformly distributed noise:

$$\sigma_n^2 = \frac{b^2}{12}$$

For an $m$-bit quantizer and a signal with maximum and minimum amplitudes $f_{\text{max}}$ and $f_{\text{min}}$, the width of a quantization level is:

$$b = \frac{f_{\text{max}} - f_{\text{min}}}{2^m} \equiv \frac{\Delta f}{2^m}$$

and by assuming that the quantization noise is uniformly distributed, the variance of the quantization noise is:

$$\sigma_n^2 = \frac{b^2}{12} = \frac{(\Delta f)^2}{12 \cdot (2^m)^2} = \frac{(\Delta f)^2}{12 \cdot 2^{2m}}$$

The resulting $\text{SNR}$ is the ratio of the variance of the signal to that of the quantization noise:

$$\text{SNR} \equiv \frac{\sigma_f^2}{\sigma_n^2} = \frac{12 \cdot 2^{2m}}{(\Delta f)^2}$$

which, when expressed on a logarithm scale, becomes:

$$\text{SNR} = 10 \log_{10} \left[ \sigma_f^2 \cdot 12 \cdot 2^{2m} \right] - 10 \log_{10} \left[ (\Delta f)^2 \right]$$

$$= 10 \log_{10} \left[ \sigma_f^2 \right] + 10 \log_{10} [12] + 20m \log_{10} [2] - 10 \log_{10} \left[ (\Delta f)^2 \right]$$

$$\approx 10 \log_{10} \left[ \sigma_f^2 \right] + 10 \cdot 1.079 + 20m \cdot 0.301 - 10 \log_{10} \left[ (\Delta f)^2 \right]$$

$$\approx 6.02 m + 10.8 + 10 \log_{10} \left[ \frac{\sigma_f^2}{(\Delta f)^2} \right] \quad [\text{dB}]$$

The third term obviously depends on both the signal and the quantizer. This equation demonstrates that the $\text{SNR}$ of quantization increases by $\geq 6 \text{ dB}$ for every bit added to the quantizer. If using the (poor) estimate that $\sigma_f^2 = (\Delta f)^2$, then the third term evaluates to zero and the approximate $\text{SNR}$ is:

$$\text{SNR for quantization to } m \text{ bits} \approx 6.02 m + 10.8 + 10 \log_{10} [1]) = 6.02 m + 10.8 \quad [\text{dB}]$$

The statistics of the signal (and thus its variance $\sigma_f^2$) may be approximated for many types of signals (*e.g.*, music, speech, realistic images) as resulting from a random process. The histograms of these signals usually are peaked at or near the mean value $\mu$ and the probability of a gray level decreases for values away from the mean; the signal approximately is the output of a Gaussian random process with variance $\sigma_f^2$. By selecting the dynamic range of the quantizer $\Delta f$ to be sufficiently larger than
\( \sigma_f \), few (if any) levels should be saturated at and clipped by the quantizer. As already stated, we assume that virtually no values are clipped if the the maximum and minimum levels of the quantizer are four standard deviations from the mean level:

\[
\mu_f - f_{\text{min}} = f_{\text{max}} - \mu_f = \frac{\Delta f}{2} = 4 \sigma_f
\]

In other words, we may choose the step size between levels of the quantizer to satisfy the criterion:

\[
\Delta f = 8 \sigma_f \implies \frac{\sigma_f^2}{(\Delta f)^2} = \frac{1}{64}
\]

The \( SNR \) of the quantization process becomes:

\[
SNR = 6.02 \, m + 10.8 + 10 \, \log_{10} \left[ \frac{1}{64} \right] \\
= 6.02 \, m + 10.8 + 10 \, (-1.806) \\
= 6.02 \, m - 7.26 \, [dB]
\]

which is 18 \( dB \) less than the estimate obtained by assuming that \( \sigma_f^2 \cong (\Delta f)^2 \). This again demonstrates that the original estimate of \( SNR \) was optimistic.

This expression for the \( SNR \) of quantizing a Gaussian-distributed random signal with measured variance \( \sigma_f^2 \) may be demonstrated by quantizing that signal to \( m \) bits over the range \( f_{\text{min}} = \mu - 4\sigma_f \) to \( f_{\text{max}} = \mu + 4\sigma_f \), and computing the variance of the quantization error \( \sigma_n^2 \). The resulting \( SNR \) should satisfy the relation:

\[
SNR = 10 \, \log_{10} \left[ \frac{\sigma_f^2}{\sigma_n^2} \right] = (6.02 \, m - 7.26) \, dB
\]

The \( SNR \) of a noise-free analog signal after quantizing to 8 bits is \( SNR_8 \cong 41 \, dB \); if quantized to 16 bits (common in CD players), \( SNR_{16} \cong 89 \, dB \). The best \( SNR \) that can be obtained from analog recording (such as on magnetic tape) is about 65 \( dB \), which is equivalent to that from a signal digitized to 12 bits per sample or 4096 gray levels.

The flip side of this problem is to determine the effective number of quantization bits after digitizing a noisy analog signal. This problem was investigated by Shannon in 1948. The analog signal is partly characterized by its bandwidth \( \Delta \nu \,[Hz] \), which is the analog analogue of the concept of digital data rate \( [\text{bits per second}] \). The bandwidth is the width of the region of support of the signal spectrum (its Fourier transform).

When sampling and quantizing a noisy analog signal, the \textit{bit rate} is determined by the signal-to-noise ratio of the analog signal. According to Shannon, the bandwidth \( \Delta \nu \) of a transmission channel is related to the maximum digital data rate \( R_{\text{max}} \) and the dimensionless signal-to-noise power ratio \( SNR \) via:

\[
R_{\text{max}} \left( \frac{\text{bits}}{\text{sec}} \right) = (2 \cdot \Delta \nu) \log_2 [1 + SNR]
\]
where Shannon defined the $SNR$ to be the ratio of the peak signal power to the average white noise power. It is very important to note that the $SNR$ in this equation is a dimensionless ratio; it is NOT compressed via a logarithm and is not measured in dB. The factor of 2 is needed to account for the negative frequencies in the signal. The quantity $\log_2 [1 + SNR]$ is the number of effective quantization bits, and may be seen intuitively in the following way: if the total dynamic range of the signal amplitude is $S$, the dynamic range of the signal power is $S^2$. If the variance of the noise power is $\sigma^2$, then the effective number of quantization transitions is the power $SNR$, or $\frac{S^2}{\sigma^2}$. The number of quantization levels is $1 + SNR$, and the effective number of quantization bits is $\log_2 [1 + SNR]$.

### 14.4 Quantizers with Memory — Error Diffusion

Another way to change the quantization error is to use a quantizer with memory, which means that the quantized value at a pixel is determined in part by the quantization error at nearby pixels. A schematic diagram of the quantizer with memory is:

![Flow chart for quantizer with memory](image)

A simple method for quantizing with memory that generally results in reduced total error without a priori knowledge of the statistics of the input image and without adding much additional complexity of computation was introduced by Floyd and Steinberg (Proc. SID, 17, pp.75-77, 1975) as a means to simulate gray level images on binary image displays and is known as error diffusion. It is easily adapted to multilevel image quantization. As indicated by the name, in error diffusion the quantization error is from one pixel is used to in the computation of the levels of succeeding pixels. In its simplest form, all quantization error at one pixel is added to the gray level of the next pixel before quantization. In the 1-D case, the quantization level at sample location $x$ is the gray level of the sample minus the error $\epsilon[x - 1]$ at the preceding pixel:

\[
\begin{align*}
    f_q[x] &= Q\{f[x] - \epsilon[x - 1]\} \\
    \epsilon[x] &= f[x] - f_q[x] \\
    &= f[x] - Q\{f[x] - \epsilon[x - 1]\}
\end{align*}
\]

In the 2-D case, the error may be weighted and propagated in different directions. A discussion of the use of error diffusion in ADC was given by Anastassiou (IEEE
The examples on the following pages demonstrate the effects of binary quantization on gray-level images. The images of the ramp demonstrate that why the binarizer with memory is often called pulse-density modulation. Note that the error-diffused images convey more information about fine detail than the images from the memoryless quantizer. This is accomplished by possibly enhancing the local binarization error.

2-D error-diffused quantization for three different gray-scale images: (a) linear ramp image, after quantizing at the midgray level, and after Floyd-Steinberg error diffusion at the midgray level; (b) same sequence for “Lincoln”; (c) same sequence for “Liberty.” The error-diffused images convey more information about the larger spatial frequencies.

14.5 Image Display Systems – Digital-to-Analog Conversion

A complete image processing system must regenerate a viewable signal from the quantized samples. This requires that the digital signal be converted back to a continuously varying brightness distribution; analog estimates of the samples of the original signal.
14.6 IMAGE INTERPOLATION

are derived by a digital-to-analog converter (DAC) and the brightness is spread over the viewing area by the interpolation of the display. Each of these processes will be discussed in turn, beginning with the DAC.

The principle of the DAC is very intuitive; each bit of the digital signal represents a piece of the desired output voltage that is generated by a voltage divider ladder network and a summing amplifier. For example, if a 4-bit digital signal is represented by the binary word ABCD, the desired output voltage is:

\[ V_{\text{out}} = V(8A + 4B + 2C + D) \]

where \( V \) is the desired voltage for a signal represented by the binary word 0001. The appropriate DAC signal is shown below:

![Digital-to-analog converter circuit for 4-bit binary input with bit values ABCD. The circuit generates an analog output voltage \( V = D + 2C + 4B + 8A \).

Variations of the circuit shown are more practical for long binary words, but the principle remains the same. Note that the output voltage is analog, but it is still quantized, i.e., only a finite set of output voltages is possible (ignoring any noise).

14.6 Image Interpolation

The image display generates a continuously varying function \( g[x, y] \) from the processed image samples \( g_q[n, m] \). This is accomplished by defining an interpolator that is placed at each sample with the same amplitude as the sample. The continuously varying reconstructed image is the sum of the scaled interpolation functions. This is analogous to the connect-the-dots puzzle for children to fill in the contours of a picture. Mathematically, interpolation may be expressed as a convolution of the output sampled image with an interpolation function (the postfilter) \( h_2 \). In 1-D:

\[ g[x] = \sum_{n=-\infty}^{\infty} g_q[n \cdot \Delta x] \cdot h_2[x - n \cdot \Delta x] = g_q[x] * h_2[x] \]

In an image display, the form of the interpolation function is determined by the hardware and may have very significant effects on the character of the displayed
image. For common cathode-ray tubes (CRTs – the television tube), the interpolation function is approximately a Gaussian function, but is often further approximated by a circle (or cylinder) function.

The effect of the interpolator on the output is illustrated by a few simple examples. In the 1-D case, the input is a sinusoid with period $X_0 = 64$ sampled at intervals $\Delta x = 8$. The interpolators are a RECT function (nearest-neighbor interpolator), triangle function (linear interpolator), cubic b-spline, and a Gaussian. Examples for 2-D images are shown on following pages.

## 14.6.1 Ideal Interpolation

In the discussion of the Whittaker-Shannon sampling theorem, we have stated that an unaliased function can be perfectly reconstructed from its unaliased ideal samples. Actually, as stated the theorem is true but a bit misleading. To be clearer, we could say the following:

Any function can be perfectly reconstructed from an infinite number of unaliased samples, i.e., samples obtained at a rate greater than two times per period of the highest frequency component in the original function.

In reality, of course, we always have a finite number of samples, and thus we cannot perfectly reconstruct an arbitrary function. Periodic functions may be reconstructed, however, because the samples of a single period will be sufficient to recover the entire function.

In the example just presented, the ideal interpolation function must be something other than a rectangle or Gaussian function. We will again assert without proof that the ideal interpolator for samples separated by a distance $\Delta x$ is:

$$h_2 [x] = \text{SINC} \left( \frac{x}{\Delta x} \right)$$

Note that the SINC function has infinite support and is bipolar; thus it is not obvious how to implement such a display. However, we can illustrate the result by using the example of the sampled cosine already considered. Note that the cosine is periodic.

![Ideal interpolation of the function $f [x] = \cos [2\pi x]$ sampled with $\Delta x = \frac{1}{16}$ unit. The weighted Dirac delta functions at each sample are replaced by weighted SINC](image-url)
functions (three shown, for \( n = 0, -1, -3 \)), which are summed to reconstruct the original cosine function.

### 14.6.2 Modulation Transfer Function of Sampling

We have just demonstrated that images may be perfectly reconstructed from unaliased and unquantized ideal samples obtained at intervals \( \Delta x \) by interpolating with \( SINC \left( \frac{x}{\Delta x} \right) \). Of course, reconstructed images obtained from a finite number of samples systems obtained from a system with averaging and quantization will not be perfect. We now digress to illustrate a common metric for imaging system quality by applying it to realistically sampled systems. Though it is not strictly appropriate, the illustration is still instructive.

Averaging by the detector ensures that the modulation of a reconstructed sinusoid \( g[x] \) will generally be less than that of the continuous input function \( f[x] \), i.e., image modulation is imperfectly transferred from the input to the reconstructed output. The transfer of modulation can be quantified for sinusoids of each frequency; because the averaging effect of the digitizer is fixed, higher-frequency sinusoids will be more affected than lower frequencies. A plot of the modulation transfer vs. spatial frequency is the modulation transfer function or MTF. Note that MTF describes a characteristic of the system, not the input or output.

For ideal sampling (and ideal reconstruction) at all frequencies less than Nyquist, the input function \( f[x] \) is perfectly reconstructed from the sample values \( f_s[n \cdot \Delta x] \), and therefore the modulation transfer function is unity for spatial frequencies less than \( \frac{1}{2} \) cycle per pixel.

**Sinusoids with frequencies \( \xi > \) the Nyquist frequency are aliased by ideal sampling.**

*The “new” frequency is less than the Nyquist frequency.*

*Because the output frequency is different from the input frequency,*

*it is not sensible to talk about the transfer of modulation for frequencies above Nyquist.*

---

**Schematic of the modulation transfer function of the cascade of ideal sampling and ideal interpolation; the MTF is unit at all spatial frequencies out to the Nyquist frequency.**
14.6.3 MTF of Realistic Sampling (Finite Detectors)

We have already demonstrated that the modulation due to uniform averaging depends on the detector width \(d\) and the spatial frequency \(\xi\) of the function as \(SINC(d\xi)\). If the detector size is half the sampling interval \((d = \frac{\Delta x}{2})\), the MTF is:

\[
SINC [d\xi] = SINC \left[ \frac{\Delta x}{2} \cdot \frac{1}{2 \cdot \Delta x} \right] = SINC \left[ \frac{1}{4} \right] = \sin \left( \frac{\pi}{4} \right) = \frac{4}{\pi} \cdot \sqrt{\frac{2}{2}} \approx 0.9 \text{ at the Nyquist frequency.}
\]

i.e., can still be reconstructed perfectly by appropriately amplifying the attenuated sinusoidal components, a process known as inverse filtering that will be considered later. In the common case of detector size equal to sampling interval \((d = \Delta x)\), the minimum MTF is \(SINC [0.5] = 0.637\) at the Nyquist frequency.

*MTF of sampling for \(d = \frac{\Delta x}{2}\) and \(d = \Delta x\).*

By scanning, we can sample the input sequentially, and it is thus possible to a detector size larger than the sampling interval. If \(d = 2 \cdot \Delta x\), then the detector integrates over a full period of a sinusoid at the Nyquist frequency; the averaged signal at this frequency is constant (usually zero, i.e., no modulation).

For larger scanned detectors, the modulation can invert, i.e., the contrast of sinusoids over a range of frequencies can actually reverse. This has already been shown for the case \(\frac{d}{\Delta x} = 1.5 \implies d = 3 \cdot \Delta x\) at the Nyquist rate.
14.7 EFFECT OF PHASE REVERSAL ON IMAGE QUALITY

MTF of scanning systems with \( d = 2 \cdot \Delta x \) and \( d = 3 \cdot \Delta x \), showing that the MTF = 0 at one frequency and is negative for larger spatial frequencies approaching the Nyquist frequency in the second case. This leads to a phase shift of the reconstructed sinusoids.

If the inputs are square waves, the analogous figure of merit is the contrast transfer function or CTF.

14.7 Effect of Phase Reversal on Image Quality

To illustrate the effect on the image of contrast reversal due to detector size, consider the examples shown below.

The input was imaged with two different systems: the MTF of the first system reversed the phase of sinusoids with higher frequencies, while the second did not. Note the sharper edges of the letters in the second image:
Effect of phase reversal on image quality. The edges are arguably “sharper” with the
phase reversal.

14.8 Summary of Effects of Sampling and Quantization

ideal sampling $\Rightarrow$ aliasing if undersampled
    realistic sampling $\Rightarrow$ aliasing if undersampled $\Rightarrow$ modulation reduced at all
nonzero spatial frequencies

quantization $\Rightarrow$ error is inherent in the nonlinear operation
    more bits, less noise $\Rightarrow$ less error
14.9 Spatial Resolution

Photographic resolution is typically measured by some figure of merit like \( \frac{\text{cycles}}{\text{mm}} \) or line pairs per mm, which are the maximum visible spatial frequency of a recorded sine wave or square wave, respectively. Visibility is typically defined by a specific value of the emulsion’s modulation transfer function (MTF, for sinusoids) or contrast transfer function (CTF, for square waves). The specific point of the modulation curve that is used as the resolution criterion may be different in different applications. For example, the resolution of imagery in highly critical applications might be measured as the spatial frequency where the modulation transfer is 0.9, while the frequency where the MTF is 0 may be used for noncritical applications. The spatial resolution of digital images may be measured in similar fashion from the MTF curve due to sampling, which we have just determined to be a function of the sampling interval \( \Delta x \) and the detector width \( d \). The maximum frequency that can be reconstructed is the Nyquist limit \( \xi_{max} = \frac{1}{2\Delta x} \), and the modulation at spatial frequency \( \xi \) varies with the detector size as \( \text{SINC}[d\xi] \). In remote sensing, it is common to use the instantaneous field of view (IFOV) and ground instantaneous field of view (GIFOV). The IFOV is the full-angle subtended by the detector size \( d \) at the entrance pupil of the optical system. The term GIFOV is inappropriate for the definition; spot size would be better. The GIFOV of a digital imaging system is the spatial size of the detector projected onto the object, e.g., the GIFOV of the French SPOT satellite is 10m.
Chapter 15

Point Operations

Once the image data has been sampled, quantized, and stored in the computer, the next task is processing to improve the image, i.e., to extract some (or more) information from the data. The various image processing operations $\mathcal{O}\{ \}$ are applied to the digital input image $f_s[n \cdot \Delta x, m \cdot \Delta y]$ to obtain a (generally) different output $g_s[n' \cdot \Delta x, m' \cdot \Delta y]$. From this point on, all images may be considered as sampled data and thus the subscript $s$ will be ignored and the coordinates will be labeled by $[x, y]$. The general operator has the form

$$\mathcal{O}\{ f[x, y] \} = g[x', y']$$

The various operators $\mathcal{O}$ can be grouped based on the number and location of pixels of the input image $f$ that affect the computation of a particular output pixel $g[x, y]$. One possible set of categories is:

1. **Point Operations on single images:** The gray value of the output image $g$ at a particular pixel $[x, y]$ depends ONLY on the gray value of the same pixel in $f$; examples of these operations include contrast stretching, segmentation based on gray value, and histogram equalization;

2. **Point Operations on multiple images:** The gray value of the output pixel $g[x, y]$ depends on the gray values of the same pixel in a set of input images $f[x, y, t_n]$ or $f[x, y, \lambda_n]$; examples are segmentation based on variations in time or color; multiple-frame averaging for noise smoothing, change detection, and spatial detector normalization;

3. **Neighborhood Operations on one image:** The gray value of $g$ at a particular pixel $[x, y]$ depends on the gray values of pixels in the neighborhood of the same pixel in $f[x, y]$; examples include convolution (as for image smoothing or sharpening), and spatial feature detection (e.g., line, edge, and corner detection);

4. **Neighborhood Operations on multiple images:** This is just a generalization of (3); the pixel $g[x, y]$ depends on pixels in the spatial and temporal
(or spectral) neighborhood of \([x, y, t_n] \) or \([\lambda_n]\). spatial / temporal convolution or spatial / spectral convolution.

5. **Operations based on Object “Shape”** (e.g., “structural” or “morphological”) operations: The gray level of the output pixel is determined by the object class to which a pixel belongs; examples include classification, segmentation, data compression, character recognition.

6. **Geometrical Operations**: The pixels \(f[x, y]\) are remapped to a new coordinate system to obtain \(g[x, y]\): image warping, cartography.

7. “Global” Operations: The gray value of the output image at a pixel depends on the gray values of all of the pixels of \(f[x, y]\); these include image transformations, e.g., Fourier, Hartley, Hough, Haar, Radon transforms.

### 15.1 Point Operations on Single Images

The gray value of each pixel in the output image \(g[x, y]\) depends on the gray value of only the corresponding pixel of the input image \(f[x, y]\). Every pixel of \(f[x, y]\) with the same gray level maps to a single (usually different) gray value in the output image.

*Figure: Schematic of a point operation on a single image: the gray value of the output pixel is determined ONLY by the gray value of the corresponding input pixel.*

In a point operation, the only available parameter that determines the output pixel is the gray value of that one input pixel. Therefore, the point operation must affect all pixels with the same input gray level \(f_0\) in the same fashion; they all change to the same output gray value \(g_0\). When designing the action of the point operation, it often is very useful to know the pixel population \(H\) as a function of gray level \(f\):
the histogram \( H[f] \) of the image. The amplitude of the histogram at gray value \( f \) is proportional to the probability of occurrence of that gray value.

### 15.1.1 Image Histograms

The histogram of the 2-D image \( f[x, y] \) plots the population of pixels with each gray level \( f \). The histogram generally is represented as a 1-D function \( H[f] \) where the independent variable is the gray value \( f \) and the dependent variable is the number of pixels \( H \) with that level. The histogram depicts a particular feature of the image: the population of gray levels. It arguably represents the simplest feature of the image, i.e., a measure of a useful characteristic of the image. The histogram may be called a feature space and its properties may be used to segment the image pixels into component groups.

The histogram often contains valuable global information about the image. For example, most pixels in a low-contrast image are contained in a narrow range of gray levels, so the histogram is concentrated within that small interval. An image with a bimodal histogram (the histogram of gray values exhibits two modes \( \Rightarrow \) a histogram with two “peaks”) often consists of a foreground object (whose pixels are concentrated around a single average gray value) on top of a background object whose pixels have a different average gray value.

Because all pixels in the image must have some gray value in the allowed range, the sum of populations of the histogram bins must equal the total number of image pixels \( N \):

\[
\sum_{f=0}^{f_{\text{max}}} H[f] = N
\]

where \( f_{\text{max}} \) is the maximum gray value (\( f_{\text{max}} = 255 \) for an 8-bit quantizer). The histogram function is a scaled replica of the probability distribution function of gray levels in that image. The discrete probability distribution function \( p[f] \) must satisfy
the constraint:

$$\sum_{f=0}^{f_{\text{max}}} p[f] = 1$$

and therefore the probability distribution and the histogram are related by the simple expression:

$$p[f] = \frac{1}{N} H[f]$$

As will be described later during the discussion of image compression and information theory, the probability distribution leads to a measure of the “quantity” of information in the image, which is the minimum number of bits of data that is required to store the image and generally is measured in bits per pixel. If the probability of gray level $f$ in the image $f[x,y]$ is represented as $p[f]$, the definition of the quantity of information in the image is:

$$I[f] = -\sum_{f=0}^{f_{\text{max}}} p[f] \log_2 (p[f]) \quad \text{[bits per pixel]}$$

From this definition, it is easy to show that the maximum information content is obtained if each gray level has the same probability; in other words, a flat histogram corresponds to maximum information content.

**15.1.2 Histograms of Typical Images**

The form of the image histogram often indicates the character of the original image: a bitonal or binary image will have only two gray levels occupied; an image composed of a small dark object on a large bright background will have a bimodal histogram; a low-contrast image will have a small number of contiguous levels occupied; and an image with a large information content will have a flat histogram.
15.1.3 Cumulative Histogram

Given an $N$-pixel image $f[x,y]$ having gray values in the range $0 \leq f \leq f_{\text{max}}$ and histogram $H[f]$, then the cumulative histogram evaluated at gray value $f_0$ is the number of pixels with gray value less than or equal to $f_0$:

$$C[f_0] = \sum_{f=0}^{f_0} H[f] = \frac{1}{N} \sum_{f=0}^{f_0} p[f]$$

The value of the cumulative histogram at the maximum gray value is the number of pixels in the image:

$$C[f_{\text{max}}] = \sum_{f=0}^{f_{\text{max}}} H[f] = N$$

In the case of a continuous probability distribution, the cumulative histogram is an integral over gray level:

$$C[f_0] = \int_{f=0}^{f_0} H[f] \, df$$

The cumulative histogram is used to derive the mapping that maximizes the global visibility of changes in image gray value (histogram equalization) and for deriving an output image with a specific histogram (histogram specification).
If $H[f]$ is flat, so that every gray level “has” the same number of pixels, then the associated cumulative histogram $C[f]$ increases by the same number of pixels for each gray value, and thus forms a linear “ramp” function at $45^\circ$. The cumulative histogram of a low-contrast image rises rapidly in the gray levels where most pixels lie and slowly over the other levels.

15.1.4 Histogram Modification for Image Enhancement

In point processing, the only parameter available in the pixel transformation is the gray value of that pixel; all pixels of the same gray level must be transformed identically by a point process. The mapping from input gray level $f$ to output level $g$ is called a lookup table, or LUT. Lookup tables can be graphically plotted as transformations $g[f]$ that relate the input gray level (plotted on the $x$-axis) to the output gray level (on the $y$-axis). One such operation is the “spreading out” of the compact
histogram from a low-contrast image over the full available dynamic range to make the image information more visible.

**Examples of Point Operations**

The output resulting from the first mapping below is identical to the input, while the output derived from the second mapping has inverted contrast, i.e., white→black.

*First row: identity lookup table* $g[f] = f$, the resulting image $g[x, y] = f[x, y]$ and its histogram. *Second row: the “negative” lookup table* $g[f] = 255 - f$, the resulting image, and its histogram, showing that the histogram is “reversed” by the lookup table.
First row: lookup table that decreases the contrast, the resulting image, and the histogram showing the concentration of pixels in the middle third of the range.  
Second row: lookup table for linear contrast enhancement, its result when applied to the low-contrast image, and the “spread-out” histogram.

**Remapping Histograms**

As already mentioned, the image histogram is proportional to the probability distribution of gray levels in the image. The action of any lookup table on an image may be modeled as a transformation of probabilities. Recall that the area under any continuous probability density \( p[f] \) or discrete probability distribution \( p_f \) is unity:

\[
\int_{0}^{\infty} p[f] \, df = 1 \implies \sum_{n=0}^{\infty} p_n = 1
\]

For histograms, the corresponding equations are:

\[
\int_{0}^{f_{\text{max}}} H[f] \, df = \sum_{f=0}^{f_{\text{max}}} H[f] = N \text{ (total number of pixels)},
\]

which merely states that every image pixel has some gray level between 0 and \( f_{\text{max}} \).

Similarly for the output image:

\[
\sum_{g=0}^{f_{\text{max}}} H[g] = N.
\]

The input and output histograms \( H[f] \) and \( H[g] \) are related to the lookup table transformation \( g[f] \) via the basic principles of probability theory. The fact that the number of pixels must be conserved requires that incremental areas under the two histograms must match, i.e., if input gray level \( f_0 \) becomes output level \( g_0 \), then:

\[
H[f_0] \, df = H[g_0] \, dg \text{ (continuous gray levels)}
\]

\[
H[f_0] = H[g_0] \text{ (discrete gray levels)}
\]

These equations merely state that all input pixels with level \( f_0 \) are mapped to level \( g_0 \) in the output.

**15.1.5 Jones Plots**

It may be useful to plot the histogram of the input image, the lookup table, and the histogram of the output image on the same *Jones plot* that shows the relationship among them. The input histogram is upside down at the lower-right; the output histogram (rotated 90° counterclockwise) is at the upper left; the lookup table is at the upper right. The new gray level \( g_0 \) is determined by mapping the value \( f_0 \)
through the curve \( g[f] \) onto the vertical axis. In this case, the gray levels of the original low-contrast image are spread out to create a higher-contrast image.

15.2 Histogram Equalization ("Flattening")

The quantity of information in an image \( f[x,y] \) was defined by Shannon to be:

\[
I[f] = - \sum_{f=0}^{f_{\text{max}}} p[f] \log_2 (p[f]) \quad \text{[bits/pixel]}
\]

The meaning of information will be considered in more detail in the discussion of image compression. The information content in an image is maximized if all gray levels are equally populated. This ensures that the differences in gray level within the image are spread out over the widest possible range and thus maximizes the ability to distinguish differences in gray values. Therefore, the act of maximizing image information results in the histogram with gray levels populated as uniformly as possible; the process is called histogram equalization or flattening. The appropriate lookup table is proportional to the cumulative histogram of the input image \( C[f] \).

The mathematical derivation of the appropriate \( g[f] \) is straightforward.

Assume the point operation (lookup table) \( O\{f[x,y]\} = g[x,y] \) equalizes the output histogram, i.e., \( H[g] \) is flat. For simplicity, assume that gray levels are continuous.
and that the lookup transformation \( g[f] \) is monotonically increasing:

\[
g[f_0] = g_0 \\
g[f_0 + \Delta f] = g_0 + \Delta g
\]

Since the lookup table \( g[f] \) must be a monotonically increasing function, then the corresponding inverse operation must exist (call it \( g^{-1} \), so that \( g^{-1}[g_0] = f_0 \). Because the number of pixels must be conserved (each pixel in \( f[x,y] \) is also in \( g[x,y] \)), then the continuous probabilities must satisfy the relation:

\[
p[f] \ df = p[g] \ dg \\
\quad \implies \quad \frac{H[f]}{N} \ df = \frac{H[g]}{N} \ dg \\
\quad \implies \quad H[f] \ df = H[g] \ dg
\]

but \( H[g] \) is constant by assumption (flat histogram), so substitute \( H[g] = k \), a constant:

\[
H[f] \ df = k \ dg
\]

Integrate both sides over the range of allowed levels from 0 to \( f_0 \). The integral evaluated a gray level \( f_0 \) is:

\[
k \cdot \int_0^{f_0} dg = \int_0^{f_0} H[f] \ df = C[f_0] \\
\quad \implies k \cdot g[f_0] - k \cdot g[f = 0] = C[f_0] \\
\quad \implies g[f_0] = \frac{1}{k} \cdot C[f_0] + g[f = 0]
\]

The proportionality constant \( k \) may be evaluated for the number \( R \) of available gray levels (dynamic range) and the image “area”

\[
k = \frac{A}{R}
\]

In the discrete case of an \( N \times N \) image, the proportionality constant is \( k = \frac{N^2}{M} \), where \( M \) is the number of available gray levels and \( N^2 \) is the number of image pixels.

The lookup table that equalizes the image histogram is:

\[
g_{flat}[f_0] = \frac{M}{N^2} C[f_0] + g[f = 0]
\]

Since all pixels with the same discrete gray level \( f_0 \) are treated identically by the transformation, the histogram is “flattened” by spreading densely occupied gray values into “neighboring,” yet sparsely occupied, gray levels. The resulting histogram is as “flat” as can be obtained without basing the mapping on features other than gray level.

The local areas under the input and flattened histograms must match; where \( H[f] \)
is large, the interval $\Delta f$ is spread out to a larger $\Delta g$, thus enhancing contrast. Where $H[f]$ is small, the interval $\Delta f$ maps to a smaller $\Delta g$, thus reducing contrast.

![Continuous Case](image)

*Jones plot for histogram equalization in the continuous case. The areas under the original and “flattened” histograms must match: where $H[f]$ is large, the interval $\Delta f$ is spread out into a larger $\Delta g$, thus enhancing contrast. Where $H[f]$ is small, the interval $\Delta f$ is compressed into a smaller range of $\Delta g$, thus reducing contrast.*

Adjacent well-populated gray levels are spread out, thus leaving gaps (i.e., unpopulated levels) in the output histogram. Pixels in adjacent sparsely populated gray levels of $f[x,y]$ often are merged into a single level in $g[x,y]$. In practice, neighboring values with few pixels may be combined into single levels, thus eliminating the gray-level differences of those pixels.
15.2.1 Example of Histogram Equalization — 1-D “Image”

Because the equalization operation acts on the histogram, and thus only indirectly on the image, the mathematical operation does not depend on the number of spatial dimensions in the input image; the process applies to images with any number of dimensions. For simplicity of presentation, consider first the equalization of a 1-D function. The “image” has the form of a decaying exponential with 256 pixels quantized to 6 bits (values $0 \leq f \leq 63$). The object is shown in (a) its histogram in (b), and its cumulative histogram in (c). Note that the histogram is significantly clustered; there are more “dark” than “light” pixels. The lookup table for histogram equalization is a scaled replica of the cumulative histogram and is shown in (d). The cumulative histogram of the equalized output image is $C[g]$ in (e), and the output histogram $H[g]$ in (f). Note that the form of the output image in (g) is approximately linear, significantly different from the decaying exponential object in (a). In other words, the operation of histogram equalization changed BOTH the spatial character as well as the quantization. The gray levels with large populations (e.g., the dark pixels) have been spread apart in the equalized image, while levels with few pixels have been compressed together.
Illustration of histogram flattening of a 1-D function: (a) 256 samples of \( f[n] \), which is a decaying exponential quantized to 64 levels; (b) its histogram \( H[f] \), showing that the smaller population of larger gray values; (c) cumulative histogram \( C[f] \); (d) Lookup table, which is scaled replica of \( C[f] \); (e) Cumulative histogram of output \( C[g] \), which more closely resembles a linear ramp; (f) histogram \( H[g] \), which shows the wider “spacing” between levels with large populations; (g) Output image \( g[n] \) after quantization.
15.2.2 Example of Histogram Equalization – 2-D “Image”

First row: low-contrast image and its concentrated histogram. Second row: nonlinear cumulative histogram which is proportional to the lookup table for histogram equalization, the resulting histogram (showing the “spreading out” of well-occupied levels and the “smooshing” of levels with small populations), and the resulting image. Third row: the histogram and image resulting from linear contrast enhancement.

15.2.3 Nonlinear Nature of Histogram Equalization

The equalization lookup table in the 1-D example just considered is not a straight line, which means that the gray value \( g \) of the “output” pixel is NOT proportional to \( f \), and thus the mapping of histogram equalization clearly is NOT linear. For subjective applications, where the visual “appearance” of the output image is the only concern, the nonlinearity typically poses no problem. However, if two images with different histograms are to be compared in a quantitatively meaningful way (e.g., to detect seasonal changes from images taken from an airborne platform), then independent
15.3 HISTOGRAM SPECIFICATION

Histogram equalization of the two images before comparison is not appropriate because the images are generally different nonlinear operations. Nonlinear operations produce unpredictable effects on the spatial frequency content of the scene, as you saw in the linear mathematics course. Images should either be compared after applying linear mappings based on pixels of known absolute “brightness”, or after the histogram specification process discussed next.

Nonlinear mappings are used deliberately in a “comandor”, which is a composite word blending “compressor” and “expander”. The process of companding is used to maintain the signal dynamic range and thus improve the “signal-to-noise” ratio in a noise reduction system. A common companding system used in audio systems is the well-known Dolby noise reduction system that is still used for recording analog audio signals on magnetic tape. Analog signals are recorded on tape by aligning appropriate percentages of magnetic domains beneath the recording head. Unavoidable statistical variations in the percentage of aligned domains generates an audible noise signal called tape “hiss” even if no signal is recorded. The Dolby system boosts the amplitude of low-level high-frequency input signals before recording; this is called “pre-emphasis.” The amount of amplification decreases with increasing level of the input signal. The complementary process of “de-emphasis” is performed on playback. The annoying tape hiss is attenuated while the recorded signal is faithfully reproduced. Compandors are also used in digital imaging systems to preserve highlights and shadow detail in digital imaging systems.

15.3 Histogram Specification

It is often useful to transform the histogram of an image to create a new image whose histogram “matches” that of some reference image $f_{\text{ref}}[x,y]$. This process of histogram specification is a generalization of histogram equalization and allows direct comparison of images perhaps taken under different conditions, e.g., LANDSAT images taken through different illuminations or atmospheric conditions. The required transformation of the histogram of $f_1$ to $H[f_{\text{ref}}]$ may be derived by first equalizing the histograms of both images:

$$O_{\text{ref}} \{ f_{\text{ref}}[x,y] \} = e_{\text{ref}}[x,y]$$
$$O_1 \{ f_1[x,y] \} = e_1[x,y]$$

where $e_n[x,y]$ is the image of $f_n[x,y]$ with a flat histogram obtained from the operator $O \{ \}$; the histograms of $e_{\text{ref}}$ and $e_1$ are “identical” (both are flat). The inverse of the lookup table transformation for the reference image is $O^{-1} \{ g_{\text{ref}} \} = f_{\text{ref}}$. The lookup table for histogram specification of the input image is obtained by first deriving the lookup tables that would flatten the histograms of the input and reference image. It should be noted that some gray levels will not be specified by this transformation.
and so must be interpolated. The functional form of the operation is:

\[ g_1[x, y] \text{ (with specified histogram)} = \mathcal{O}_{REF}^{-1} \{ \mathcal{O}_1 \{ f_1 \} \} \]

\[ = [\mathcal{O}_{REF}^{-1} \cdot \mathcal{O}_1] \{ f_1 \} \propto \mathcal{C}_{REF}^{-1} \{ \mathcal{C}_1 \{ f_1 \} \} \]

Schematic of Histogram Specification: given input image \( f_1[x, y] \) and desired “reference” histogram \( H[f_0] \), the input gray value is mapped through its cumulative histogram \( C[f_1] \) and the “inverse” of the reference cumulative histogram \( C[f_0] \) to find the “output” gray value \( f_0 \).

15.4 Application of Histograms to Tone-Transfer Correction

Histogram specification may be used to compensate for a nonlinear tone-transfer curve to ensure that the overall tone scale is linear. The recorded image \( g_1[n \cdot \Delta x, m \cdot \Delta y] \) is obtained from the sampled input image \( f[n \cdot \Delta x, m \cdot \Delta y] \) through the transfer curve (lookup table) \( g_1[f] \), which may be measured by digitizing a linear step wedge. The inverse of the transfer curve may be calculated and cascaded as a second lookup table.
15.5 APPLICATION OF HISTOGRAMS TO IMAGE SEGMENTATION

\( g_2 \) to linearize the total transfer curve:

\[
g_2 [g_1 (f [n \cdot \Delta x, m \cdot \Delta y])] = f [n \cdot \Delta x, m \cdot \Delta y] \\
\Rightarrow g_2 [g_1] = 1 \\
\Rightarrow g_2 [x] = g_1^{-1} [x]
\]

Note that the display may be linearized in similar fashion. Consider a nonlinear digitizer transfer curve of the form \( g_1 [f] = \sqrt{f} \). The correction curve necessary to linearize the system is:

\[
g_2 [f_1 [f]] = g_2 [\sqrt{f}] = f \\
\Rightarrow g_2 [x] = x^2
\]

15.5 Application of Histograms to Image Segmentation

Obviously, histograms may be used to distinguish among objects in the image that differ in gray level; this is the simplest example of segmentation in a feature space. Consider the bimodal histogram that often indicates the presence of a brighter object on a darker background. A gray value \( f_T \) may be determined from the histogram and used as a threshold to segment the “foreground” object. If the histogram clusters overlap (as they seemingly always do), then there are bound to be some false identifications. If background pixels that should be thresholded to black appear as white, we speak of “false positives”, whereas foreground pixels that are classified as background are “false negatives.”

Bimodal histogram, showing the intermixing of the “tails” of the two object classes, which produces false identifications in the image created by the thresholding lookup table.

The threshold lookup table maps all pixels with gray levels greater than \( f_T \) to white and all others to black. If the histogram clusters are disjoint and the threshold is well chosen (and if the image really contains a bright foreground object), a binary image
of the foreground object will result. In this case, the histogram likely is composed of two overlapping Gaussian clusters, and thus some pixels likely will be misclassified by the threshold. Segmentation based on gray level only will be imperfect; there will be false positive pixels (background pixels classified as foreground), and false negative (foreground classified as background). Consider the crude $64 \times 64$ 5-bit image, which shows several distinguishable objects even though the histogram exhibits only two obvious clusters. Segmentation based on this histogram will be unsatisfactory. A theme of the study of image processing operations will be to improve segmentation by gathering or processing data to create histograms with compact and distinguishable clusters.

Segmentation of noisy from histogram; the histogram contains four obvious “clusters”; the lookup table segmented the clusters at level 158, which segmented the sky, clouds, and door from the grass, house, and tree, but some white pixels appear in the grass (“false positives”) and some black pixels in the sky (“false negatives”).

This result illustrates the goal of histogram segmentation; to find some “feature space” (histogram) where the clusters of pixels from the various objects are “compact” and “far apart.”

Other nonlinear mappings may be used for segmentation. For example, the upper LUT on the left maps background pixels to black and foreground pixels to their original gray level. The other is a level slicer; gray levels below $f_1$ and above $f_2$ map to zero while those with $f_1 < f[x, y] < f_2$ are thresholded to white.
The output pixel $g[x, y]$ is a function of the gray value of that pixel in several input images. The input frames may differ in time, wavelength, depth (if they are slices of a 3-D scene), resolution, etc. Most commonly, the gray values of the multiple inputs are combined by arithmetic operations (e.g. addition, multiplication); binary images (i.e. two gray values) may be combined via logical operators (e.g. AND, XOR, etc.). It is also very common to generate a multidimensional histogram from the multiple inputs and use the interpreted data to segment the image via multispectral thresholding.

**Applications:**

1. Image segmentation using multispectral information
2. Averaging multiple frames for noise reduction
3. Change detection by subtraction
4. Windowing images by mask or template multiplication
5. Correct for detector nonuniformity by division

We begin this discussion by immediately digressing to the most common class of multiple-image system, that of color vision where the images differ in the wavelength $\lambda$. 
15.7 Digression: Introduction to Vision and Color

One of the more common systems in this category is the combination of three monochrome images to create a color image. To understand this principle, we need to introduce the human visual system (HVS).

15.7.1 The Eye

The eye is (obviously, and no pun intended) the human sensor that collects radiant energy and forms an image. It contains two positive lens arrangement that generates a real image on the light-sensitive retina. Kepler (1604) described vision in terms of the image projected onto the retina. Sheiner confirmed Kepler’s description by looking at the image created by an eye in 1625.

![Structures in the eye](http://www.tedmontgomery.com/the_eye/)

The eye is nearly spherical (24 mm by 22 mm across). The vitreous humor (behind the lens) “supports” the eyeball, much as air inflates a balloon. It contains microparticles of cellular debris floating within, that produce entoptic perception. Within the sclera is the choroid, a dark layer that absorbs stray light in the same manner as the black coating inside a camera. The retina is a thin layer of light receptor cells that cover the inner surface of the choroid. The retinas of (at least most) human eyes have four kinds of receptors: rods and three kinds of cones. The receptors work via a photochemical reaction in a photopigment.

Humans have \( \approx 75 - 150 \) million rods distributed over the retinal surface. They are arrayed in groups of several rods connected to a single nerve ending. This feature increases the sensitivity of the eye, but decreases the spatial resolution discernible by these receptors. Rods increase in density from the center to about 20° off axis and then decrease in density out to extreme periphery. Thus they provide an overall picture of the field of view. The rods contain rhodopsin, which is a “blue-green” pigment, but they are not sensitive to color and are used at low levels of illumination (scotopic vision). Rod vision is better in low light situations because a ganglion cell will fire when certain threshold signal for all the sensors is reached. It is easier to reach the
threshold with more receptors collecting light. Objects that appear brightly colored in daylight are seen as colorless in dim light because only the rods are stimulated.

Humans have \( \approx 6 - 7 \) million of cones in each eye that are concentrated in the central portion of the retina, the \textit{fovea}. The cones are the color receptors. Each cone is connected to its own nerve ending, thus ensuring a better spatial resolution than the ganged-together rods. Vision generated by cones is called \textit{photopic} and is applicable at “normal” (daylight) levels of illumination. Under these conditions, the eye motion muscles rotate the eyeball until the image of an object falls on the fovea, where the cones are located give color and high resolution. The eye motion “jiggles” the image on the retina; if the image would fade out if kept stationary on a given spot of photoreceptors. Without the fovea the eye would lose 90% of its capability, retaining only peripheral vision.

\[ \text{Angular distribution of rods and cones across the retina, also showing the location of the “blind spot” due to the optic nerve.} \]

For simplicity, we often think of the three cone receptors as sensitive to red, green, and blue light, though this is not strictly correct. The English scientist Thomas Young (a contributor to several areas of optics) hypothesized in 1801 that the eye has three color receptors. This was the basis for his theory of \textit{trichromacy}, which observed that the three independent attributes of colored light (hue, saturation, and lightness) suggests that the eye is sensitive to three independent color input signals. Helmholtz hypothesized that the three types of cones are primarily sensitive to \textit{short}, \textit{medium} (or \textit{intermediate}), and \textit{long} wavelengths (S, M or I, and L), though the response curves were assumed to overlap. The three cones contain pigments with peak response at \( \lambda \approx 447 \text{ nm} \) (“short”, or blue), 540 nm (“medium”, or green), and 577 nm (“long”, or red). Approximate sensitivity curves are shown in the figure:
Of course, some humans (usually males) are born without the use of one of the cone types (due to missing connections to some nerves or connections to the wrong nerves), or a cone type contains an incorrect photopigment. Such folks suffer from *color blindness*.

**Eye Sensors**

![Schematic of a rod (top) and a cone (bottom); light is incident from the left. The shape of the “outer sections” of the receptors (to the right) leads to the names. The photochemical is contained in these outer sections.](image)

The transduction (transformation) of light energy into electrical energy occurs through a chemical reaction via a photosensitive dye. The photochemical in rods is called *rhodopsin* ("visual purple"), which is derived from vitamin A (hence the parental enticement to "eat your carrots"). The interaction of light with the molecules of visual purple causes the electrons to oscillate and change the shape of the molecule. The shape change creates an electrical signal that is transmitted through the nerve synapse to the brain. The “new” molecular shape is not stable and it returns to the original shape after some time delay. The process in cones is similar, but not identical. Because cones work under bright illumination, their absorption process is less efficient; some light is “discarded” by scattering from the cones themselves.

**Latency**

Because the absorption process is chemical, the response of receptor cells is not instantaneous when light arrives or is removed. In the first case, the result is the “latency effect”, while the continuation in response after light is removed is the “persistence response”. The latter effect is the reason why movies and video can convey the illusion of continuous motion from time-sampled data.

**Brightness adaptation and discrimination:**

Digital images are displayed as a discrete set of intensities. Important to understand how the eye differentiate between different intensity levels. The human visual system adapts to a huge irradiance range of $10^{10}$, from the scotopic threshold (dimmest light) to the bright glare limit. It is described by a logarithmic function of the incident light irradiance.
In photonic vision alone, the range of irradiances is about $10^6$. The transition from scotopic to photopic is gradual from $0.001 - 0.1$ millilamberts (or -3 to -1 mL in log scale). This huge dynamic range is accomplished by changing the overall sensitivity of the eye to the average brightness via the process of adaptation. It also is important to understand is the ability of the eye to discriminate between irradiance changes at any specific adaptation level:

Assume that the retina is flooded with a uniform field of illumination of level $\ell$. Then imagine a short, localized flash in the center of the field of view with level $\ell + \Delta \ell$:

The Weber ratio is:

$$\text{Weber ratio} = \frac{\Delta \ell_c}{\ell}$$

where $\Delta \ell_c$ is the increment that a subject detects 50% of the time. A large Weber ratio implies poor sensitivity, meaning that a large percentage change in intensity is required for perception.
At low levels of $\log [f]$ (i.e., in the dark range), $\log \left[ \frac{\Delta V}{V} \right]$ is large, which means that the ability to discriminate brightnesses is poor. As $\log [f]$ increases, so that the background illumination increases, the Weber Ratio decreases, which means that the ability to discriminate brightnesses improves.

Visual system tends to undershoot or overshoot around the boundary of regions of different intensities.

Note that Weber ratio is large at low levels of illumination. The two branches mean that at low levels of illumination vision is carried out by the rods, at high levels is a function of the cones.

### 15.7.2 Convergence

The large number of sensors must be connected to the brain. Since the processing capacity of the brain is limited (though prodigious!), the signals from multiple sensors are combined into a smaller number of nerves by the process of *convergence*. The “degree” of convergence is very different for rods and cones: $\approx 120$ rods and $\approx 6$ cones converge together; in the fovea. Some cones may have their own ganglia.

**Spatial Effect of Convergence**

The individual sensors are tied together by a *neural network* of ganglion cells within the retina.
The lateral connections between cells are weighted to diminish the response from neighbors if one cell is stimulated by a strong signal. This lateral inhibition has the effect of introducing an analogue of an impulse response to the imaging system, where the response “on axis” is positive and those of immediate neighbors are negative. If we think of the system as linear (which it most certainly is not, but is an effective model if the eye is well adapted to a small range of “brightnesses”) and shift-invariant, then we can construct a corresponding transfer function via the Fourier transform. A measurement of the eye response under these conditions leads to some valid conclusions.
The Campbell-Robson chart shown in the figure consists of a “chirped” sinusoidal grating whose frequency increases linearly to the right and whose modulation decreases vertically. When the chart is viewed at a fixed distance, the observer typically can draw a line on the chart where the grating modulation “disappears;” this line typically has a peak value of the modulation at a nonzero spatial frequency.
Approximate contrast sensitivity function of the HVS, showing peak monochromatic ("luminance") response at a spatial frequency of approximately 6 cycles per angular degree ($\cong 1.7 \cdot 10^{-3}$ cycles per arcsecond).

This observation demonstrates that the response (i.e., the output modulation) of the HVS is larger than zero at DC, increases to its peak response at a frequency of about six cycles per angular degree ($\cong 0.003$ cycles per radian), and falls off at larger spatial frequencies, as shown in (b). The corresponding impulse response resembles that shown below in (a).

The model of the “impulse response” of the eye (a) and the corresponding transfer function (b). The impulse response shows the lateral inhibition.

In this model, the impulse response of the HVS is positive at the origin, but becomes negative within a short distance. In words, this indicates that the response of an eye receptor to a bright source actually subtracts from the neighboring receptors; we actually say that the response of one receptor inhibits the responses of those nearby. In other words, these neighboring receptors must be stimulated more strongly to elicit the same response as the first. The HVS “processor” that generates this response is the network of neural connections behind the retina, the so-called visual neural net. We can use this linear shift-invariant model of the HVS to calculate the visual response to a specific input image. If the stimulus is a simple “stairstep” input irradiance, with regular steps of increasing brightness, the output exhibits “overshoots” at the edges. This is the phenomenon of Mach Bands, which result in a nonuniform visual appearance of uniform areas in the vicinity of a transition in gray level. The eye response enhances edges.
The effect on the image of lateral inhibition: (a) the input is a “stairstep” function; (b) the output shows the “overshoot” characteristic of edge enhancement.

15.7.3 Eye Motions

The photoreceptors do not respond to luminance, but rather to changes in luminance; in other words, the eye is an “AC system.”

Saccadic motion

Ballistic point-to-point jumps that the eye executes 1 to 3 times per second when viewing a scene.

Drift, tremor, flicker

Minute, involuntary motions that the eye executes continuously. Necessary to maintaining vision. If you hold eyes in a fixed position image would fade.

15.7.4 Eye Lens

The power of the eye is due to the refractive effect of the cornea and of the eye lens. As we saw during our discussion of optics, the cornea actually provides more of the power, while the lens varies the power of the system.

Accommodation

Accommodation (fine focusing) is performed by the crystalline lens, which is suspended behind the iris by ligaments connected to the ciliary muscles. When relaxed, these muscles pull outward to bring the lens into its “flattest” configuration, so that the radii of the surfaces lengthen, producing a longer focal length. For a perfect eye, from an object at infinity will be focused on the retina if the ciliary muscles are completely relaxed. If the object is positioned closer to the eye, the ciliary muscles
contract, relieving the external tension on the periphery of lens. The radii of the lens surfaces decreases, and thus so does the focal length. This ensures that the image distance remains unchanged. The closest location of an object where the eye can accommodate is the near point, which tends to increase with age (170 mm for a preteenager, \( \approx 250 \) mm for young adult, \( \approx 500 \) mm at middle age, and \( \approx 5000 \) mm at 60 years of age). People whose near points are abnormally close to the eye are nearsighted. This trait was actually valuable in ancient times because nearsighted people could see fine detail up close, as when working on jewelry. Since the ability of the eye to accommodate has some hereditary basis, it was common for families to have a tradition of this kind of precision work.

Normal wavelength range of human vision is in the wavelength \( 390 \text{ nm} \leq \lambda \leq 780 \text{ nm} \) (which actually seems to be too large, particularly on the red end). The limitation on sensitivity to ultraviolet wavelengths is due to absorption by crystalline lens.

**Nearsightedness (myopia)**

Myopic eyes bring parallel rays to focus in front of the retina; in other words, the power of the lens is too large. Myopia is corrected by placing another lens in front of the eye such that the focal point of the combined power lens-eye system is on the retina without changing the power of the eye+auxiliary lens system. We saw in the section on optics that the system power is unchanged if an additional lens is placed at the front focal point of the original system. This can be seen by applying the two-lens equation with the separation distance \( \tau = (\varphi_{\text{eye}})^{-1} \)

\[
\varphi_{\text{system}} = \varphi_{\text{corrector}} + \varphi_{\text{eye}} - \varphi_{\text{corrector}} \cdot \varphi_{\text{eye}} \cdot (\varphi_{\text{eye}})^{-1} \\
= \varphi_{\text{corrector}} + \varphi_{\text{eye}} - \varphi_{\text{corrector}} \\
\varphi_{\text{system}} = \varphi_{\text{eye}}
\]

If you are nearsighted and wear glasses, note that the images are the same size with and without correction. Since corrective contact lenses are placed directly upon the cornea, we can assume that the distance between the corrective lens and the power due to the first surface of the cornea is zero, thus the power of the corrected cornea is the sum of the powers of the two lenses:

\[
\varphi'_{\text{cornea}} = \varphi_{\text{corrector}} + \varphi_{\text{cornea}}
\]

**Farsightedness (hyperopia)**

In this condition, the second focal point lies behind the retina, i.e., the lens power is insufficient, or is placed too close to the retina. To increase the bending of the rays a positive lens is place in front of the eye. A farsighted person can see distant objects sufficiently well, but cannot bring objects close to our eye to increase the angular subtense of fine structure on the retina because the near point is farther away than normal.
Astigmatism

This is an aberration of the eye system due to different radii of curvature of the cornea along different axial directions, thus resulting in different focal lengths for horizontal and vertical objects.

15.7.5 Spatial resolution

The metric of “visual acuity”

- Seek the ability to characterize the resolving power of the eye.
- Snellen tests used by optometrists
- Test for visual acuity where the patient reads Snellen’s chart at a certain distance with one eye, then with the other, and then with both eyes. Acuity is given by the longest distance at which the patient is able to read the letters, divided with the distance at which he should normally be able to read them.

15.8 Color Vision

The three cones contain pigments with peak response at $\lambda \approx 447\text{ nm}$ (“short”, or blue), 540 nm (“medium”, or green), and 577 nm (“long”, or red). Approximate sensitivity curves are shown:

![The response of short, medium, and long cones.](image)

Curves have been determined through physiological measurements of absorption spectra of individual cones in vitro, and psychophysical color matching experiments. The curves are normalized.

The responses of the cones differ, and also the number of short, medium, and long (SML) cones are not equal. There are significantly fewer $S$ cones than the others, so that the $S$ cones contribute little to the overall brightness sensation.

The sensitivity curves provide the basis for color vision: Assume two lights with two wavelengths ($\lambda_1 = 500\text{ nm}$ and $\lambda_2 = 600\text{ nm}$). The first generates a response from $M$ cones that is twice as large as its $L$ response. The second light produces
approximately twice the response in $L$ cones than $M$. Now change relative intensities of the two lights: the individual $L$ and $M$ responses will be different (corresponding to difference in luminance) but their ratios will be constant, which means that we can separate changes in intensity from changes in color. One single type of detector could not accomplish this!

### 15.8.1 Color Matching, Metamers

A consequence of color matching is that two colors can appear identical to the eye even though they have different spectral compositions. These are called *metamers*. They appear to match under one illumination but mismatch when viewed under another. This is because color is a sensation rather than a property of an object; the cones can register the same sensation from an infinite variety of combinations of different light frequencies and amplitudes.

**Grassman’s Law**

The eye can distinguish only three attributes of a light: *hue*, *brightness*, and *saturation*. Hue is that psychological dimension of color which roughly corresponds to wavelength. Brightness is the psychological dimension of color which most closely relates to physical intensity and saturation is the amount of hue a color possesses and is most closely correlated with spectral purity. Two lights with two different colors added to two other lights with the same colors produce mixtures of the same color: if $a = b$ and $c = d$, then $a + c = b + d$. Two lights of the same color each subtracted respectively from two other light of the same color will leave mixtures of the same color: if $a = b$ and $c = d$ then $a - c = b - d$. If one unit of light $a$ has the same color as one unit of $b$, then $ka = kb$. Luminance produced by the additive mixture of a number of lights is the sum of luminances produced separately by each light.

Mathematically Grassman’s laws define a vector space implies we can represent any color as a vector in a 3-D space, where the basis vectors are the primary colors. It’s obvious that the primary colors have to be linearly independent, meaning that any one cannot be created as a weighted sum of the others. In this vector space, an arbitrary color $C$ can be matched by appropriate quantities of the three primaries $R$, $G$, $B$:

$$C = r_1R + g_1G + b_1B$$

where $r_1, g_1, b_1$ are the weights (number of units) applied to each primary to make the match.

**Trichromancy**

The weighted sum means that every colored light $C$ is the physical sum of a number of essentially pure spectral components and can be described by a function $E[\lambda]$ (radiant power per unit wavelength). The result of the experiment across a wide range of test wavelengths produces the so-called *color matching functions* for primaries at 436 nm, 546 nm, and 700 nm. The color matching functions reveal the fundamental fact that
human color vision is trichromatic. Trichromancy simply means that any visible color can be matched by a mixture of three primary colors.

\[
\text{Color matching functions}
\]

Note that the color matching functions include negative weights. For example, this means that \( \lambda = 500 \text{ nm} \) cannot be matched by the three primaries unless some red is subtracted from the mixture. No colors exist that can be used as primaries such that the color matching functions are wholly positive. In 1931, CIE adopted a set of fictitious primaries that result in wholly positive functions:

\[
\text{CIE synthesized color matching functions with positive weights.}
\]

### 15.8.2 XYZ Color Space

To find the \( X, Y, Z \) values corresponding to a particular color object let \( R[\lambda] \) be the spectral transmittance or reflectance of the object, and let \( P[\lambda] \) be the spectral power per unit wavelength interval of the illuminating source. Also let \( \bar{x}_\lambda, \bar{y}_\lambda, \text{ and } \bar{z}_\lambda \)
be the CIE positive-definite color matching functions:

\[ X = k \int_{380}^{780} P(\lambda) R(\lambda) \bar{x}\lambda \, d\lambda \]
\[ Y = k \int_{380}^{780} P(\lambda) R(\lambda) \bar{y}\lambda \, d\lambda \]
\[ Z = k \int_{380}^{780} P(\lambda) R(\lambda) \bar{z}\lambda \, d\lambda \]

where \( k \) is a normalizing factor such that \( Y = 100 \) when the object is a perfect white diffuser or a perfect transmitter of light for the entire visible band. \( Y \) is therefore proportional to total luminance and matches the photopic response of the eye. The discrete forms of the equations are:

\[ X = k \sum_{\lambda=380}^{780} P(\lambda) R(\lambda) \bar{x}\lambda \]
\[ Y = k \sum_{\lambda=380}^{780} P(\lambda) R(\lambda) \bar{y}\lambda \]
\[ Z = k \sum_{\lambda=380}^{780} P(\lambda) R(\lambda) \bar{z}\lambda \]

**Determination of X, Y, Z**

\[
\begin{bmatrix}
X \\
Y \\
Z
\end{bmatrix} =
\begin{bmatrix}
r_1 & g_1 & b_1 \\
r_2 & g_2 & b_2 \\
r_3 & g_3 & b_3
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

The color can be characterized independent of its luminance via:

\[ x = \frac{X}{X+Y+Z} \]
\[ y = \frac{Y}{X+Y+Z} \]
\[ z = 1 - (x + y) = \frac{Z}{X+Y+Z} \]

which are known as chromaticity coordinates. The normalization reduces the 3D color space to a 2D plane that satisfies the constraint \( x + y + z = 1 \). Purely additive combinations are found inside the triangle:
All colors that are possible by a combination of two primaries will lie on a straight line connecting the primaries in this space. Any color that can be derived from any three primaries will lie inside the triangle whose apexes are the primaries; this triangle represents the gamut of the primaries.

15.8.3 CIE Chromaticity Diagram

Additive mixtures of colors lie along a straight line connecting those colors. The complement of any color is found by extending a straight line form that color through white to the opposite side of the CIE diagram.
**Definition 15.1 Chromaticity Diagram:** A two-dimensional Cartesian plot that depicts the multidimensional subjective relationship among colors perceived by the normal human visual system (eyes and nervous system, including the brain) when additively stimulated by (two or more; usually three) discrete monochromatic visible sources (wavelengths). Note: The familiar CIE chromaticity diagram depicts perceived colors plotted as a function of the normalized relative intensity of a defined red (increasingly red with increased ”X”, or abscissa, value) versus the normalized relative intensity of a defined green (increasingly green with increased ”Y”, or ordinate, value). With respect to a given perceived color, as plotted on the chromaticity diagram, the normalized relative intensity of a defined blue at any point is obtained by adding the normalized relative intensities of the red and green, and subtracting the total from 1.

Additive mixtures of colors lie along a straight line connecting those colors. The complement of any color is found by extending a straight line form that color through white to the opposite side of the CIE diagram.

http://www.atis.org/tg2k/_chromaticity_diagram.html

In 1931, the CIE defined three standard primaries (X, Y, Z). The Y primary was intentionally chosen to be identical to the average luminous-efficiency function.
of human eyes. The three standard CIE primary colors are ($\lambda = 435.8 \text{ nm, } 546.1 \text{ nm and } 700 \text{ nm}$).

CIE has recommended the use of other color spaces, all derived from XYZ. $L*a*b*$ is used for non-luminous objects such as textiles, paints, plastics, etc. $L*u*v*$ is helpful in the registration of color differences experienced with flashes, photography, television screen, etc. These systems are useful in specifying small differences between color stimuli. The motivation was to find coordinates that relate in a linear fashion to the perceptual attributes of color (perceptually uniform color spaces)

$L*a*b*$

- $L^*$ is the lightness axis and extends from 0 (black) to 100 (white).
- $a^*$ represents the redness-greenness
- $b^*$ represents the yellowness-blueness

\[
L^* = 116 \left( \frac{Y}{Y_n} \right)^{\frac{1}{3}} - 16
\]
\[
a^* = 500 \left[ \left( \frac{X}{X_n} \right)^{\frac{1}{3}} - \left( \frac{Y}{Y_n} \right)^{\frac{1}{3}} \right]
\]
\[
b^* = 200 \left[ \left( \frac{Y}{Y_n} \right)^{\frac{1}{3}} - \left( \frac{Z}{Z_n} \right)^{\frac{1}{3}} \right]
\]

where $X_n, Y_n, Z_n$ are the coordinates of the reference white.

15.8.4 Color Reproduction

A color imaging system usually goes through three steps in order to reproduce an image:

1. Color separation: Derive R, G, B signals from the input scene, just like the eye would do. System need more than one kind of detector

2. Processing: The system must convert the R, G, B signals from the detectors into output signals that are suitable for the color reconstruction stage. For example a current that controls the amount of red dye to a nozzle. Signal processing may be electronic (TV), chemical (photography), or more complex combination.

3. Reconstruction: Image must be produced by whatever means are appropriate: dyes, phosphors, etc. There are two basic methods of color reproduction: subtractive and additive.
Subtractive Color Reproduction

Color printing uses three colored dyes: cyan, magenta and yellow. These can be thought of as complementary to the additive primaries red, green and blue. In other words, cyan absorbs the red part of the spectrum so by varying its concentration we can vary the absorption of red while having small effect on other parts of the spectrum.

Additive Color Reproduction

Used with lights (e.g., phosphors) of the three additive primary colors. red, green, and blue. There are phosphors with smooth spectral distributions in the blue and green, but red phosphors tend to have “spikey” spectral distributions. As a consequence the full range of human color perception cannot be reproduced.

According to Hunt, there are 6 different types of color reproduction:

1. *Spectral*: Spectral reflectance or transmittance of the image matches the original exactly. This is not possible in normal reproductions since the color gamut does not encompass the entire “horseshoe”.

2. *Colorimetric*: Reproduction matches the original in chromaticity and relative luminance, i.e., the original and reproduced colors are metamers. Of course, whether a reproduction is colorimetric varies with the observer.

3. *Exact*: In addition to colorimetric reproduction we have equality of absolute luminances, i.e., the appearance of colors does not depend on the illuminant intensity.

4. *Equivalent*: Chromaticities, relative luminance, and absolute luminance are adjusted to achieve equality of appearance, i.e., we assume the viewing conditions and adjust by making the reproduction colorimetrically incorrect but perceptually correct.

5. *Corresponding*: Similar to “equivalent” but does not require that absolute luminances match, i.e., adjust chromaticity and relative luminance so as to achieve equality of appearance as if the original was lit by the reproduction illuminant.

6. *Preferred*: Issue with Caucasians who don’t like their skin tone.

15.8.5 Color Spaces

Red, Green, Blue

This representation is usually graphed in a Cartesian system analogous to \([x, y, z]\), as shown:
RGB coordinates (8 bit) displayed as a Cartesian system. Locations with the same value in each coordinate are “neutral”, e.g., [0,0,0] is “black”, [255,255,255] is “white”, and others are “gray”. Pure colors appear along the axes.

Hue, Saturation, Lightness (or Brightness, or Value):

This is the representation that led to Young’s theory:

- Hue corresponds to the common definition of color, e.g., “red”, “orange”, “violet” etc., specified by the dominant wavelength in a spectrum distribution, though a “dominant” may not actually be present

- Saturation (also called chroma): an expression for the “strength” or “purity” of a color. The intensity of a very saturated color is concentrated near the dominant wavelength. Looked at another way, saturation is measured as the amount of “gray” in proportion to the hue. All colors formed from one or two primaries have 100% saturation; if some amount of the third primary is added to a color formed from the other two, then there is at least some “gray” in the color and the saturation decreases. The saturation of a pure white, gray, or black scene (equal amounts of all three primaries) is zero. A mixture of a purely saturated color (e.g., “red”) and white produces a “desaturated red”, or
“pink”. Saturation is reduced if surface reflections are present.

\[
\begin{align*}
R &= 0 & H &= 116 \\
G &= 204 & S &= 255 & S = \frac{255}{255} = 100\% \\
B &= 153 & L &= 102 \\
R &= 25 & H &= 115 \\
G &= 204 = 25 + 179 & S &= 199 & S = \frac{199}{255} = 78\% \\
B &= 153 & L &= 115 \\
S &= \frac{\text{hue} - \text{gray}}{\text{hue}} &= \frac{115 - 25}{115} = 78\%
\end{align*}
\]

In more scientific terms, the saturation is the relative bandwidth of the visible output from a light source. A source with a large saturation has a narrow bandwidth, and vice versa. As saturation increases, colors appear more “pure.” As saturation decreases, colors appear more “washed-out”.

- **Brightness**: sensation of intensity of a light, from dark through dim to bright.
- **Lightness**: a relative expression of the intensity of the energy output reflected by a surface; “blackness”, “grayness”, “whiteness” of a visible light source. It can be expressed as a total energy value or as the amplitude at the wavelength where the intensity is greatest.

HSB is often represented in a cylindrical coordinate system analogous to \((r, \theta, z)\). The saturation coordinate is plotted along the radial axis, the hue as the azimuthal coordinate, and the lightness as the vertical \((z)\) axis. The hue determines the frequency of light, the position in the spectrum, or the relative amounts of red, green, and blue. It is a continuous and periodic scale that often is measured in an “angle” in angular degrees (e.g., the “hue angle”), though it also may be normalized to be compatible with 8-bit numerical representations. Hues located at the extrema (e.g., angles of \(\pm 180^\circ\)) are identical, as shown in the figure taken from the hue adjustment in Adobe Photoshop™. A pure hue is 50% luminosity, 100% saturation. The hue angles are shown, where red corresponds to an angle of \(0^\circ\).

![Hue representation](image)

*The hue representation used in Adobe Photoshop™. The hue at angle 0° is “red”*
To illustrate the continuity of the hue circle, it may be rotated about the azimuth and to show how it “wraps around” at cyan \( (\theta = 180^\circ) \), the complementary color to red.

The hue axis after rotation by 180°, showing the “wraparound” at the edge of the axis.

Hue is represented by 8-bit integer values in other applications, such as Powerpoint™. A list of the primary colors for different hue angles is shown in the table. Note that the additive primaries are located at 0° and ±120°, while the subtractive primaries are at ±60° and 180°. Colors at opposite sides of the hue circle (separated by 180°) are complementary, so that the sum of two complementary colors produces white.

The sum of monochromatic yellow (\( \lambda = 580 \text{ nm} \)) and monochromatic blue (\( \lambda = 480 \text{ nm} \)) produces white light that looks just as white as the sum of all visible wavelengths.

<table>
<thead>
<tr>
<th>Color</th>
<th>Photoshop™ Hue Angle ( \theta ) (°)</th>
<th>Powerpoint™ Hue ([0, 255])</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyan</td>
<td>±180°</td>
<td>127 ( \equiv \frac{1}{2} )</td>
</tr>
<tr>
<td>green</td>
<td>+120°</td>
<td>85 ( \equiv \frac{1}{3} )</td>
</tr>
<tr>
<td>yellow</td>
<td>+60°</td>
<td>42 ( \equiv \frac{1}{6} )</td>
</tr>
<tr>
<td>red</td>
<td>0°</td>
<td>0</td>
</tr>
<tr>
<td>magenta</td>
<td>−60°</td>
<td>213 ( \equiv \frac{5}{6} )</td>
</tr>
<tr>
<td>blue</td>
<td>−120°</td>
<td>170 ( \equiv \frac{2}{3} )</td>
</tr>
</tbody>
</table>
## 15.8 COLOR VISION

### Color Vision Diagram

- **Lightness (Value, Intensity):**
  - White
  - Gray
  - Black

- **Saturation:**

- **Hue Angle:**

---

### Color Values Table

<table>
<thead>
<tr>
<th>Color</th>
<th>R</th>
<th>G</th>
<th>B</th>
<th>H</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyan</td>
<td>0</td>
<td>255</td>
<td>255</td>
<td>127</td>
<td>255</td>
<td>128</td>
</tr>
<tr>
<td>green</td>
<td>0</td>
<td>255</td>
<td>0</td>
<td>85</td>
<td>255</td>
<td>128</td>
</tr>
<tr>
<td>yellow</td>
<td>255</td>
<td>255</td>
<td>0</td>
<td>42</td>
<td>255</td>
<td>128</td>
</tr>
<tr>
<td>red</td>
<td>255</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>255</td>
<td>128</td>
</tr>
<tr>
<td>magenta</td>
<td>255</td>
<td>0</td>
<td>255</td>
<td>213</td>
<td>255</td>
<td>128</td>
</tr>
<tr>
<td>blue</td>
<td>0</td>
<td>0</td>
<td>255</td>
<td>170</td>
<td>255</td>
<td>128</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Color</th>
<th>R</th>
<th>G</th>
<th>B</th>
<th>H</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>yellow</td>
<td>255</td>
<td>255</td>
<td>0</td>
<td>42</td>
<td>255</td>
<td>128</td>
</tr>
<tr>
<td>yellow</td>
<td>131</td>
<td>128</td>
<td>0</td>
<td>42</td>
<td>255</td>
<td>64</td>
</tr>
<tr>
<td>yellow</td>
<td>129</td>
<td>255</td>
<td>255</td>
<td>42</td>
<td>255</td>
<td>192</td>
</tr>
<tr>
<td>yellow</td>
<td>246</td>
<td>240</td>
<td>0</td>
<td>42</td>
<td>255</td>
<td>120</td>
</tr>
<tr>
<td>yellow</td>
<td>195</td>
<td>192</td>
<td>64</td>
<td>42</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>gray</td>
<td>128</td>
<td>128</td>
<td>128</td>
<td>42</td>
<td>0</td>
<td>128</td>
</tr>
</tbody>
</table>
Microsoft Windows color dialogs also use HSB but call the third dimension “luminosity” or “lightness”. It ranges from 0% (black) to 100% (white).

The RGB model is quite simple, so it is natural to consider the advantages of the HSL color model:

1. You can generate grey scales using only one parameter – the luminosity (set saturation to 0).
2. You can vary the color without changing the brightness – vary the hue alone.
3. You can fade or darken several colors, or whole bitmaps, such that the lightness (or darkness) stay in step.

The HSL model is easier to use visually because it suits the eye, whereas the RGB model is easier to use in programming.

15.8.6 Conversion from RGB to HSL

Recall the transformations between Cartesian coordinates \([x, y, z]\) and cylindrical coordinates \((r, \theta, z)\):

\[
\begin{align*}
  r &= \sqrt{x^2 + y^2} \quad x = r \cos \theta \\
  \theta &= \tan^{-1} \left( \frac{y}{x} \right) \quad y = r \sin \theta \\
  z &= z 
\end{align*}
\]

The transformation from Cartesian to cylindrical coordinates is nonlinear and thus cannot be written as a matrix-vector product. The scheme for computing HSL from RGB is:

1. Normalize the three values \([R, G, B]\) to \([0, 1]\)
2. Find the maximum and minimum of the three values; these are \(color_{\text{max}}\) and \(color_{\text{min}}\)
3. If all three RGB values are identical, then the hue and saturation are both 0
4. Compute the lightness \(L\) via

\[
L = \frac{color_{\text{max}} + color_{\text{min}}}{2}
\]

5. Test \(L\):

\[
\begin{align*}
  \text{If } L < 0.5 \text{ then } S &= \frac{color_{\text{max}} - color_{\text{min}}}{color_{\text{max}} + color_{\text{min}}} \\
  \text{If } L < 0.5 \text{ then } S &= \frac{color_{\text{max}} - color_{\text{min}}}{2 - (color_{\text{max}} + color_{\text{min}})}
\end{align*}
\]
6. Compute hue $H$ via:

(a) If $\text{color}_{\text{max}} = R$ then

$$H = \frac{G - B}{\text{color}_{\text{max}} - \text{color}_{\text{min}}}$$

(b) If $\text{color}_{\text{max}} = G$ then

$$H = 2 + \frac{B - R}{\text{color}_{\text{max}} - \text{color}_{\text{min}}}$$

(c) If $\text{color}_{\text{max}} = B$ then

$$H = 4 + \frac{R - G}{\text{color}_{\text{max}} - \text{color}_{\text{min}}}$$

7. Convert $L$ and $S$ back to percentages, and $H$ into an angle in degrees (i.e., scale it from 0-360).

Example:

$$R = 25 \rightarrow \frac{25}{255} \approx 0.098$$

$$G = 204 \rightarrow \frac{204}{255} = 0.8$$

$$B = 53 \rightarrow \frac{53}{255} \approx 0.208$$

$$\text{color}_{\text{max}} = \frac{G}{255} = 0.800$$

$$\text{color}_{\text{min}} = \frac{R}{255} = 0.098$$

$$L = \frac{0.8 + 0.098}{2} = 0.449 < 0.5$$

$$L = 0.449 \cdot 255 = 115$$

$$S = \frac{0.8 - 0.098}{0.8 + 0.098} \cdot 255 = 0.782 \cdot 255 = 199$$

$$H = 2 + \frac{0.208 - 0.098}{0.8 - 0.098}$$

$$\approx 2.157 \text{ radians} \approx +123.6^\circ$$
Conversion from HSL to RGB

1. If $S = 0$, define $L = R = G = B$, otherwise, test $L$:
   
   - If $L < 0.5$, then $\alpha = L \cdot (S + 1)$
   - If $L \geq 0.5$, then $\alpha = L + S - L \cdot S$

2. Set
   
   $\beta = 2.0 \cdot L - \alpha$

3. Normalize hue angle $H$ to the range $[0, 1]$ by dividing by $360^\circ$

4. For each of $R$, $G$, $B$, compute $\gamma$ as follows:
   
   - for $R$, $\gamma = H + \frac{1}{3}$
   - for $G$, $\gamma = H$
   - for $B$, $\gamma = H - \frac{1}{3}$

5. If $\gamma < 0$, then $\gamma = 1 + \gamma$

6. For each of $[R, G, B]$, do the following test:
   
   - If $\gamma < \frac{1}{6}$, then $\text{color} = \beta + (\alpha - \beta) \cdot 6 \cdot \gamma$
   - If $\frac{1}{6} \leq \gamma < \frac{1}{2}$, then $\text{color} = \alpha$
   - If $\frac{1}{2} \leq \gamma < \frac{2}{3}$, then $\text{color} = \beta + (\alpha - \beta) \cdot (4 - 6\gamma)$

7. Scale back to the range $[0, 100]$

8. Repeat steps 6 and 7 for the other two colors.

Other Descriptors of Colors

- Hue: a “pure” color, i.e., one containing no black or white.
- Shade: a “dark” color, i.e., one produced by mixing a hue with black
- Tint: a “light” color, i.e., one produced by mixing a hue with white
- Tone: color produced by mixing a hue with a shade of grey.
15.8.7 Phenomena of Color Vision

Afterimages

This is a well-known optical phenomenon due to the photochemical sensitivity of the eye. If exposed to a stationary scene for an extended period of time, the photochemical dyes that are sensitive to that color become “depleted.” If exposed to a neutral white background, the eye “sees” the complement to that color, because only the dyes sensitive to that color are available. The phenomenon often is illustrated by viewing a representation of the American flag rendered in the colors that lie opposite on the hue angle (cyan for red, yellow for blue, and black for white). Afterimages may be created in Adobe Photoshop™ by using the “inverse” option that cascades a rotation of the hue by 180° and an “inversion” (complementing of) the lightness.

Constructing a specimen for demonstrating the “afterimage:” rotate the hue by 180° and complement the lightness, or just use the “inverse” operation in Adobe Photoshop™. Test it by staring at the processed image for 30+ seconds and then look at a blank sheet of white paper.

15.9 Pixel Processing of Multiple Images

We now return to consider pixel processing of multiband images, that may include images differing in color, time, or other parameter. The gray value of each pixel in the output image is determined from the gray values of the corresponding input pixel in the various bands. For example we can decompose a color or monochromatic luminance (lightness) image from the gray values of the image in each of the three additive primary colors via:
\[ g(x, y) = \alpha f(x, y, \lambda_r) + \beta f(x, y, \lambda_g) + \gamma f(x, y, \lambda_b) \]
\[ \equiv \alpha f_R(x, y) + \beta f_G(x, y) + \gamma f_B(x, y), \]

where the coefficients \([\alpha, \beta, \gamma]\) are functions of the spectral filtration and sensitivity of the recording process. The gray values of the same pixel in the three images are generally different but correlated. For example, a red object will be bright in the red image and darker in the green and blue. The decomposition of a crude color image into its component \(RGB\) images is shown below; note that the histograms do not exhibit easily distinguishable clusters and thus it will be difficult to segment the objects from the image effectively.

Simple 3-band color image, the individual monochromatic images, and their histograms. Note that the gray values of the “tree” are approximately equal to those of “sky” in the red band and approximately equal to those of both “sky” and “grass” in the green band. For this reason, the tree is difficult to segment from a single color channel.

The three 1-D histograms do not exhibit distinct pixel clusters for each of the five object colors (red house, green tree, pale green grass, blue sky, and white clouds and door). In other words, the clusters of pixels belonging to these five classes overlap. This observation forces the conclusion that the five objects cannot be segmented successfully from a single band. That said, we can observe that the spectral reflectance
of pixels belonging to the “house” is significantly different from the other objects, and those pixels may be segmented from the other objects fairly easily, (e.g., by a simple threshold in the blue image, as shown after selecting a threshold level of 112). However, the reflectance of the pixels belonging to the tree in the blue image is only slightly different from those belonging to sky. Attempts to segment pixels belonging to the “tree” from the blue image are shown, which lead to either a “noisy” segmentation, or the misidentification of the “tree” and the “grass.”

Two attempts to segment “tree” from blue image alone. The first attempt sets all pixels to black that lie in the range $112 \leq b \leq 140$; this produces a very noisy “tree.” Pixels in the second image are thresholded to black if they lie in the range $112 \leq b \leq 167$. The segmented pixels include the “grass.”

15.10 Multispectral Histograms for Feature Extraction

As already described, visible-light color images are represented by triads of monochrome images. The decomposition of color images into $RGB$ bands is very common. Within the neural network of the eye, the three cone signals are weighted and combined to generate the three channels that are transmitted to the brain. Roughly speaking, one channel corresponds to the lightness or luminance of the scene (i.e., the black-and-white video signal), and is a weighted sum (integral) of $S,M$, and $L$. This
information is transmitted to the brain with full resolution. The other two transmitted signals are weighted differences (derivatives) of the $S$, $M$, and $L$ and describe the chrominance of the scene; these are transmitted with reduced resolution, thus preserving information deemed more important during evolutionary development. Broadcast transmission of color video signals is roughly modeled on the weighted summing of cone signals for transmission to the brain.

In digital imaging, the three raw cone signals generated by the human visual system can be represented as three digital images: roughly the brightness in blue, green, and red light. The weighted summing of the cone signals for transmission to the brain may be modeled as a linear operation applied to the 3-element vector $[R, G, B]$. The operation is implemented by an invertible $3 \times 3$ matrix. The requirement that the matrix be invertible ensures that the 3-element output vector may be computed from the output vector. However, note that if the output values are quantized, as required before subsequent digital processing, then the cascade of forward and inverse transformations may not yield the identical triplet of $[R, G, B]$.

The variations of colors of the objects in an image may be used to segment a multispectral image $f[n, m, \lambda_i]$. To successfully segment the “tree” pixels from the color image, it is necessary to use the multispectral information simultaneously. One way is to use a multidimensional (2-D or 3-D) histogram generated from two or three of the gray values at each pixel. Often only two colors are used to generate a 2-D histogram because of the difficulty of displaying three dimensions of information by conventional means. For example, pixels having a particular gray level $f_R$ in the red image and a level $f_G$ in the green image are counted in bin $[f_R, f_G]$. The resulting matrix is the image of a 2-D feature space, i.e., the bin with the largest number of pixels can be displayed as white and unpopulated bins as black. The histogram can be used for image segmentation as before, but the extra information obtained from the second color usually ensures better results.
15.10 MULTISPECTRAL HISTOGRAMS FOR FEATURE EXTRACTION

15.10.1 2-D Histograms of Simple Color Image

The three 2-D histograms of each pair of channels of the three color bands. You should be able to identify the objects in each of the histograms. Four clusters can be identified in the R-G and five in the G-B histograms.

These 2-D histograms (also called “scatterplots”) of pairs of grayscale images were plotted using the Hypercube Image Analysis Software (available free for many platforms from http://www.tec.army.mil/Hypercube/).

The cluster of pixels with large gray values in all images due to the white pixels in the original are easily identified in the three 2-D histograms. You should be able to identify the clusters that belong to the house, tree, sky, etc. We can segment the image by thresholding those pixels within certain intervals of red, green, and blue (to white, say), and thresholding the others to black.

Note that histogram concept can be extended easily to more than three dimensions, though visual representation is more difficult. This is the basis for multispectral segmentation in many areas of image processing.

Segmentation from 3-D Histogram

It is often easier to identify distinct and well-separated clusters from the multidimensional histogram rather than from the individual images. The segmented image of the tree obtained from the 3-D histogram is:
Segmentation of “tree” pixels directly from the 3-D histogram of the RGB image. The black pixels satisfy the three constraints: \( r \geq 106, g \geq 195, \) and \( b \leq 167. \)

15.10.2 Principal Component Analysis – PCA

Reference: Schowengerdt, *Remote Sensing*

The gray values of a pixels in the different bands of a multispectral image (e.g., an RGB color image) often are highly correlated, meaning that some (or many) bands may be visually similar. In other words, the information content of a multispectral image often is quite “redundant.” It often is convenient to reduce or even eliminate this redundancy by expressing the image in terms of “axes” other than the original “bands.” The data image is transformed to a new coordinate system by rigidly rotating axes to align with these other “directions” and the image data then projected onto these new axes. In principal components, the rotation produces a new multispectral image with a diagonal covariance matrix, so that there is no covariance between the various bands. One axis of the rotated coordinate system is aligned with the direction of maximum variance of the image, the second axis is perpendicular to the first and aligned with the direction of the second largest variance, etc. The bands in the principal component image are thus arranged in order from largest to smallest variance.

To illustrate, consider the principal components of a 2-band image created from the blue and green bands of the simple color image; the values in the red band were replaced with zeros. The 2-D histogram, blue vs. green, is shown to locate the (approximate) axes of the principal components, that were evaluated using “Hypercube.” Since the third component image was black, only two principal components are needed to fully represent the image. In the outputs, note that the “lightest” objects in the image (the white clouds and door) also are lightest in the first PC. The darkest structure in both channels is the house, which appears “black” in the 1st PC. The gray values are projected onto the orthogonal axis in the second PC image. The red house pixels and the white pixels belonging to the clouds and door are projected to the same “mid-gray” value in the second principal component, and thus are indistinguishable in that band (the white door has “disappeared into the dark house”).
Principal components of a two-band image, computed in “Hypercube”: the image is created by inserting zeros in the red band. The 2-D histogram blue vs. green is shown. The first principal component projects the gray values onto the “cyan” axis that includes most of the image variance. The second principal component projects the data onto the magenta axis. The images are shown. Note that the first principal component shows the clouds and door as white and the house as black. The door is not visible in the second principal component, as its gray value is the same as that of the house.

The three PCs of the complete RGB image also are shown along with the three eigenvalues. The first PC of the 3-D scene has a very dark “sky” because it exhibits the largest contrast relative to the white objects (clouds and door). The 3rd PC shows the smallest range of variance, which is dominated by image noise.
The three principal components of the RGB house + tree image. Note that the third component image is quite noisy.

15.11 Time-Sequence Images: Video

The other common example a 3-D image plots time on the third image axis. Perhaps the most obvious use is in motion pictures, where the different frames of the movie are the time samples of the 3-D spatial image \( f [x, y, t] \). The illusion of continuous motion is created because of the photochemical response of the rods and cones in the retina. The time duration of the process is the source of the phenomenon of “persistence of vision.” The persistence is shorter if image is brighter (movie projectors have rotary shutters that show each frame two or even three times). Movies were originally taken at 16 frames per second, later increased to 24 fps in US. This is the reason why motion in old movies is too fast. The frame rate in Europe \( ) \) is 25 fps, related to the AC frequency of rate is 50 Hz. For this reason, American movies shown in Europe finish more quickly than they do in the US.

The second most familiar example of time-sampled imagery is video, where the 3-D scene \( f [x, y, t] \) is sampled in both time and space to convert the scene to a 1-D function of time \( s [t] \). It took 50 or more years to develop the hardware to scan scenes. The first systems were mechanical, based either on a system that scanned the light reflected from an illuminated scene, or an illumination system that scanned a beam of light over the scene and collected the reflected light. One of the primary developers of mechanically scanned video was the Scotsman John Logie Baird. The hardware of electronic scanning was developed through the efforts of such “illuminaries” as the American Philo T. Farnsworth, who demonstrated a working video system in the 1920s.

Video systems commonly use an “interlaced scan” that alternates scans of the even and odd lines of the full frame, so that the eye is presented with half of the image information every 1/60 s. This is less objectionable to the eye than the original “progressive-scanning” systems that presented a full frame every 1/30 s.

Note that lines number 248 to 263 and 511 to 525 are typically blanked to provide time for the beam to return to the upper left hand corner for the next scan; other signals (such as closed captioning or the second audio program) are transmitted during this “flyback” time.
15.11 TIME-SEQUENCE IMAGES: VIDEO

Figure 15.1: Interlaced format of NTSC video raster: one frame in 1/30 second is composed of two fields, each taking 1/60 second. The first “field” includes 262.5 odd lines and the second “field” has 262.5 even lines. Lines 248-263 in the first field and lines 511-525 in the second field are “blanked” — this is the “retrace” time for the CRT beam and is the time when additional information (e.g., closed captioning) is transmitted.

15.11.1 Color-Space Transformations for Video Compression

References:
Falk, Brill, Stork, Seeing the Light, §10
Glassner, Principles of Digital Image Synthesis, §1

Particular color transformations have been developed for use in many different applications, including various schemes for image transmission. Consider the transformation used in the color video standard in North America, that was developed by the National Television Standards Committee (NTSC), which converts RGB values to three other channels by a linear transformation: the “luminance” $Y$ (used by “black-and-white” receivers), and two “chrominance” channels $I$ and $Q$ via:

\[
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
0.596 & -0.274 & -0.322 \\
0.211 & -0.523 & 0.312
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
=
\begin{bmatrix}
Y \\
I \\
Q
\end{bmatrix}
\]

Note that the sum of the weights applied to the luminance channel $Y$ is $0.299 + 0.587 + 0.114 = 1.0$, while the sums of the weights of the two chrominance channels are both 0. In other words, the luminance channel is a weighted sum of RGB (analogous to an integral), while the chrominance channels are weighted differences (similar to derivatives). In the context of linear systems, $Y$ is the result of “spectral lowpass filtering,” while $I$ and $Q$ are the outputs of what may be loosely described as “spectral highpass filters.”

If the input $R$, $G$, and $B$ are in the range $[0, 255]$, so will be the range of $Y$. Both chrominance channels of any gray input pixel (where $R = G = B$) is zero, and the range of allowed chrominances is bipolar and fills the available dynamic range,
e.g., \(-103 \leq I \leq 152, +133 \leq Q \leq -122\), totaling 256 levels for 8-bit RGB inputs. The positive polarity of \(I\) isreddish (often described as “orange”), and its negative polarity is “green+blue” or cyan; hence the \(I\) information is sometimes called the “orange-cyan” axis. The positive polarity of \(Q\) is “red+blue” or purple, and the negative polarity is green, so the \(Q\) information is the “purple-green” axis.

The transformation of a “mid-gray” pixel where the red, green, and blue images are identically \(\alpha\) is:

\[
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
0.596 & -0.274 & -0.322 \\
0.211 & -0.523 & 0.312
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\alpha \\
\alpha
\end{bmatrix}
= 
\begin{bmatrix}
\alpha \\
0 \\
0
\end{bmatrix}
\]

so that the luminance is the gray value of the three colors and the two chrominance channels are both zero.

The transformation from \(YIQ\) back to \(RGB\) is the inverse of the forward matrix operator:

\[
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
0.596 & -0.274 & -0.322 \\
0.211 & -0.523 & 0.312
\end{bmatrix}^{-1}
\approx
\begin{bmatrix}
1.0 & 0.956 & 0.621 \\
1.0 & -0.273 & -0.647 \\
1.0 & -1.104 & 1.701
\end{bmatrix}
(\text{rounded})
\]

\[
\begin{bmatrix}
1.0 & 0.956 & 0.621 \\
1.0 & -0.273 & -0.647 \\
1.0 & -1.104 & 1.701
\end{bmatrix}
\begin{bmatrix}
Y \\
I \\
Q
\end{bmatrix}
= 
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

Note that the \(R\), \(G\), and \(B\) channels all include 100% of the luminance channel \(Y\).

Other color transformations are used in other video systems. The two common systems used in Europe and Asia are \(PAL\) (“Phase Alternation by Line”) and \(SECAM\) (“Systeme Electronique Couleur Avec Memoire”). Each broadcasts 625 lines at 25 frames per second and uses the \(YUV\) triad of luminance and chrominance, where the luminance is the same combination of \(RGB\) but the chrominance channels are slightly different. The \(RGB\) transformation to \(YUV\) is:

\[
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
-0.148 & -0.289 & 0.437 \\
0.615 & -0.515 & -0.100
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
= 
\begin{bmatrix}
Y \\
U \\
V
\end{bmatrix}
\]

The luminance calculations for \(YIQ\) and \(YUV\) are identical. The ranges of allowed chrominances are bipolar: for 8-bit \(RGB\) inputs \(-144 \leq U \leq 111\) and \(-98 \leq V \leq 157\) (each with 256 levels). The NTSC selected the \(YIQ\) standard over \(YUV\) because tests indicated that more \(I-Q\) data could be discarded without affecting the subjective
image quality.

Color image decomposition, all images displayed with full dynamic range (i.e., maximum possible gray value is mapped to white and minimum to black): first row RGB; second row YIQ; third row YUV.

Obviously there exists an infinite number of invertible $3 \times 3$ matrix transformations, and thus of invertible color transformations. The 3-D histograms of a particular input image before and after transformation generally will differ. Therefore segmentation of objects with particular similar colors likely will be easier or more successful in a particular color space.
15.11.2 Multiple-Frame Averaging

Consider a series of video or movie images of an invariant (i.e., stationary) object \( f[x,y] \) corrupted by additive noise that changes from pixel to pixel and frame to frame:

\[
g[x,y,t_i] = f[x,y] + n[x,y,t_i]
\]

where \( n[x,y,t_i] \) is a random number selected from a Gaussian distribution with \( \mu = 0 \). The additive noise will tend to obscure the “true” image structure \( f[x,y] \). One common problem in image processing is to enhance the visibility of the invariant objects in the image by attenuating the noise. If the gray values \( n \) are truly random, i.e., all values of \( n \) in the range \((-\infty, \infty)\) can exist with equal likelihood, then little can be done to improve the image. Fortunately, in realistic imaging problems the probability of each value of \( n \) is determined by some probability density function (histogram) \( p[n] \) and we say that the noise is stochastic. The most common probability distributions in physical or imaging problems are the uniform, Poisson, and normal. Less common, but still physically realizable, distributions are the Boltzmann (negative exponential) and Lorentzian. A general discussion of stochastic functions is beyond the scope of the immediate discussion, though we will go into more detail later while reviewing statistical filters. Interested readers should consult Frieden’s book *Probability, Statistical Optics, and Data Testing* (Springer-Verlag, 1991) for detailed discussions of physically important stochastic processes. At this point, we will state without proof that the central limit theorem determines that the most common probability density function in physical problems is the normal distribution \( N[\mu, \sigma^2] \). The histogram of noise gray values in the normal distribution is:

\[
p[n] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(n-\mu)^2}{2\sigma^2}\right] \equiv N[\mu, \sigma^2]
\]

The normal distribution is completely characterized by two parameters: the mean value \( \mu \) and the variance \( \sigma^2 \) (or its equivalent, the standard deviation \( \sigma \)). This is the equation for the familiar bell curve with maximum value located at \( n = \mu \) and with full width of \( 2\sigma \) measured at approximately 20% of the maximum amplitude. In the special case where the mean value \( \mu = 0 \), the normal distribution commonly is called a Gaussian distribution. The remainder of the discussion in this section will assume that the additive noise has been selected at random from a normal distribution.

It is probably clear intuitively that an image created by averaging a collection of noise images \( n[x,y,t_i] \) over time will tend toward a uniform image whose gray level is the mean \( \mu \) of the noise, i.e., the variations in gray level about the mean will “cancel out”:

\[
\frac{1}{N} \sum_{i=1}^{N-1} n[x,y,t_i] \cong \mu \cdot 1[x,y]
\]

If the sequence of input images includes an invariant object on a background of additive normal noise, the visibility of the object will be enhanced in the average
15.11 TIME-SEQUENCE IMAGES: VIDEO

image:

\[
\frac{1}{N} \sum_{i=1}^{N-1} (f[x,y] + n[x,y,t_i]) = \frac{1}{N} \sum_{i=1}^{N-1} f[x,y] + \frac{1}{N} \sum_{i=1}^{N-1} n[x,y,t_i] \\
\cong f[x,y] + \mu \cdot 1[x,y]
\]

This result is proven directly below, but may be seen more easily if the reader is familiar with the statistical concept that the probability density function of the sum of \(N\) random variables is the \(N\)-fold convolution of the individual probability density functions.

To quantify the visibility of the object in a noisy image, it is necessary to quantify the visibility of the noise, i.e., the variability of gray level due to the stochastic signal. The average gray value due to noise is:

\[
\langle n[x,y] \rangle = \int_{-\infty}^{+\infty} n[x,y] \ p[n] \ dn = \mu
\]

where \(\langle \ \rangle\) denotes the averaged value of the quantity and \(p[n]\) is the probability density function of the noise. The mean value obviously is not an appropriate measure because it does not describe the variability of gray value. A useful quantity is the variance of the noise which describes the average of the difference between the square of the gray value due to the noise and the mean:

\[
\sigma^2[n] = \langle (n - \mu)^2 \rangle = \int_{-\infty}^{+\infty} (n - \mu)^2 \ p[n] \ dn \\
= \int_{-\infty}^{+\infty} (n^2 - 2n\mu + \mu^2) \ p[n] \ dn \\
= \int_{-\infty}^{+\infty} n^2 \ p[n] \ dn - 2\mu \int_{-\infty}^{+\infty} n \ p[n] \ dn + \mu^2 \int_{-\infty}^{+\infty} p[n] \ dn \\
= \int_{-\infty}^{+\infty} n^2 \ p[n] \ dn - 2\mu \cdot \mu + \mu^2 \cdot 1 \\
= \langle n^2 \rangle - 2\mu^2 + \mu^2 \\
\sigma^2[n] = \langle n^2 \rangle - \mu^2
\]

A measure of the relative visibility of the signal and the noise is the ratio of the signal to the noise variance, and is often called the signal-to-noise ratio (\(S/N\) or \(SNR\)). It may be expressed as power or amplitude:

\[
\text{Power } SNR \equiv \frac{f^2[x,y]}{\sigma^2[n]} \\
\text{Amplitude } SNR \equiv \sqrt{\frac{f^2[x,y]}{\sigma^2[n]}} = \frac{f[x,y]}{\sigma[n]}
\]
After averaging $P$ frames of signal plus noise, the gray level at the pixel $[x, y]$ will approach:

$$g[x, y] = \frac{1}{P} \sum_{i=1}^{P} (f[x, y] + n[x, y, t_i])$$

$$= \sum_{i=1}^{P} \frac{f[x, y]}{P} + \sum_{i=1}^{N} \frac{n[x, y, t_i]}{P}$$

$$\approx f[x, y] + \mu$$

The variance of the noise at pixel $[x, y]$ after averaging $P$ frames is:

$$\sigma^2[n] = \langle n^2[x, y] \rangle - \mu^2 = \sum_{i=1}^{P} \left( \frac{n[x, y, t_i]}{P} \right)^2 - \mu^2$$

$$= \frac{1}{P^2} \sum_{i=1}^{P} n[x, y, t_i] \sum_{j=1}^{P} n[x, y, t_j] - \mu^2$$

$$= \frac{1}{P^2} \sum_{i=1}^{P} (n[x, y, t_i])^2 + \frac{2}{P^2} \sum_{i>j}^{P} (n[x, y, t_i] \cdot n[x, y, t_j]) - \mu^2$$

If the noise values are selected from the same distribution, all the terms $n[x, y, t_i]$ in the first sum on the right are identical:

$$\frac{1}{P^2} \sum_{i=1}^{P} (n[x, y, t_i])^2 = \frac{1}{P^2} \left[ \sum_{i=1}^{P} (\sigma^2_i + \mu^2) \right]$$

$$= \frac{1}{P^2} \left[ P \cdot (\sigma^2_i + \mu^2) \right]$$

$$= \frac{(\sigma^2_i + \mu^2)}{P}$$

Because the noise values are uncorrelated by assumption, the second sum on the right is just the square of the mean:

$$\frac{2}{P^2} \sum_{i>j}^{P} (n[x, y, t_i] \cdot n[x, y, t_j]) = \frac{2}{P^2} \left( \frac{P^2}{2} \cdot \mu \cdot \mu \right) = \mu^2$$
So the variance of the average of $P$ samples of the noise is:

$$\sigma^2[n] = \frac{1}{P^2} \sum_{i=1}^{P} (n[x, y, t_i])^2 + \frac{2}{P^2} \sum_{i>j}^{P} (n[x, y, t_i] \cdot n[x, y, t_j] - \mu^2)$$

$$= \frac{(\sigma_i^2 + \mu^2)}{P} + \mu^2 - \mu^2$$

$$= \frac{\sigma_i^2 + \mu^2}{P}$$

If the mean value of the noise is $\mu = 0$ (Gaussian noise), then the variance of the sum is reduced by a factor of $P$ and standard deviation is reduced by $\sqrt{P}$:

$$\sigma^2[n] = \frac{\sigma_i^2}{P}$$

$$\sigma[n] = \sqrt{\sigma^2[n]} = \frac{1}{\sqrt{P}} \sigma_i[n]$$

The amplitude SNR of the averaged image is:

$$SNR_{out} = \frac{f[x, y]}{\sqrt{\sigma^2[n]}} = \frac{f[x, y]}{\left(\frac{\sigma_i[n]}{\sqrt{P}}\right)} = \sqrt{P} \cdot \frac{f[x, y]}{\sigma_i} = \sqrt{P} \cdot SNR_{in}$$

$$SNR_{out} = \sqrt{P} \cdot SNR_{in}$$

Thus the effect of averaging multiple frames which include uncorrelated additive noise from a Gaussian distribution is to decrease the width of the histogram of the noise by a factor of $\left(\sqrt{P}\right)^{-1}$, which increases the signal-to-noise ratio of the image by $\sqrt{P}$. For example, a video image from a distant TV station often is contaminated by random noise (“snow”). If the image is invariant over time, its signal-to-noise ratio can be increased by averaging; if 90 frames of video ($\approx 3$ sec) are averaged, the SNR of the output image will increase by a factor of $\sqrt{90} \approx 9.5$.

If the noise is correlated to some degree from frame to frame (i.e., its spatial structure is partially correlated from image to image), then averaging will not improve the SNR so rapidly. For example, consider imaging of a submerged object through a water surface. Wave motion will distort the images but there will be some correlation between frames. The SNR might improve only as, say, $P^{1/4} \approx 3$ for $P = 90$ frames.
Averaging of independent noise samples: (a) signal $f[x]$; (b) one realization of Gaussian noise $n_1[x]$ with $\mu = 0$ and $\sigma = 1$; (c) $f[x] + n_1[x]$; (d) $\frac{1}{9} \sum_{i=1}^{9} (f[x] + n_i[x])$, showing the improved signal-to-noise ratio of the averaged image.

15.11.3 Required Number of Bits for image Sums, Averages, and Differences

If two 8-bit images are added, then the gray value of the output image lies in the interval $[0, 510]$, with 511 gray values requiring 9 bits of data to represent fully. If constrained to 8 useful bits of data in the output image, half of the gray-scale variations data in the summation must be discarded. In short, it is necessary to requantize the summation image. If two 8-bit images are averaged, then the gray values of the resulting image are in the interval $[0, 255]$, but half-integer values are virtually guaranteed, thus ensuring that the average image also has 9 bits of data unless requantized. The central limit theorem indicates that the histogram of a summation or average image should approach a Gaussian form.
15.12 Image Subtraction for Change Detection

Subtracting images of the same scene recorded at different times will highlight pixels whose gray value has changed in the interval:

\[ g[x, y] = f[x, y, t_1] - f[x, y, t_0]. \]

Invariant pixels will subtract to 0, pixels that have \{brightened\} will have \{positive\} gray level. This technique is applied to motion/change detection and may be interpreted as a discrete version of the time derivative.

\[ \frac{\partial f[x, y, t]}{\partial t} = \lim_{\Delta t \to 0} \frac{f[x, y, t + \Delta t] - f[x, y, t]}{\Delta t} \]

For multitemporal digital images, the smallest nonvanishing time interval \(\Delta t\) is the interval between frames \((\Delta t = t_1 - t_0)\). The time derivative is the difference image of adjacent frames:

\[ \frac{\partial f[x, y, t]}{\partial t} \to \partial_t \{f[x, y, t]\} = \frac{f[x, y, t_0 + (t_1 - t_0)] - f[x, y, t_0]}{t_1 - t_0} = \frac{f[x, y, t_1] - f[x, y, t_0]}{t_1 - t_0}. \]

Note that the difference image is bipolar, white > 0, gray = 0, black < 0. The difference image \(g[x, y]\) is bipolar and must be scaled to fit the available discrete dynamic range \([0, 255]\) for display. Often \(g_0 = 0\) is mapped to the mid-level gray (e.g., 127), the maximum negative level is mapped to 0, the brightest level to 255, and the intervening grays are linearly compressed to fit.

\[ f[x, y, t_1] \quad f[x, y, t_2] \quad f[x, y, t_2] - f[x, y, t_1] \]

\[ t_2 \approx t_1 + 3 \text{ sec.} \]

Extrema \(\Rightarrow\) translations

Difference of two images of the same scene taken approximately 3 seconds apart. The images are from a movie taken by James Noel during the Mallory Everest Expedition of 1924. The one image has been translated by a small distance, thus
creating “edge” regions. Differences of zero are mapped to “midgray”, while regions where the second image is brighter or darker map to white and black, respectively.

In this example, the difference image may be used to help determine how well registered the two images are; if perfectly registered, only pixels that have really changed between images will have gray values other than 0.

### 15.13 Difference Images as Features

In feature extraction and recognition applications (e.g. remote sensing), linear combinations (i.e. sums and differences) of multispectral imagery may supply additional useful information for classification. For example, the difference image of spectral band 3 (red) and 4 (infrared) (out of a total of seven bands) imaged by LANDSAT helps classify urban vs. rural features in the image. In the simple house-tree image, the visibility of the house is enhanced in Red-Green and Blue-Red images.

![Green - Red](image1) ![Red - Blue](image2) ![Blue - Green](image3)

*Color difference images: the “house” is more obvious in the green-red image, and the tree in the red-blue image.*

Note that the difference images are noticeably noisier, especially Blue-Green; difference images enhance all variations in gray value, whether desired or not. The concerns about displaying bipolar gray values of time derivatives exist here as well.

#### 15.13.1 Number of Bits in Difference Image

If two 8-bit images (with gray values $f_n$ satisfying the constraint $0 \leq f_n \leq 255$) are subtracted, then the gray values of the resulting image $g[x,y]$ lie in the range $-255 \leq g \leq +255$, for a total of 511 possible values requiring 9 bits of data. To obtain 8 useful bits of data in the output image, half of the data in the difference image must be discarded (usually the least-significant bit, though if the pixels are well correlated, then the most-significant bit may be zero).

The principles of statistical distributions tell us that the difference of images with data generated from two statistical distributions is the convolution of the two distributions centered about the difference in the mean values.
15.14 “Mask” or “Template” Multiplication

Pixel-by-pixel multiplication is an essential part of local neighborhood and global image operations, e.g., the Fourier transform. It also is useful to mask out sections of an image, perhaps to be replaced by objects from other images. This is occasionally useful in segmentation and pattern recognition, but is essential in image synthesis, such as for special effects in movies.

15.15 Image Division

Images recorded by a system with spatially nonuniform response are functions of both the input distribution \( f[x, y] \) and the spatial sensitivity curve \( s[x, y] \), \( 0 \leq s[x, y] \leq 1 \):

\[
g[x, y] = f[x, y] \cdot s[x, y]
\]

This is a deterministic multiplicative degradation of the image; the image may be restored by dividing out the noise. An estimate of the true image brightness can be computed at each pixel by division:

\[
\hat{f}[x, y] = \frac{g[x, y]}{s[x, y]} \approx f[x, y]
\]

(n.b., no image information is recorded at pixels where \( s[x, y] = 0 \) and thus the true value cannot be recovered at those pixels). This technique has been applied to remotely sensed imagery where information about image brightness is critical. Note that errors in the sensitivity function greatly distorts the recovered signal. Similarly, additive noise creates big problems in image division.

15.15.1 Image Division to Correct Spatial Sensitivity

Imaging systems often suffer from a consistent multiplicative error. One very common example is the variation in sensitivity of the pixels in a CCD sensor due to the intrinsic variability in the substrate properties or manufacturing. These multiplicative errors are measured and corrected via a subsequent division.

Consider a biased sine wave \( f[x] \) recorded by an imaging system whose sensitivity falls off away from the origin. The image may be recovered completely if the sensitivity curve is nonzero everywhere and there is no noise in either the recorded signal or the estimate of the sensitivity.
1-D simulation of spatial compensation for sensitivity correction: (a) original signal $f[x]$ is a biased sinusoid; (b) sensitivity function $s[x]$; (c) $g[x] = f[x] \cdot s[x]$; (d) correction $\hat{f}[x] = \frac{g[x]}{s[x]}$

Noise in the estimate of the sensitivity results in distortion of the recovered signal; the effect is more severe where the SNR is low. The deviation of the added noise is 0.005.
15.15 IMAGE DIVISION

Spatial correction in presence of noise: (a) $g[x] + n[x]$, where $n[x]$ is zero-mean Gaussian noise with $\sigma = 0.005$; (b) $\hat{f}[x] = \frac{g[x] + n[x]}{s[x]}$, showing the large errors due to incorrect division of two small values.

15.15.2 Image Division to Enhance Low-Contrast Imagery

Astronomers would often need to examine fine structure (i.e. significant brightness differences) that are hidden by a larger-scale brightness gradient of the object. Included among the significant low-contrast structural features are streamers in the solar corona that radiate outward from the solar surface; the brightness details provide clues about the physical nature of the solar atmosphere and magnetic fields. However, the radial brightness gradient of the corona makes it very difficult to image the full length of the coronal streamers. The overall dynamic range of ground-based imagery of the solar corona is approximately three orders of magnitude (limited by atmospheric sky brightness). Imagery from air- or spacecraft may add an order of magnitude for a total dynamic range of $10^4$. This may be recorded on a wide-range photographic negative ($\Delta D > 4$), but it is not possible to print that dynamic range by normal photographic techniques.

The problem of enhancing the visibility of small changes in coronal image brightness is similar to correction of detector sensitivity just considered and may be attacked by digital methods. The brightness gradient of the corona is analogous to the 1-D sensitivity function $s[x]$: division of the original image by the brightness gradient “equalizes” or “flattens” the gray-level variations in the background, thus allowing the smaller-scale variations across the image to be displayed on a device with limited dynamic range. An estimate of the coronal brightness gradient may be determined by averaging fitted curves of the radial brightness or by making a low-resolution (blurred) image. The latter is more commonly used, as it may be derived via simple local neighborhood digital operations to be considered next. The recovered image is the ratio of the measured high-resolution image and the image of the brightness gradient. This technique may be applied to archival photographic negatives of historical eclipses to provide additional information about the history of solar conditions and thus their effects on the earth’s climate.
Examples of Image Division for Local Contrast Enhancement

Examples of recovered information from old imagery of a solar eclipse and a comet are shown. The pictures are from *Enhancement of Solar Corona and Comet Details* by Matuska, et al., *Proc. SPIE, 119*, pp. 28-35, 1977 and in *Optical Engineering, 17*(6), 661-665, 1978. The images are scans of xerographic prints from microfilm, which is the reason for the poor quality (courtesy of Wallace Memorial Library!)

*Average of four “raw” images of the 1973 solar eclipse, showing the corona with little discernable detail.*

*Image of solar corona obtained by “pseudounsharp masking,” this is the ratio of of the original average of four images image and a lowpass-filtered replica.*

*Image of solar corona obtained by dividing the original image by a lowpass-filtered replica and following with highpass filtering to “sharpen” the fine detail.*
Original image of comet 1957V (Mrkos)

Enhanced image of Comet Mrkos via “pseudounsharp” masking.
Chapter 16

Local Operations

\[ g[x, y] = \mathcal{O}\{f[x \pm \Delta x, y \pm \Delta y]\} \]

In many common image processing operations, the output pixel is a weighted combination of the gray values of pixels in the neighborhood of the input pixel, hence the term local neighborhood operations. The size of the neighborhood and the pixel weights determine the action of the operator. This concept has already been introduced when we considered image prefiltering during the discussion of realistic image sampling. It will now be formalized and will serve as the basis of the discussion of image transformations.

Schematic of a local operation applied to the input image \( f[x, y] \) to create the output image \( g[x, y] \). The local operation weights the gray values in the neighborhood of the input pixel.
16.1 Window Operators — Correlation

You probably have already been exposed to window operations in the course on linear systems. An example of a window operator acting on the 1-D continuous function \( f[x] \) is:

\[
O \{ f[x] \} = g[x] = \int_{-\infty}^{+\infty} f[\alpha] \gamma[\alpha - x] \, d\alpha
\]

The resulting function of \( x \) is the area of the product of two functions of \( \alpha \): the input \( f \) and a second function \( \gamma \) that has been translated (shifted) by the distance \( x \). Different results are obtained by substituting different functions \( \gamma[x] \).

The process may be recast in a different form by defining a new variable of integration \( u \equiv \alpha - x \):

\[
\int_{-\infty}^{+\infty} f[\alpha] \gamma[\alpha - x] \, d\alpha \rightarrow \int_{u=-\infty}^{u=+\infty} f[x + u] \gamma[u] \, du
\]

which differs from the first expression in that the second function \( \gamma[u] \) remains fixed in position and the input function \( f \) is shifted, now by \(-x\). If the amplitude of the function \( \gamma \) is zero outside some interval in this second expression, then the integral need be computed only over the region where \( \gamma[u] \neq 0 \). The region where the function \( \gamma[x] \) is nonzero is called the support of \( \gamma \), and functions that are nonzero over only a finite domain are said to exhibit finite or compact “support.”

The 2-D versions of these expressions are:

\[
O \{ f[x,y] \} = \iint_{-\infty}^{+\infty} f[\alpha,\beta] \gamma[\alpha - x, \beta - y] \, d\alpha \, d\beta
\]

\[
= \iint_{-\infty}^{+\infty} f[x + u, y + v] \gamma[u,v] \, du \, dv.
\]

The analogous process for sampled functions requires that the integral be converted to a discrete summation:

\[
g[n, m] = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} f[i, j] \gamma[i - n, j - m]
\]

\[
= \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} f[i + n, j + m] \gamma[i, j].
\]

In words, this process scales the shifted function by the values of the matrix \( \gamma \), and thus computes a weighted summation of gray values of the input image \( f[n, m] \). The operation defined by this last equation is called the crosscorrelation of the image with the window function \( \gamma[n, m] \). The correlation operation often is denoted by a
five-pointed star ("pentagram"), e.g.,

\[ g[n, m] = f[n, m] \star [n, m] \]

\[ = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} f[i, j] \gamma[i - n, j - m] \]

The output image \( g \) at pixel indexed by \([n, m]\) is computed by centering the window \( \gamma[n, m] \) on that pixel of the input image \( f[n, m] \), multiplying the window and input image pixel by pixel, and summing the products. This operation produces an output extremum at shifts \([n, m]\) where the gray-level pattern of the input matches that of the window.

In the common case where the sampled function \( \gamma \) is zero outside a domain with compact support of size \( 3 \times 3 \) samples, the function may be written in the form of a \( 3 \times 3 \) matrix or window function:

\[
\gamma[n, m] = \begin{bmatrix}
\gamma_{-1,1} & \gamma_{0,1} & \gamma_{1,1} \\
\gamma_{-1,0} & \gamma_{0,0} & \gamma_{1,0} \\
\gamma_{-1,-1} & \gamma_{0,-1} & \gamma_{1,-1}
\end{bmatrix}
\]

### 16.1.1 Examples of \( 3 \times 3 \) Crosscorrelation Operators

Consider the action of these \( 3 \times 3 \) window functions:

\[
\gamma_1[n, m] = \begin{bmatrix}
0 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}
\]

\[
\gamma_2[n, m] = \begin{bmatrix}
0 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 0
\end{bmatrix}
\]

\[
\gamma_3[n, m] = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
\]

- \( \gamma_1 \) – the only pixel that influences the output \( g[n, m] \) is the identical pixel in the input \( f[n, m] \) – this is the identity operator.

- \( \gamma_2 \) – the output pixel has twice the gray value of the input pixel – this is a uniform contrast stretching operator.
• $\gamma_3$— the output pixel is identical to its right-hand neighbor in the input image – this operator translates the image one pixel to the left.

Once the general crosscorrelation algorithm is programmed, many useful operations on the image $f[n,m]$ can be performed simply by specifying different values for the window coefficients.

### 16.2 Convolution

A mathematically equivalent but generally more convenient neighborhood operation is the **convolution**, which has some very nice mathematical properties. The convolution of two 1-D continuous functions, the input $f[x]$ and the **impulse response** (or kernel, or **point spread function**, or **system function**) $h[x]$ is:

$$g[x] = f[x] * h[x] \equiv \int_{-\infty}^{\infty} d\alpha \ f[\alpha] \ h[x-\alpha].$$

where $\alpha$ is a dummy variable of integration. As for the crosscorrelation, the function $h[x]$ defines the action of the system on the input $f[x]$. By changing the integration variable to $u \equiv x - \alpha$, an equivalent expression for the convolution is found:

$$g[x] = \int_{-\infty}^{\infty} f[\alpha] \ h[x-\alpha] \ d\alpha$$

$$= \int_{u=+\infty}^{u=-\infty} f[x-u] \ h[u] \ (-du)$$

$$= \int_{-\infty}^{\infty} f[x-u] \ h[u] \ du$$

$$= \int_{-\infty}^{\infty} h[\alpha] \ f[x-\alpha] \ d\alpha$$

where the dummy variable was renamed from $u$ to $\alpha$ in the last step. Note that the roles of $f[x]$ and $h[x]$ have been exchanged between the first and last expressions, which means that the input function $f[x]$ and system function $h[x]$ can be interchanged.

The convolution of a continuous 2-D function $f[x,y]$ with a system function $h[x,y]$ is denoted by an asterisk “$*$” and defined as:

$$g[x,y] = f[x,y] * h[x,y]$$

$$\equiv \iint_{-\infty}^{\infty} f[\alpha,\beta] \ h[x-\alpha, y-\beta] \ d\alpha \ d\beta$$

$$= \iint_{-\infty}^{\infty} f[x-\alpha, y-\beta] \ h[\alpha, \beta] \ d\alpha \ d\beta$$

Note the difference between the first forms for the convolution and the crosscorrela-
16.2 CONVOLUTION

\[ f[x, y] \star [x, y] = \int\int_{-\infty}^{\infty} f[\alpha, \beta] \gamma[\alpha - x, \beta - y] \, d\alpha \, d\beta \]

\[ f[x, y] * h[x, y] \equiv \int\int_{-\infty}^{\infty} f[\alpha, \beta] \, h[x - \alpha, y - \beta] \, d\alpha \, d\beta \]

and between the second forms:

\[ f[x, y] \star [x, y] \equiv \int\int_{-\infty}^{\infty} f[x + u, y + v] \gamma[u, v] \, du \, dv \]

\[ f[x, y] * h[x, y] \equiv \int\int_{-\infty}^{\infty} f[x - \alpha, y - \beta] \, h[\alpha, \beta] \, d\alpha \, d\beta \]

The changes of the order of the variables in the first pair says that the function \( \gamma \) is just shifted before multiplying by \( f \) in the crosscorrelation, while the function \( h \) is flipped about its center (or equivalently rotated about the center by 180°) before shifting. In the second pair, the difference in sign of the integration variables says that the input function \( f \) is shifted in different directions before multiplying by the system function \( \gamma \) for crosscorrelation and \( h \) for convolution. In convolution, it is common to speak of filtering the input \( f \) with the kernel \( h \). For discrete functions, the convolution integral becomes a summation:

\[ g[n, m] = f[n, m] * h[n, m] \equiv \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} f[i, j] \, h[n - i, m - j] . \]

Again, note the difference in algebraic sign of the action of the kernel \( h[n, m] \) in convolution and the window \( \gamma_{ij} \) in correlation:

\[ f[n, m] \star [n, m] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} f[i, j] \gamma[i - n, j - m] \]

\[ f[n, m] * h[n, m] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} f[i, j] \, h[n - i, m - j] . \]

This form of the convolution has the very useful property that convolution of an input in the form of an impulse function \( \delta_d[n, m] \) with \( h[n, m] \) yields \( h[n, m] \), hence the name for \( h \) as the impulse response:

\[ \delta_d[n, m] * h[n, m] = h[n, m] \]
where the discrete Dirac delta function $\delta_d[n,m]$ is defined:

$$
\delta_d[n,m] \equiv \begin{cases} 
1 & \text{if } n = m = 0 \\
0 & \text{otherwise}
\end{cases}
$$

16.2.1 Evaluating Discrete Convolutions

1-D

Schematic of the sequence of calculations in 1-D discrete convolution. The 3-pixel kernel $h[n] = [1 \ 2 \ 3]$ is convolved with the input image that is 1 at one pixel and zero elsewhere. The output is a replica of $h[n]$ centered at the location of the impulse.
Schematic of 2-D discrete convolution with the 2-D kernel $h[n,m]$.

$$
\delta[i-n, j-m] \ast h[n,m] = h[n,m]
$$

Examples of 2-D Convolution Kernels

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>+1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$\Rightarrow$ identity

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>+1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$\Rightarrow$ shifts image one pixel to right

Discrete convolution is linear because it is defined by a weighted sum of pixel gray values:

$$
\begin{align*}
    f[n,m] \ast (h_1[n,m] + h_2[n,m]) & = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} f[i,j] \cdot (h_1[n-i,m-j] + h_2[n-i,m-j]) \\
    & = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} (f[i,j] \cdot h_1[n-i,m-j] + f[i,j] \cdot h_2[n-i,m-j]) \\
    & = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} f[i,j] \cdot h_1[n-i,m-j] + \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} f[i,j] \cdot h_2[n-i,m-j]
\end{align*}
$$
The linearity of convolution allows new kernels to be created from sums or differences of other kernels. For example, consider the sum of three $3 \times 3$ kernels whose actions have already been considered:

$$h[n,m] = \frac{1}{3} \begin{pmatrix} 0 & 0 & 0 \\ 0 & +1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 0 & +1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & +1 & 0 \end{pmatrix}$$

The output image $g[n,m]$ is the average of three images: the input and copies shifted one pixel up and down. Therefore, each pixel in $g[n,m]$ is the average of three pixels in a vertical line; $g[n,m]$ is blurred vertically. Note that the kernels have been normalized so that the sum of the elements is unity. This ensures that the gray level of the filtered image will fall within the dynamic range of the input image, but they may not be integers. The output of a lowpass filter must typically be requantized.

### 16.2.2 Convolutions – Edges of the Image

Because a convolution is the sum of weighted gray values in the neighborhood of the input pixel, there is a question of what to do near the edge of the image, i.e., when the neighborhood of pixels in the kernel extends “over the edge” of the image. The common solutions are:

1. consider any pixel in the neighborhood that would extend off the image to have gray value “0”;
2. consider pixels off the edge to have the same gray value as the edge pixel;
3. consider that the convolution in any such case to be undefined; and
4. define any pixels over the edge of the image to have the same gray value as pixels on the opposite edge.

On the face of it, the fourth of these alternatives may seem to be ridiculous, but it is simply a statement that the image is assumed to be periodic, i.e., that:

$$f[n,m] = f[n + kN, m + \ell M]$$

where $N$ and $M$ are the numbers of pixels in a row or column, respectively, and $k, \ell$ are integers. In fact, this is the most common case, and will be treated in depth when global operators are discussed.
Possible strategies for dealing with the edge of the image in 2-D convolution: (a) the input image is padded with zeros; (b) the input image is padded with the same gray values "on the edge;" (c) values "off the edge" are ignored; (d) pixels off the edge are assigned the values on the opposite edge, this assumes that the input image is periodic.

The $3 \times 3$ image $f[n,m]$ is outlined by the bold-face box and the assumed gray values of pixels off the edge of the image are shown in light face for four cases; the presence of an “x” in a convolution kernel indicates that the output gray value is undefined.

16.2.3 Convolutions — Computational Intensity

Evaluating convolutions with large kernels in a serial processor used to be very slow. For example, convolution of a $512^2$-pixel image with an $M \times M$ kernel requires: $2 \cdot 512^2 \cdot M^2$ operations (multiplications and additions) for a total of $4.7 \cdot 10^6$ operations with a $3 \times 3$ kernel and $25.7 \cdot 10^6$ operations with a $7 \times 7$ (these operations generally are performed on floating-point data). The increase in computations as $M^2$ ensures that convolution of large images with large kernels is not very practical by serial brute-force means. In the discussion of global operations to follow, we will introduce an alternative method for computing convolutions via the Fourier transform that requires many fewer operations for large images.

16.2.4 Smoothing Kernels — Lowpass Filtering

If all elements of a convolution kernel have the same algebraic sign, then the operator $O$ sums weighted gray values of input pixels in the neighborhood; if the sum of the elements is one, then the process computes a weighted average of the gray values. Averaging reduces the variability of the gray values of the input image; it "smoothes" the function:
Local averaging decreases the “variability” (variance) of pixel gray values

Put another way, a local averaging filter “pushes” the gray values of pixels towards the mean, and thus reduces the variance of the gray values. For a uniform averaging kernel of a fixed size, functions that oscillate over a period just longer than the kernel (e.g., short-period, high-frequency sinusoids) will be averaged more than slowly varying terms. In other words, local averaging attenuates the high sinusoidal frequencies while passing the low frequencies relatively undisturbed – local averaging operators are lowpass filters. If the kernel size doubles, input sinusoids with twice the period (half the spatial frequency) will be equivalently affected. This action was discussed in the section on realistic sampling; a finite detector averages the signal over its width and reduces modulation of the output signal to a greater degree at higher frequencies.

Local averaging operators are lowpass filters

Obviously, averaging kernels reduce the visibility of additive noise by spreading the difference in gray value of noise pixel from the background over its neighbors. By analogy with temporal averaging, spatial averaging of noise increases SNR by the square-root of the number of pixels averaged if the noise is random and the averaging weights are identical.

The action of an averager can be directional:

\[
\begin{align*}
h_1 [n, m] &= \frac{1}{3} \begin{bmatrix} 0 & +1 & 0 \\ 0 & +1 & 0 \\ 0 & +1 & 0 \end{bmatrix} \quad \text{averages vertically} \\
h_2 [n, m] &= \frac{1}{3} \begin{bmatrix} 0 & 0 & 0 \\ +1 & +1 & +1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{blurs horizontally} \\
h_3 [n, m] &= \frac{1}{3} \begin{bmatrix} +1 & 0 & 0 \\ 0 & +1 & 0 \\ 0 & 0 & +1 \end{bmatrix} \quad \text{blurs diagonally}
\end{align*}
\]

The “rotation” or “reversal” of the convolution kernel means that the action of \( h_3 [n, m] \) blurs diagonally along the direction at 90° that in the kernel.
The elements of an averaging kernel need not be identical, e.g.,

\[
    h[n, m] = \frac{1}{3} \begin{bmatrix}
        +\frac{1}{4} & +\frac{1}{4} & +\frac{1}{4} \\
        +\frac{1}{4} & +1 & +\frac{1}{4} \\
        +\frac{1}{4} & +\frac{1}{4} & +\frac{1}{4}
    \end{bmatrix}
\]

averages over the entire window but the output is primarily influenced by the center pixel; the output blurred less than in the case when all elements are identical.

Other 2-D discrete averaging kernels may be constructed by “orthogonal multiplication,” e.g., we can construct the common 3 × 3 uniform averager via the product of two orthogonal 1-D uniform averagers:

\[
    \begin{bmatrix}
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3}
    \end{bmatrix} \cdot \begin{bmatrix}
        +\frac{1}{3} \\
        +\frac{1}{3} \\
        +\frac{1}{3}
    \end{bmatrix} = \begin{bmatrix}
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3}
    \end{bmatrix}
\]

The associated transfer function is the orthogonal product of the individual 1-D transfer functions. Note that the 1-D kernels can be different, such as a 3-pixel uniform averager along the n-direction and a 2-pixel uniform averager along m:

\[
    \begin{bmatrix}
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3}
    \end{bmatrix} \cdot \begin{bmatrix}
        +\frac{1}{2} \\
        +\frac{1}{2} \\
        +\frac{1}{2}
    \end{bmatrix} = \begin{bmatrix}
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \\
        +\frac{1}{3}
    \end{bmatrix}
\]

**Lowpass-Filtered Images**

*Examples of lowpass filtering: (a) f[n,m]; (b) after uniform averaging over a 3 × 3 neighborhood; (c) after uniform averaging over a 5 × 5 neighborhood. Note that the “fine structure” (such as it is) becomes less visible as the neighborhood size increases.*
16.2.5 Frequency Response of 1-D Averagers

The impulse response of the linear operator that averages uniformly is a unit-area rectangle:

\[ h[x] = \frac{1}{b} \text{RECT} \left( \frac{x}{b} \right) \]

where \( b \) is the width of the averaging region. The corresponding continuous transfer function is:

\[ H[\xi] = \text{SINC} \left( \frac{\xi}{b} \right) \]

In the discrete case, the rectangular impulse response is sampled and the width \( b \) is measured in units of \( \Delta x \). If \( \frac{b}{\Delta x} \) is even, the amplitudes of the endpoint samples are \( \frac{1}{2} \). We consider the cases where \( b = 2 \cdot \Delta x, 3 \cdot \Delta x, \) and \( 4 \cdot \Delta x \). The discrete impulse responses of uniform averagers that are two and three pixels wide have three nonzero samples:

\[
\begin{align*}
    b = 2 \cdot \Delta x &: h_2 [n] = \frac{1}{4} \left( \delta_d [n+1] + 2 \cdot \delta_d [n] + \delta_d [n-1] \right) \\
    b = 3 \cdot \Delta x &: h_3 [n] = \frac{1}{3} \left( \delta_d [n+1] + \delta_d [n] + \delta_d [n-1] \right)
\end{align*}
\]

The linearity of the DFT ensures that the corresponding transfer function may be constructed by summing transfer functions for the identity operator and for translations by one sample each to the left and right. The resulting transfer functions may be viewed as discrete approximations to the continuous \( \text{SINC} \) functions:

\[
\begin{align*}
    H_2[k] &= \frac{1}{4} e^{-2\pi i k \cdot \Delta x} + \frac{1}{2} + \frac{1}{4} e^{+2\pi i k \cdot \Delta x}; \quad -\frac{N}{2} \leq k \leq \frac{N}{2} - 1 \\
    &= \frac{1}{2} \left( 1 + \cos \left( 2\pi \left( k \cdot \Delta x \right) \right) \right) \equiv \text{SINC}_d \left( k; \frac{N}{2} \right) \\
    H_3[k] &= \frac{1}{3} e^{-2\pi i k \cdot \Delta x} + \frac{1}{3} + \frac{1}{3} e^{+2\pi i k \cdot \Delta x} \\
    &= \frac{1}{3} \left( 1 + 2 \cos \left( 2\pi \left( k \cdot \Delta x \right) \right) \right) \equiv \text{SINC}_d \left( k; \frac{N}{3} \right)
\end{align*}
\]

Note that both \( H_2[k] \) and \( H_3[k] \) have unit amplitude at the origin because of the discrete central ordinate theorem. The “zero-crossings” of the two-pixel averager would be located at \( \xi = \pm \frac{1}{2} \) cycle per sample, i.e., at the positive and negative Nyquist frequencies, but the index convention admits only the negative index, \( k = -\frac{N}{2} \). This sinusoid oscillates with a period of two samples and is thus averaged to zero by the two-pixel averager. The three-pixel averager should “block” any sinusoid with a period \( 3 \cdot \Delta x \), which can certainly be constructed because the Nyquist condition is satisfied. The zero crossings of the transfer function “should” be located at \( \xi = \pm \frac{1}{3} \) cycle per sample, which “would” occur at noninteger indices \( k = \pm \frac{N}{3} \). However,
because $\pm \frac{N}{3}$ is not an integer when $N$ is even, the zero crossings of the spectrum occur between samples in the discrete spectrum. In other words, the discrete transfer function of the three-pixel averager has no zeros and thus must “pass” all sampled and unaliased sinusoidal components (albeit with attenuation). This seems to be a paradox – a sinusoid with a period of three pixels can be sampled without aliasing but this function is not “blocked” by a three-pixel averager, whereas the two-pixel sinusoid is blocked by a two-pixel averager. The reason is because there are a noninteger number of periods of length $3 \cdot \Delta x$ in an array where $N$ is even. Thus there will be “leakage” in the spectrum of the sampled function. The resulting “spurious” frequency components pass reach the output. As a final observation, note also that the discrete transfer functions approach the edges of the array “smoothly” (without “cusps”) in both cases, as shown in the figure.

Comparison of 2- and 3-pixel uniform averagers for $N = 64$: (a) central region of the impulse response $h_2[n] = \frac{1}{2} \text{RECT} \left[ \frac{n}{2} \right]$; (b) discrete transfer function $H_2[k]$ compared to the continuous analogue $H_2[\xi] = \text{SINC} \left[ 2\xi \right]$ out to the Nyquist frequency $\xi = \frac{1}{2} \text{cycle/sample}$; (c) $h_3[n] = \frac{1}{3} \text{RECT} \left[ \frac{n}{3} \right]$, which has the same support as $h_2[n]$; (d) $H_3[k]$ compared to $H_3[\xi] = \text{SINC} \left[ 3\xi \right]$, showing the “smooth” transition at the edge of the array.

The discrete impulse response of the four-pixel averager has five nonzero pixels:
The linearity of the DFT ensures that the corresponding transfer function may be constructed by summing the transfer function of the three-pixel averager scaled by $\frac{3}{4}$ with the transfer functions for translation by two samples each to the left and right:

$$H_4[k] = \frac{1}{4} \left( \frac{1}{2} e^{-2\pi ik \cdot 2\Delta \xi} + e^{-2\pi ik \cdot \Delta \xi} + 1 + e^{+2\pi ik \cdot \Delta \xi} + \frac{1}{2} e^{+2\pi ik \cdot 2\Delta \xi} \right)$$

$$= \frac{1}{4} \left( 1 + 2 \cos \left[ 2\pi \left( k \cdot \Delta \xi \right) \right] + \cos \left[ 2\pi \left( k \cdot 2\Delta \xi \right) \right] \right)$$

which also may be thought of as a discrete approximation of a \(\text{SINC}\) function: \(SINCD \left[ k; \frac{N}{4} \right]\). This discrete transfer function has zeros located at $\xi = \pm \frac{1}{4}$ cycle per sample, which correspond to $k = \pm \frac{N}{4}$. Therefore the four-pixel averager “blocks” any sampled sinusoid with period $4 \cdot \Delta x$ from reaching the output. Again the transfer function has “smooth” transitions of amplitude at the edges of the array, thus preventing “cusps” in the periodic spectrum, as shown in the figure.

*Four-pixel averager for \(N = 64\): (a) central region of impulse response $h_4[n] = \frac{1}{4} \text{RECT} \left[ \frac{n}{4} \right]$; (b) Discrete transfer function $H_4[k]$ compared to the continuous transfer function $\text{SINC} \left[ 4\xi \right]$, showing the smooth decay of the discrete case at the edges of the array.*
The general expression for the discrete \( \text{SINC} \) function in the frequency domain suggested by these results for \(-\frac{N}{2} \leq k \leq \frac{N}{2} - 1\) is:

\[
\text{SINC}_d \left[ k; \frac{N}{w} \right] = \begin{cases} 
\frac{1}{w} \left( 1 + 2 \sum_{\ell=1}^{\frac{w-1}{2}} \cos \left[ 2\pi \left( k \cdot \ell \cdot \Delta \xi \right) \right] \right) & \text{if } w \text{ is odd} \\
\frac{1}{w} \left( 1 + \cos \left[ 2\pi \left( k \cdot \frac{w}{2} \cdot \Delta \xi \right) \right] + 2 \sum_{\ell=1}^{\frac{w-1}{2}} \cos \left[ 2\pi \left( k \cdot \ell \cdot \Delta \xi \right) \right] \right) & \text{if } w \text{ is even}
\end{cases}
\]

### 16.2.6 2-D Averagers

Effect of Lowpass Filtering on the Histogram

Because an averaging kernel reduces pixel-to-pixel variations in gray level (and hence the variance of additive random noise in the image), we would expect that clusters of pixels in the histogram of an averaged image to be taller and thinner than in the original image. It should be easier to segment objects based on average gray level from the histogram of an averaged image. To illustrate, we reconsider the example of the house-tree image. The image in blue light and its histogram before and after averaging with a \( 3 \times 3 \) kernel are shown below: Note that there are four fairly distinct clusters in the histogram of the averaged image, corresponding to the house, grass+tree, sky, and clouds+door (from dark to bright). The small clusters at the ends are more difficult to distinguish on the original histogram.
Effect of blurring on the histogram: the $64 \times 64$ color image, the histograms of the 3 bands, and the 3 2-D histograms are shown at top; the same images and histograms after blurring with a $3 \times 3$ kernel are at the bottom, showing the concentration of histogram clusters resulting from image blur.

Note that the noise visible in uniform areas of the images (e.g., the sky in the blue image) has been noticeably reduced by the averaging, and thus the widths of the histogram clusters have decreased. The downside of this process is that pixels on the boundaries of objects now exhibit blends of the colors of both bounding objects, and thus will not be as easy to segment.

### 16.2.7 Differencing Kernels — Highpass Filters

A kernel with both positive and negative terms computes differences of neighboring pixels. From the previous discussion, it is probably apparent that the converse of the
statement about local averaging also is true.

**Local Differencing increases the variance of pixel gray values**

The difference of adjacent pixels with identical gray levels will cancel out, while differences between adjacent pixels will be amplified. Since high-frequency sinusoids vary over shorter distances, differencing operators will enhance them and attenuate slowly varying (i.e., lower-frequency) terms.

**Differencing operators are highpass filters**

because a differencing operator will “attenuate” low-frequency sinusoids (and generally “block” constant terms) while “passing” high-frequency sinusoids. A true differencing operator that outputs a mean value of 0 will “push” the gray values away from the “new” mean, thus increasing the variances.

Subtraction of adjacent pixels can result in negative gray values; this is the spatial analogy of temporal differencing for change detection. The gray values of the output image must be biased “up” for display by adding some constant gray level to all image pixels, e.g., if \( g_{\text{min}} < 0 \), then the negative gray values may be displayed by adding the level \(|g_{\text{min}}|\) to all pixel gray values.

The discrete analogue of differentiation may be derived from the definition of the continuous derivative:

\[
\frac{df}{dx} \equiv \lim_{\tau \to 0} \left( \frac{f(x + \tau) - f(x)}{\tau} \right)
\]

In the discrete case, the smallest nonzero value of \( \tau \) is the sampling interval \( \Delta x \), and thus the discrete derivative is:

\[
\frac{1}{\Delta x} (f[(n+1) \cdot \Delta x] - f[n \cdot \Delta x]) = \frac{1}{\Delta x} f[n \cdot \Delta x] * (\delta[n+1] - \delta[n])
\]

In words, the discrete derivative is the scaled difference of the value at the sample indexed by \( n+1 \) and by \( n \). By setting \( \Delta x = 1 \) sample, the leading scale factor may be ignored. The 1-D derivative operator may be implemented by discrete convolution with a 1-D kernel that has two nonzero elements; we will write it with three elements to clearly denote the sample indexed by \( n = 0 \).

\[
f[n] * (\delta[n+1] - \delta[n]) = f[n] * \begin{bmatrix} +1 & -1 & 0 \end{bmatrix} \equiv f[n] * \partial[n]
\]

where \( \partial[n] \equiv \begin{bmatrix} +1 & -1 & 0 \end{bmatrix} \) is the discrete impulse response of differentiation, which is perhaps better called a differencing operator. Note that the impulse response may
be decomposed into its even and odd parts.

\[
\partial_{\text{even}}[n] = \begin{bmatrix}
\frac{1}{2} & -1 & \frac{1}{2}
\end{bmatrix}
= \left(\frac{1}{2}\right) \begin{bmatrix}
+1 & +1 & +1
\end{bmatrix} - \left(\frac{3}{2}\right) \begin{bmatrix}
0 & 1 & 0
\end{bmatrix}
\]

\[
\partial_{\text{odd}}[n] = \begin{bmatrix}
\frac{1}{2} & 0 & -\frac{1}{2}
\end{bmatrix} = \left(\frac{1}{2}\right) \begin{bmatrix}
+1 & 0 & -1
\end{bmatrix}
\]

The even part is a weighted difference of the identity operator and the three-pixel averager, while the odd part computes differences of pixels separated by two sample intervals.

The corresponding 2-D derivative kernel in the \(x\)-direction is:

\[
[0 \ 0 \ 0]
[+1 \ -1 \ 0]
[0 \ 0 \ 0]
\]

\[
\equiv \partial_x
\]

The resulting image is equivalent to the difference between an image translated one pixel to the left and an unshifted image, i.e.,

\[
\frac{\partial}{\partial x} f[x, y] = \lim_{\Delta x \to 0} \frac{f[x + \Delta x, y] - f[x, y]}{\Delta x}
\]

\[
\implies \frac{\partial}{\partial x} f[x, y] \equiv f[(n + 1) \cdot \Delta x, m \cdot \Delta y] - f[n \cdot \Delta x, m \cdot \Delta y]
\]

\[
\implies \partial_x \ast f[n, m] = f[n + 1, m] - f[n, m]
\]

because the minimum nonzero value of the translation \(\Delta x = 1\) sample. The corresponding discrete partial derivative in the \(y\)-direction is:

\[
\partial_y \ast f[n, m] \equiv f[n, m + 1] - f[n, m]
\]

The partial derivative in the \(y\)-direction is the difference of a replica translated “up” and the original:

\[
\partial_y = \begin{bmatrix}
0 & 0 & 0
0 & -1 & 0
0 & +1 & 0
\end{bmatrix}
\]
This definition of the derivative effectively “locates” the edge of an object at the pixel immediately to the right or above the “crack” between pixels that is the actual edge.

“Symmetric” (actually antisymmetric or odd) versions of the derivative operators are sometimes used that evaluate the difference across two pixels:

\[
\begin{align*}
\partial_x &= \begin{bmatrix}
0 & 0 & 0 \\
+1 & 0 & -1 \\
0 & 0 & 0
\end{bmatrix} \\
\partial_y &= \begin{bmatrix}
0 & +1 & 0 \\
0 & 0 & 0 \\
0 & -1 & 0
\end{bmatrix}
\end{align*}
\]

These operators locate the edge of an object between two pixels symmetrically, but produce an output that is two pixels wide centered about the location of the edge:
First derivatives as edge detectors: (a) 1-D object with both positive and negative edges (“dark to bright” and “bright to dark”); (b) bipolar output from discrete first derivative \( \begin{bmatrix} +1 & -1 & 0 \end{bmatrix} \), the dashed red line showing location of actual edge; (c) output of antisymmetric first derivative \( \begin{bmatrix} +1 & -2 & +1 \end{bmatrix} \), showing that the “edges” are two pixels wide. The dynamic ranges of the output images are ±1 and so are usually rescaled to render the maximum value with white, “0” with gray, and the theoretical minimum with black.

Higher-Order Derivatives

The kernels for higher-order derivatives are easily computed because convolution is associative. The convolution kernel for the 1-D second derivative is obtained by
autoconvolving the kernel for the 1-D first derivative:

\[
\frac{\partial^2}{\partial x^2} f[x, y] = \frac{\partial}{\partial x} \left( \frac{\partial}{\partial x} f[x, y] \right)
\]

\[
= \frac{\partial}{\partial x} \left( \lim_{\Delta x \to 0} \frac{f[x + \Delta x, y] - f[x, y]}{\Delta x} \right)
\]

\[
= \lim_{\Delta x \to 0} \left( \frac{f[x + 2 \cdot \Delta x, y] - f[x + \Delta x, y]}{\Delta x} \right) - \lim_{\Delta x \to 0} \left( \frac{f[x + \Delta x, y] - f[x, y]}{\Delta x} \right)
\]

\[
= \lim_{\Delta x \to 0} \left( \frac{f[x + 2 \cdot \Delta x, y] - 2f[x + \Delta x, y] + f[x, y]}{\Delta x} \right)
\]

\[
\Rightarrow \partial^2_x \ast f[n, m] = f[n + 2, m] - 2f[n + 1, m] - f[n, m]
\]

which may be evaluated by convolution with a three-element kernel, which is displayed in a five-pixel window or a 5 × 5 2-D window to identify the center pixel:

\[
\partial^2_x = \begin{bmatrix} +1 & -2 & +1 & 0 & 0 \end{bmatrix}
\]

The kernel usually is “centered” in a 3 × 3 array by translating the kernel one pixel to the right and lopping off the zeros:

\[
\hat{\partial}^2_x = \begin{bmatrix} 0 & 0 & 0 \\ +1 & -2 & +1 \\ 0 & 0 & 0 \end{bmatrix}
\]

Except for the one-pixel translation, this operation generates the same image produced by the cascade of two first-derivative operators. The corresponding 2-D second partial derivative kernels in the y-direction are:

\[
\partial^2_y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & +1 & 0 \end{bmatrix}
\]
CHAPTER 16 LOCAL OPERATIONS

\[
\hat{\partial}_y^2 = \begin{bmatrix}
0 & +1 & 0 \\
0 & -2 & 0 \\
0 & +1 & 0 \\
\end{bmatrix}
\]

1-D rectangle object and the location of its edges obtained by convolving with “centered” second-derivative kernel \[+1 \quad -2 \quad +1\] in the x-direction; the location of the edge transitions are shown as red dashed lines. The edge is identified by a pair of impulses with opposite signs. Again the dynamic range of the output image is bipolar and must be rescaled so that “0” is rendered as midgray.

The derivation may be extended to derivatives of still higher order by convolving kernels to obtain the kernels for the 1-D third and fourth derivatives. The third
16.2 CONVOLUTION

The gray-level extrema of the image produced by a differencing operator indicate pixels in regions of rapid variation, e.g., at the edges of distinct objects. The visibility of these pixels can be further enhanced by a subsequent contrast enhancement or thresholding operation.

Because they compute weighted differences in pixel gray value, differencing operators also will enhance the visibility of noise in an image. Consider the 1-D example where the input image \( f [x] \) is a 3-bar chart with added noise, so that the signal-to-noise ratio (SNR) of 4. The convolution of the input \( f [x] \) with an averaging kernel \( h_1 [x] = \frac{1}{3} +1 +1 +1 \) and with differencing kernel \( h_2 [x] = -1 +3 -1 \) are shown:
Effect of averaging and sharpening on an image with noise: (a) \( f[n] + \text{noise} \); (b) after averaging over a 3-pixel neighborhood with \( h_1[n] = \frac{1}{3} [1 \ 1 \ 1] \); (c) after applying a sharpening operator \( h_2[n] = [-1 \ 3 \ -1] \), showing the increase in the variance.

Note that the noise is diminished by the averaging kernel and enhanced by the differencing kernel.

### 16.2.8 Frequency Responses of 2-D Derivative Operators:

#### First Derivatives

The 2-D \( x \)-derivative operator denoted by \( \partial_x \):

\[
\partial_x = \begin{bmatrix}
0 & 0 & 0 \\
+1 & -1 & 0 \\
0 & 0 & 0 
\end{bmatrix}
\]

has the associated discrete transfer function obtained via the discrete Fourier transform:

\[
\mathcal{F}_2 \{\partial_x\} = H_{\partial_x}[k, \ell] = \left( -1 + e^{i2\pi\frac{k}{N}} \right) \cdot 1[\ell] \\
= \left( -1[k, \ell] + \cos \left[ 2\pi \frac{k}{N} \right] \cdot 1[\ell] \right) + i \left( \sin \left[ 2\pi \frac{k}{N} \right] \cdot 1[\ell] \right)
\]
The differencing operator in the $y$-direction and its associated transfer function are obtained by rotating the expressions just derived by $+\frac{\pi}{2}$ radians. The kernel is:

$$
\partial_y = \begin{bmatrix}
0 & 0 & 0 \\
0 & -1 & 0 \\
0 & +1 & 0
\end{bmatrix}
$$

The corresponding transfer function is the rotated version of $\mathcal{F}_2 \{ \partial_x \}$:

$$
\mathcal{F}_2 \{ \partial_y \} = H_{(\partial_y)}[k, \ell] = 1[k] \cdot \left( -1 + e^{+2\pi i \frac{\ell}{N}} \right)
$$

We can also define 2-D derivatives along angles. We can also define differences along the diagonal directions:

$$
\partial_{(\theta=+\frac{\pi}{4})} = \begin{bmatrix}
0 & 0 & 0 \\
0 & -1 & 0 \\
+1 & 0 & 0
\end{bmatrix}
$$

$$
\partial_{(\theta=+\frac{3\pi}{4})} = \begin{bmatrix}
0 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & +1
\end{bmatrix}
$$

The angle in radians has been substituted for the subscript. Again, the continuous distance between the elements has been scaled by $\sqrt{2}$. 
1-D discrete derivative operator and its transfer function: (a) impulse response \( h[n] = \delta_d[n + 1] - \delta_d[n] \); The samples of the discrete transfer function are shown as (b) real part, (c) imaginary part, (d) magnitude, and (e) phase, along with the corresponding continuous transfer function \( H[\xi] = i2\pi\xi \). Note that real part of the discrete transfer function is not zero. The imaginary part is forced to be periodic by the DFT and the discrete MTF “rolls off” when approaching the Nyquist frequency.
16.2 CONVOLUTION

1-D Antisymmetric Differentiation Kernel

We can also construct a discrete differentiator with odd symmetry by placing the components of the discrete “doublet” at samples \( n = \pm 1 \):

\[
(\partial_x)_2 = \begin{bmatrix} +1 & 0 & -1 \end{bmatrix}
\]

This impulse response is proportional to the odd part of the original 1-D differentiator. The corresponding transfer function is again easy to evaluate via the appropriate combination of translation operators. Because \((\partial_x)_2\) is odd, the real part of the discrete transfer function is zero, as shown in the figure:

\[
H[k] = \exp \left[ +2\pi i \left( \frac{k}{N} \right) \right] - \exp \left[ -2\pi i \left( \frac{k}{N} \right) \right] = 2i \sin \left( 2\pi \frac{k}{N} \right)
\]

\[
|H[k]| = 2 \left| \sin \left( 2\pi \frac{k}{N} \right) \right|
\]

\[
\Phi \{ H[k] \} = \frac{\pi}{2} \left( SGN[k] - \delta_d \left[ k + \frac{N}{2} \right]\right)
\]

Note that this transfer function evaluated at the Nyquist frequency is:

\[
\left| H \left[ k = -\frac{N}{2} \right] \right| = 2 \left| \sin \left[ -\pi \right] \right| = 0
\]

which means that this differentiator also “blocks” the Nyquist frequency. This may be seen by convolving \((\partial_x)_2\) with a sinusoid function that oscillates with a period of two samples. Adjacent positive extrema are multiplied by \(\pm 1\) in the kernel and thus cancel. Also note that the transfer function amplifies lower frequencies more and larger frequencies less than the continuous transfer function.
1-D discrete antisymmetric derivative operator and its transfer function: (a) 1-D impulse response $h[n]$; Because $h[n]$ is odd, $H[k]$ is imaginary and odd. The samples of the discrete transfer function are shown as (b) real part, (c) imaginary part, (d) magnitude, and (e) phase, along with the corresponding continuous transfer function $H[\xi] = i2\pi\xi$. Note that the magnitude of the discrete transfer function is attenuated near the Nyquist frequency and that its phase is identical to that of the transfer function of the continuous derivative.
1-D Second Derivative

The impulse response and transfer function of the continuous second derivative are easily obtained from the derivative theorem:

\[ h[x] = \delta''[x] \]
\[ H[\xi] = (2\pi i\xi)^2 = -4\pi^2\xi^2 \]

Again, different forms of the discrete second derivative may be defined. One form is obtained by differentiating the first derivative operator via discrete convolution of two replicas of \( \partial_x \). The result is a five-pixel kernel including two null weights:

\[
\begin{bmatrix}
+1 & -1 & 0 & +1 & -1 & 0
\end{bmatrix}
\times
\begin{bmatrix}
+1 & -1 & 0 & +1 & -1 & 0
\end{bmatrix}
= \begin{bmatrix}
+1 & -2 & +1 & 0 & 0
\end{bmatrix}
\]

The corresponding discrete transfer function is obtained by substituting results from the translation operator:

\[
H[k] = e^{+2\pi i\frac{2k}{N}} - 2 e^{+2\pi i\frac{k}{N}} + 1[k]
\]
\[
= e^{+2\pi i\frac{k}{N}} \left( e^{+2\pi i\frac{k}{N}} - 2 + e^{-2\pi i\frac{k}{N}} \right)
\]
\[
= 2 e^{+2\pi i\frac{k}{N}} \left( \cos \left[ 2\pi \frac{k}{N} \right] - 1 \right)
\]

The leading linear phase factor usually is discarded to produce the real-valued and symmetric discrete transfer function:

\[
H[k] = 2 \left( \cos \left[ 2\pi \frac{k}{N} \right] - 1 \right)
\]

Deletion of the linear phase is the same as translation of the original discrete second derivative kernel by one pixel to the right. The discrete impulse response for this symmetric discrete kernel is also real valued and symmetric:

\[
h[n] = \partial_x^2 \equiv 0 +1 -2 +1 0 = +1 -2 +1
\]
\[
= \delta_d[n+1] - 2\delta_d[n] + \delta_d[n-1]
\]

and the magnitude and phase of the transfer function are:
\[
|H[k]| = 2 \left( 1 - \cos \left( \frac{2\pi k}{N} \right) \right)
\]
\[
\Phi \{ H[k] \} = \begin{cases} 
-\pi & \text{for } k \neq 0 \\
0 & \text{for } k = 0 
\end{cases} = \pi ( -1 + \delta_d [k])
\]
as shown in the figure. The amplitude of the discrete transfer function at the Nyquist frequency is:
\[
H \left[ k = \frac{-N}{2} \right] = 2 \cdot (\cos [-\pi] - 1) = -4
\]
while that of the continuous transfer function is \(-4\pi^2 \left( -\frac{1}{2} \right)^2 = -\pi^2 \cong -9.87\), so the discrete second derivative does not amplify the negative amplitude at the Nyquist frequency as much as the continuous second derivative. The transfer function is a discrete approximation of the parabola and again approaches the edges of the array in a manner that is smoothly periodic.

Higher-order discrete derivatives may be derived by repeated discrete convolution of \( \partial_x \), after discarding any linear-phase factors.

1-D Discrete second derivative: (a) Impulse response \( \partial_x^2 \); (b) comparison of corresponding continuous transfer functions of the second derivative with the DFT of the discrete impulse response. Note the periodicity of the discrete transfer function.

16.2.9 Laplacian Operator

The Laplacian operator for continuous functions was introduced in the discussion of electromagnetism. It is the sum of orthogonal second partial derivatives:
\[
\nabla^2 f [x, y] \equiv \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) f [x, y]
\]
The associated transfer function is the negative quadratic that evaluates to 0 at DC, which again demonstrates that constant terms are blocked by differentiation:

\[ H[\xi, \eta] = -4\pi^2 (\xi^2 + \eta^2) \]

The discrete Laplacian operator is the sum of the orthogonal 2-D second-derivative kernels:

\[
\partial_x^2 + \partial_y^2 \equiv \nabla_d^2 = \begin{pmatrix}
0 & 0 & 0 \\
1 & -2 & 1 \\
0 & 0 & 0
\end{pmatrix} + \begin{pmatrix}
0 & +1 & 0 \\
0 & -2 & 0 \\
0 & +1 & 0
\end{pmatrix} = \begin{pmatrix}
0 & +1 & 0 \\
1 & -4 & 1 \\
0 & +1 & 0
\end{pmatrix}
\]

The discrete transfer function of this “standard” discrete Laplacian kernel is:

\[
H[k, \ell] = 2 \left( \cos \left[ \frac{2\pi k}{N} \right] - 1 \right) + 2 \left( \cos \left[ \frac{2\pi \ell}{N} \right] - 1 \right)
\]

\[
= 2 \left( \cos \left[ \frac{2\pi k}{N} \right] + \cos \left[ \frac{2\pi \ell}{N} \right] - 2 \right)
\]

The amplitude at the origin is \( H[k = 0, \ell = 0] = 0 \) and decays in the horizontal or vertical directions to \(-6\) at the “edge” of the discrete array and to \(-8\) at its corners.

**Rotated Laplacian**

The sum of the second-derivative kernels along the diagonals creates a rotated version of the Laplacian, which is “nearly” equivalent to rotating the operator \( \nabla_d^2 \) by \( \theta = +\frac{\pi}{4} \) radians:

\[
\left( \partial_{(\pm\frac{\pi}{4})}^2 + \partial_{(\pm\frac{3\pi}{4})}^2 \right) = \nabla_d^2(\pm\frac{\pi}{4}) = \begin{pmatrix}
0 & 0 & +1 \\
0 & -2 & 0 \\
+1 & 0 & 0
\end{pmatrix} + \begin{pmatrix}
+1 & 0 & 0 \\
0 & -2 & 0 \\
0 & 0 & +1
\end{pmatrix} = \begin{pmatrix}
+1 & 0 & +1 \\
0 & -4 & 0 \\
+1 & 0 & +1
\end{pmatrix}
\]

Derivation of the transfer function is left to the student; its magnitude is zero at the origin and its maximum negative values are located at the horizontal and vertical edges, but the transfer function zero at the corners.
The real and symmetric transfer functions of Laplacian operators: (a) “normal” Laplacian from Eq.(20.50); (b) rotated Laplacian from Eq.(20.51), showing that the amplitude rises back to 0 at the corners of the array.

A 2-D example is shown in the figure, where the input function is nonnegative and the bipolar output $g[n,m]$ is displayed as amplitude and as magnitude $|g[n,m]|$, which shows that the response is largest at the edges and corners.
Action of the 2-D discrete Laplacian. The input amplitudes are nonnegative in the interval $0 \leq f \leq 1$, and the output amplitude is bipolar in the interval $-2 \leq g \leq +2$ in this example. As shown in the magnitude image, the extrema of output amplitude occurs at the edges and corners of the input.
Isotopic Laplacian:

A commonly used “isotropic” Laplacian is obtained by summing the original and rotated Laplacian kernels:

\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & -4 & 1 \\
0 & 1 & 0
\end{pmatrix} + \begin{pmatrix}
1 & 0 & 1 \\
0 & -4 & 0 \\
1 & 0 & 1
\end{pmatrix} = \begin{pmatrix}
1 & 1 & 1 \\
1 & -8 & 1 \\
1 & 1 & 1
\end{pmatrix} = \begin{pmatrix}
+1 & +1 & +1 \\
+1 & -8 & +1 \\
+1 & +1 & +1
\end{pmatrix}
\]

The linearity of the DFT ensures that the transfer function of the isotropic Laplacian is the real-valued and symmetric sum of the “normal” and rotated Laplacians.

Generalized Laplacian

The isotropic Laplacian just considered may be written as the difference of a $3 \times 3$ average and a scaled discrete delta function:

\[
\begin{pmatrix}
+1 & +1 & +1 \\
+1 & -8 & +1 \\
+1 & +1 & +1
\end{pmatrix} = \begin{pmatrix}
+1 & +1 & +1 \\
+1 & +1 & +1 \\
+1 & +1 & +1
\end{pmatrix} - 9 \cdot \begin{pmatrix}
0 & 0 & 0 \\
0 & +1 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

which suggests that the Laplacian operator may be generalized to include any operator that computes the difference of a scaled original image and replicas that were blurred by some averaging kernel. For example, the impulse response of the averager may be the 2-D circularly symmetric and continuous Gaussian impulse response normalized to unit volume:

\[
h[x, y] = \frac{1}{|b|^2} \exp \left[ -\pi \left( \frac{x^2 + y^2}{b^2} \right) \right]
\]

where the decay parameter $b$ determines the rate at which the values of the kernel decrease away from the center and the amplitude parameter $A$ often is selected to normalize the sum of the elements of the kernel to unity, thus ensuring that the process computes a weighted average. A normalized discrete approximation of the Gaussian kernel with $b = 2 \cdot \Delta x$ is:
The corresponding generalized Laplacian operator is this difference of this quantized Gaussian and the $5 \times 5$ identity kernel:

$$h_1[n, m] \approx \frac{1}{2047} \begin{bmatrix} 1 & 11 & 23 & 11 & 1 \\ 11 & 111 & 244 & 111 & 11 \\ 23 & 244 & 535 & 244 & 23 \\ 11 & 111 & 244 & 111 & 11 \\ 1 & 11 & 23 & 11 & 1 \end{bmatrix}$$

$$h_2[n, m] = \frac{1}{21} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 1 & 2 & 5 & 2 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} - \alpha \cdot \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\alpha$ is the weighting of the identity image. Of course the leading scale factor of $\frac{1}{21}$ can be ignored; it just scales the gray values of the image. Note that the sum of the elements of this generalized Laplacian kernel is zero because of the normalization of the Gaussian kernel, and so this operator applied to an input with uniform gray value will yield a null image.

### 16.2.10 Discrete “Sharpening” Operators

#### 1-D Case

A “sharpening” operator passes all sinusoidal components with no change in phase while amplifying those with large spatial frequencies. This action will tend to compensate for the effect of lowpass filtering. One example of a 1-D continuous sharpener is constructed from the second derivative; the amplitude of its transfer function is unity at the origin and rises as $\xi^2$ for larger spatial frequencies:

$$H[\xi] = 1 + 4\pi^2\xi^2$$
The corresponding continuous 1-D impulse response is the difference of the identity and second-derivative kernels

\[ h[x] = \delta[x] - \delta''[x] \]

A discrete version of the impulse response may be generated by substituting the discrete Dirac delta function and the “centered” discrete second derivative operator:

\[
\begin{align*}
  h[n] &= \delta_d - \partial_x^2 = \begin{bmatrix} 0 & +1 & 0 & -1 & -2 & +1 \end{bmatrix} \\
  &= \begin{bmatrix} -1 & +3 & -1 \end{bmatrix} = \begin{bmatrix} 0 & +4 & 0 \end{bmatrix} - \begin{bmatrix} +1 & +1 & +1 \end{bmatrix} \\
  &= 4 \cdot \delta_d[n] - RECT\left[\frac{n}{3}\right]
\end{align*}
\]

The transfer function of the discrete sharpener is:

\[
H[k] = 4 \cdot 1[k] - \left( 1 + 2 \cos \left[ 2\pi \frac{k}{N} \right] \right) \]

\[
= 3 - 2 \cos \left[ 2\pi \frac{k}{N} \right]
\]

The amplitudes of the transfer function at DC and at the Nyquist frequency are:

\[
H[k = 0] = +1 \\
H\left[ k = \frac{-N}{2} \right] = +5
\]

In words, the “second-derivative sharpener” amplifies the amplitude of the sinusoidal component that oscillates at the Nyquist frequency by a factor of 5.

The action of this sharpening operator on a “blurry” edge is shown in the figure. The slope of the edge is “steeper” after sharpening, but the edge also “overshoots” the correct amplitude at both sides.
Action of 1-D 2nd-derivative sharpening operator on a “blurry” edge. The angle of the slope of the sharpened edge is “steeper”, but the amplitude “overshoots” the correct value on both sides of the edge. The output is not the ideal sharp edge, so this operator only approximates the ideal inverse filter.

This interpretation may be extended to derive other 1-D sharpeners by computing the difference between a scaled replica of the “original” (blurred) input image and an image obtained by passing through a different lowpass filter.

2-D Sharpening Operators

We can generalize the 1-D discussion to produce a 2-D sharpener based on the Laplacian. The discrete version often is used to sharpen digital images that have been blurred by unknown lowpass filters. The process is ad hoc; it is not “tuned” to the details of the blurring process and so is not an “inverse” filter. It cannot generally reconstruct the original sharp image, but it “steepens” the slope of pixel-to-pixel changes in gray level, thus making the edges appear “sharper.” The impulse response of the 2-D continuous Laplacian sharpener is:

\[ f(x, y, 0) \cong f(x, y, z) - \alpha \cdot \nabla^2 f(x, y, z) \]

where \( \alpha \) is a real-valued free parameter that allows the sharpener to be “tuned” to the amount of blur. Obviously, the corresponding discrete solution is:

\[ g[n, m] = f[n, m] - \alpha \cdot \nabla^2_d f[n, m] \]

\[ = (\delta_d[n, m] - \alpha \cdot \nabla^2_d) \ast f[n, m] \]

where \( \nabla^2_d[n, m] \) is a Laplacian kernel that may be selected from the variants already considered. A single discrete sharpening kernel \( h[n, m] \) may be constructed from the
simplest form for the Laplacian:

\[
\begin{pmatrix}
0 & 0 & 0 \\
0 & +1 & 0 \\
0 & 0 & 0 \\
\end{pmatrix} - \alpha \cdot
\begin{pmatrix}
0 & +1 & 0 \\
+1 & -4 & +1 \\
0 & +1 & 0 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 & -\alpha & 0 \\
-\alpha & 1 + 4\alpha & -\alpha \\
0 & -\alpha & 0 \\
\end{pmatrix}
\]

The parameter \(\alpha\) may be increased to enhance the sharpening by steepening the slope of the edge profile and also increasing the “overshoot.” Selection of \(\alpha = +1\) produces a commonly used sharpening kernel:

\[
\begin{pmatrix}
0 & -1 & 0 \\
-1 & +5 & -1 \\
0 & -1 & 0 \\
\end{pmatrix}
\]

The weights in the kernel sum to unity, which means that the average gray value of the image is preserved. In words, this process amplifies differences in gray level of adjacent pixels while preserving the mean gray value.

The corresponding discrete transfer function for the parameter \(\alpha\) is:

\[
H_1[k, \ell; \alpha] = (1 + 4\alpha) - 2\alpha \left( \cos \left[ \frac{2\pi k}{N} \right] + \cos \left[ \frac{2\pi \ell}{N} \right] \right)
\]

In the case \(\alpha = +1\), the resulting transfer function is:

\[
H_1[k, \ell; \alpha = 1] = 5 - 2 \left( \cos \left[ \frac{2\pi k}{N} \right] + \cos \left[ \frac{2\pi \ell}{N} \right] \right)
\]

which has its maximum amplitude of \((H_1)_{\text{max}} = 9\) at the corners of the array.

A sharpening operator also may be derived from the isotropic Laplacian:

\[
\begin{pmatrix}
-\alpha & -\alpha & -\alpha \\
-\alpha & 1 + 8\alpha & -\alpha \\
-\alpha & -\alpha & -\alpha \\
\end{pmatrix}
\]

Again, the sum of the elements in the kernel is unity, ensuring that the average gray value of the image is preserved by the action of the sharpener. If the weighting factor is again selected to be unity, the kernel is the difference of a scaled original and a
3 × 3 blurred copy:

\[
\begin{bmatrix}
-1 & -1 & -1 \\
-1 & +9 & -1 \\
-1 & -1 & -1 \\
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & +10 & 0 \\
0 & 0 & 0 \\
\end{bmatrix}
= \begin{bmatrix}
0 & 0 & 0 \\
0 & +1 & +1 \\
0 & +1 & +1 \\
\end{bmatrix}
\]

This type of process has been called *unsharp masking* by photographers, who have often printed a “sandwich” of a negative of the original image and a blurred positive transparency to produce a sharpened image of the original. This difference of the blurred image and the original is easily implemented in a digital system as a single convolution.

An example of 2-D sharpening is shown in the figure. Note the “overshoots” at the edges in the sharpened image. The factor of +9 ensures that the dynamic range of the sharpened image can be as large as from +9 to −8 times the maximum gray value, or −2040 ≤ ̂f ≤ +2295 for an 8-bit image. This would only happen for an isolated bright pixel at the maximum surrounded by a neighborhood of black pixels, and vice versa. In actual use, the range of values is considerably smaller. The image gray values either have to be biased up and rescaled or “clipped” at the maximum and minimum, as was done here.
CHAPTER 16 LOCAL OPERATIONS

Action of the 2-D sharpening operator based on the Laplacian. The original image \( f[n,m] \) has been blurred by a \( 3 \times 3 \) uniform averager to produce \( g[n,m] \). The action of the \( 3 \times 3 \) Laplacian sharpener on \( g[n,m] \) produced a bipolar image with much larger dynamic range. The output clipped at the original dynamic range is \( \hat{f}[n,m] \).

The “overshoots” at the edges conveys the impression of a sharper image.

16.2.11 2-D Gradient

The gradient of a 2-D continuous function \( f[x,y] \) also was defined in the discussion of electromagnetism. It constructs a 2-D vector at each location in a scalar (gray-scale) image. The \( x \)-and \( y \)-components of the vector at each location are the \( x \)- and \( y \)-derivatives of the image:

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

The image \( f[n,m] \) is a scalar function which assigns a numerical gray value \( f \) to each coordinate \([n,m]\). The gray value \( f \) is analogous to terrain “elevation” in a
map. This process calculates a vector at each coordinate \([x, y]\) of the scalar image whose Cartesian components are \(\frac{\partial f}{\partial x}\) and \(\frac{\partial f}{\partial y}\). Note that the 2-D vector \(\nabla f\) may be represented in polar form as magnitude \(|\nabla f|\) and direction \(\Phi\{\nabla f\}\):

\[
|\nabla f[x, y]| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}
\]

\[
\Phi\{\nabla f[n, m]\} = \tan^{-1}\left[\frac{\frac{\partial f}{\partial y}}{\frac{\partial f}{\partial x}}\right]
\]

The 2-D vector at each location has the values:

\[
g[n, m] = \nabla f[n, m] = \begin{bmatrix}
\partial_x * f[n, m] \\
\partial_y * f[n, m]
\end{bmatrix}
\]

This vector points “uphill” in the direction of the maximum “slope” in gray level. The magnitude \(|\nabla f|\) is the “slope” of the 3-D surface \(f\) at pixel \([n, m]\). The azimuth angle (often called the phase by analogy with complex numbers) of the gradient \(\Phi\{\nabla f[n, m]\}\) is the compass direction toward which the slope points “uphill.”

The discrete version of the gradient magnitude also is a useful operator in digital image processing, as it will take on extreme values at edges between objects. The magnitude of the gradient often is approximated as the sum of the magnitudes of the components:

\[
|\nabla f[n, m]| = \sqrt{(\partial_x * f[n, m])^2 + (\partial_y * f[n, m])^2}
\]

\[
\cong |\partial_x * f[n, m]| + |\partial_y * f[n, m]|
\]

The gradient is not a linear operator, and thus can neither be evaluated as a convolution nor described by a transfer function. The largest values of the magnitude of the gradient correspond to the pixels where the gray value “jumps” by the largest amount, and thus the thresholded magnitude of the gradient may be used to identify such pixels. In this way the gradient may be used as an “edge detection operator.”
Example of the discrete gradient operator $\nabla f[n,m]$. The original object is the nonnegative function $f[n,m]$ shown in (a), which has amplitude in the interval $0 \leq f \leq +1$. The gradient at each pixel is the 2-D vector with bipolar components $\left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]$. These component images are shown in (b) and (c). The gradient also may be displayed as magnitude $\sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}$ and angle $\phi = \tan^{-1}\left(\frac{\partial f}{\partial x}/\frac{\partial f}{\partial y}\right)$. The extrema of the magnitude are located at corners and edges in $f[n,m]$. 

![Image of the discrete gradient operator](image.png)
16.2.12 Pattern Matching

It often is useful to design kernels to locate specific gray-level patterns, such as edges at particular orientations, corners, isolated pixels, particular shapes, or what have you. Particularly in the early days of digital image processing when computers were less capable than they are today, the computational intensity of the calculation often was an important issue. It was desirable to find the least intensive method for common tasks such as pattern detection, which generally meant that the task was performed in the space domain using a small convolution kernel rather than calculating a better approximation to the ideal result in the frequency domain. That said, the process of designing and applying a pattern-matching kernel illuminates some of the concepts and thus is worth some time and effort.

A common technique for pattern matching convolves the input image with a kernel that is the same size as the “reference” pattern. The process and its limitations will be illustrated by example. Consider an input image \( f[n,m] \) that is composed of two replicas of some real-valued nonnegative pattern of gray values, \( p[n,m] \), centered at coordinates \([n_1, m_1]\) and \([n_2, m_2]\) with respective amplitudes \( A_1 \) and \( A_2 \). The image also includes a bias \( b \cdot 1[n,m] \):

\[
f[n,m] = A_1 \cdot p[n-n_1, m-m_1] + A_2 \cdot p[n-n_2, m-m_2] + b \cdot 1[n,m]
\]

The appropriate kernel of the discrete filter is:

\[
\hat{m}[n,m] = p[-n,-m]
\]

which also is real valued and nonnegative within its region of support. The output from this matched filter autocorrelation of the pattern centered at those coordinates:

\[
g[n,m] = f[n,m] \ast \hat{m}[n,m] \\
= A_1 \cdot p[n,m] \ast p[n,m]|_{n=n_1, m=m_1} + A_2 \cdot p[n,m] \ast p[n,m]|_{n=n_2, m=m_2} + b \cdot (1[n,m] \ast p[-n,-m]) \\
= A_1 \cdot p[n,m] \ast p[n,m]|_{n=n_1, m=m_1} + A_2 \cdot p[n,m] \ast p[n,m]|_{n=n_2, m=m_2} + b \cdot \sum_{n,m} p[n,m]
\]

The last term is the constant output level from the convolution of the bias with the matched filter, which produces the sum of the product of the bias and the weights at each sample. The spatially varying autocorrelation functions rest on a bias proportional to the sum of the gray values \( p \) in the pattern. If the output bias is large, it can reduce the “visibility” of small (but significant) variations in the autocorrelation in exactly the same way as small modulations of a nonnegative sinusoidal function with a large bias are difficult to see. It is therefore convenient to construct a matched filter kernel whose weights sum to zero. It only requires subtraction of the average
value from each sample of the kernel:

\[
\hat{m}[n, m] = p[-n, -m] - \overline{p} \Rightarrow \sum_{n,m} \hat{m}[-n, -m] = \sum_{n,m} \hat{m}[n, m] = 0
\]

Thus ensuring that the constant bias vanishes. This result determines the strategy for designing convolution kernels that produce outputs that have large magnitudes at pixels centered on neighborhoods that contain these patterns and small magnitudes in neighborhoods where the feature does not exist. For example, consider an image containing an “upper-right corner” of a brighter object on a darker background:

\[
\begin{array}{cccccc}
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & 50 & 50 & 50 & 50 & 50 \\
\vdots & 50 & 50 & 50 & 50 & 50 \\
\vdots & 100 & 100 & 100 & 50 & 50 \\
\vdots & 100 & 100 & 100 & 50 & 50 \\
\vdots & 100 & 100 & 100 & 50 & 50 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

The task is to design a 3 × 3 kernel for “locating” this pattern:

\[
p[n, m] = \begin{bmatrix}
50 & 50 & 50 \\
100 & 100 & 50 \\
100 & 100 & 50 \\
\end{bmatrix}
\]

In other words, we want to construct an operator that produces a large output value when it is centered over this “upper-right corner” pattern. The recipe for convolution tells us to rotate the pattern by \(\pi\) radians about its center to create \(p[-n, -m]\):

\[
p[-n, -m] = \begin{bmatrix}
50 & 100 & 100 \\
50 & 100 & 100 \\
50 & 50 & 50 \\
\end{bmatrix}
\]

The average weight in this 3 × 3 kernel is \(\frac{650}{9} \approx 72.222\), which is subtracted from each element:
The multiplicative factor may be ignored since it just scales the output of the convolution by this constant. Thus one realization of the unamplified $3 \times 3$ matched filter for upper-right corners is:

$$
\hat{m}[n,m] \cong \begin{bmatrix}
-1 & +1.25 & +1.25 \\
-1 & +1.25 & +1.25 \\
-1 & -1 & -1
\end{bmatrix}
$$

Though it is not really an issue any longer (given the advanced state of computing technology), it was once more convenient to restrict the weights in the kernel to integer values so that all calculations were performed by integer arithmetic. This may be achieved by redistributing the weights slightly. In this example, the fraction of the positive weights often is concentrated in the center pixel to produce the Prewitt corner detector:

$$
\hat{m}[n,m] \cong \begin{bmatrix}
-1 & +1 & +1 \\
-1 & +2 & +1 \\
-1 & -1 & -1
\end{bmatrix}
$$

Note that the upper-right corner detector contains a bipolar pattern that looks like a lower-left corner because of the rotation (“reversal”) inherent in the convolution. Because $\hat{m}$ is bipolar, so generally is the output of the convolution with the input $f[n,m]$. The linearity of convolution ensures that the output amplitude at a pixel is proportional to the contrast of the feature.

If the contrast of the upper-right corner is large and “positive,” meaning that the corner is much brighter than the dark background, the output at the corner pixel will be a large and positive extremum. Conversely, a dark object on a very bright background will produce a large negative extremum. The magnitude of the image shows the locations of features with either contrast. The output image may be thresholded to specify the pixels located at the desired feature.

This method of feature detection is not ideal. The output of this unamplified filter at a corner is the autocorrelation of the feature rather than the ideal 2-D discrete Dirac delta function. If multiple copies of the pattern with different contrasts are present in the input, it will be difficult or impossible to segment the desired features by thresholding the convolution alone. Another consequence of the unamplified matched filter is that features other than the desired pattern produce nonnull outputs, as shown in the output of the corner detector applied to a test object consisting of “E” at two different amplitudes as shown in the figure. The threshold properly locates the
upper-right corners of the bright “E” and one point on the sampled circle, but misses the corners of the fainter “E”. This shows that corners of some objects are missed (false negatives). If the threshold were set at a lower level to detect the corner of the fainter “E”, other pixels will be incorrectly identified as corners (false positives). A simple method for reducing misidentified pixels is considered in the next section.

Thresholding to locate features in the image: (a) $f[n,m]$, which is the nonnegative function with $0 \leq f \leq 1$; (b) $f[n,m]$ convolved with the “upper-right corner detector”, producing the bipolar output $g[n,m]$ where $-5 \leq g \leq 4$. The largest amplitudes occur at the upper-right corners, as shown in the image thresholded at level 4, shown in (c) along with the “ghost” of the original image. This
demonstrates that the upper-right corners of the high-contrast “E” and of the circle were detected, but corner of the low-contrast “E” was missed.

Kernels Commonly Used to Detect Particular Features

**Prewitt Edge Detectors:** Horizontal Edge (Vertical Transition of Gray)

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

**Vertical Edge (Horizontal Transition of Gray)**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>2</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>2</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+2</td>
<td></td>
</tr>
</tbody>
</table>

**Sobel Edge Detectors:** Horizontal Edge (Vertical Transition of Gray)

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>-2</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>+2</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

**Vertical Edge (Horizontal Transition of Gray)**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

**Line Detectors (One Pixel Wide)**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>+2</td>
<td>+2</td>
<td>+2</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>2</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

16.2.13 Normalization of Contrast of Detected Features

The recipe just developed allows creation of kernels for detecting pixels in neighborhoods that are “similar” to some desired pattern. However, the sensitivity of the process to feature contrast that was also demonstrated can significantly limit its usefulness. A simple modification to “normalize” the correlation measure can improve the classification significantly. Ernest Hall defined the normalized correlation measure $R[n, m]$: 
\[ R[n, m] = \frac{f[n, m] * h[n, m]}{\sqrt{\sum_{n,m} (f[n, m])^2} \sqrt{\sum_{n,m} (h[n, m])^2}} \]

where the sums in the denominator are over ONLY the elements of the kernel. The sum of the squares of the elements of the kernel \( h[n, m] \) results in a constant scale factor \( k \) and may be ignored, thus producing the formula:

\[ R[n, m] = k \left( \frac{f[n, m]}{\sum_{n,m} (f[n, m])^2} \right) * h[n, m] \]

where again the summation is ONLY over the size of the kernel. In words, this operation divides the convolution by the geometric sum of gray levels under the kernel and by the geometrical sum of the elements of the kernel. The modification to the filter makes the entire process shift variant and thus may not be performed by a simple convolution. The denominator may be computed by convolving \( (f[n, m])^2 \) with a uniform averaging kernel of the same size as the original kernel \( h[n, m] \) and then evaluating the square root

\[ R[n, m] = k \frac{f[n, m] * h[n, m]}{\sqrt{(f[n, m])^2 * s[n, m]}} \]

The upper-right corner detector with normalization is shown in the figure, where the features of both “E”s are located with a single threshold.
16.3 NONLINEAR FILTERS

The action of the nonlinear normalization of detected features using the same object: (a) $f[n,m] * \begin{bmatrix} -1 & 1 & 1 \\ -1 & 2 & 1 \\ -1 & -1 & -1 \end{bmatrix}$ convolved with the upper-right corner detector with normalization by the image amplitude, producing the bipolar output $g[n,m]$ where $-0.60 \leq g \leq +0.75$; (b) image after thresholding at $g = 0.6$, with a “ghost” of the original image, showing the detection of the upper-right corners of both “E”s despite the different image contrasts.

16.3 Nonlinear Filters

Any filtering operation that cannot be evaluated as a convolution must be either nonlinear or space variant, or both.

16.3.1 Median Filter

Probably the most useful nonlinear statistical filter is the local median, i.e., the gray value of the output pixel is the median of the gray values in a neighborhood, which is obtained by sorting the gray values in numerical order and selecting the middle value. To illustrate, consider the $3 \times 3$ neighborhood centered on the value “3” and the 9 values sorted in numerical order; the median value of “2” is indicated by the
box and replaces the “3” in the center of the window:

\[
\begin{array}{ccc}
1 & 2 & 6 \\
3 & 2 & 5 \\
1 & 5 & 2 \\
\end{array}
\]

The nonlinear nature of the median can be recognized by noting that the median of
the sum of two images is generally not equal to the sum of the medians. For example,
the median of a second 3 × 3 neighborhood is “3”

\[
\begin{array}{ccc}
4 & 5 & 6 \\
3 & 1 & 2 \\
2 & 4 & 3 \\
\end{array}
\]

⇒ ordered sequence is

\[1 \; 1 \; 2 \; 2 \; 3 \; 4 \; 4 \; 5 \; 6\]

The sum of the two medians is 2 + 3 = 5, but the sum of the two 3 × 3 neighborhoods
produces a third neighborhood whose median of “6”:

\[
\begin{array}{ccc}
5 & 7 & 12 \\
6 & 3 & 7 \\
3 & 9 & 5 \\
\end{array}
\]

⇒ ordered sequence is

\[3 \; 3 \; 5 \; 5 \; 6 \; 7 \; 7 \; 9 \; 12\]

confirming that the median of the sum is not the sum of the medians and that the
median is a nonlinear operation.

The median requires sorting and thus may not be computed as a convolution. Its
computation typically requires more time than a mean filter, but it has the advantage
of reducing the modulation of signals that vary or oscillate over a period less than the
width of the window while preserving the gray values of signals which are constant or
monotonically varying on a scale larger than the window size. This implies that the
variance of additive noise will be reduced by the median in a fashion similar to the
mean filter, while preserving sharp transitions in gray value. Also note that, unlike
the mean filter, all gray values generated by the median exist in the original image,
thus obviating the need for requantization.

The statistics of the median-filtered image depend on the probability density func-
tion of the input signal, including the deterministic part and any noise. Thus predic-
tions of the effect of the filter cannot be as specific as for the mean filter, i.e., given
an input image with known statistics (mean, variance, etc.), the statistics of the output
image are more difficult to predict. However, Frieden (Probability, Statistical
Optics, and Data Testing, Springer-Verlag, 1983, pp. 254-258). has analyzed the
statistical properties of the median filter by modeling it as a limit of a large number
of discrete trials of a binomial probability distribution (Bernoulli trials). The median
of N samples (odd number) for a set of gray values \(f_i\) taken from an input distribu-
tion with probability law (i.e. histogram) \( p_f [x] \) must be determined. Frieden applied
the principles of Bernoulli trials to determine the probability density of the median
of several independent sets of numbers. In other words, he sought to determine the
probability that the median of the \( N \) numbers \( \{ f_n \} \) is \( x \) by evaluating the median
of many independent such sets of \( N \) numbers selected from a known probability dis-
tribution \( p_f [x] \). Frieden reasoned that, for each placement of the median window, a
specific amplitude \( f_n \) of the \( N \) values is the median if three conditions are satisfied:

1. one of the \( N \) numbers satisfies the condition \( x \leq f_n < x + \Delta x \)
2. of the remaining \( N - 1 \) numbers, \( \frac{N-1}{2} \) exceed \( x \), and
3. \( \frac{N-1}{2} \) of the remaining numbers are less than \( x \).

The probability of the simultaneous occurrence of these three events is the proba-
bility density of the output of the median window. For an arbitrary \( x \), any one value
\( f_n \) must either lie in the interval \( (x \leq f < x + \Delta x) \), be larger than \( x \), or less than
\( x \). In other words, each trial has three possible outcomes. These conditions define a
sequence of Bernoulli trials with three outcomes, which is akin to the task of flipping
a “three-sided” coin where the probabilities of the three outcomes are not equal. In
the more familiar case, the probability that \( N \) coin flips with two possible outcomes
that have associated probability \( p \) and \( q \) will produce \( m \) “successes” (say, \( m \) heads) is:

\[
P_N [m] = \frac{N!}{(N-m)!m!} p^m (1-p)^{N-m}
\]

The formula is easy to extend to the more general case of \( N \) experiments with three
possible outcomes; the probability that the result yields \( m_1 \) instances of the first
possible outcome (say, “head #1), \( m_2 \) of the second outcome (“head #2”) and \( m_3 =
N - (m_1 + m_2) \) of the third (“tails”) is

\[
P_N [m_1, m_2, m_3] = \frac{N!}{m_1!m_2! (N - (m_1 + m_2))!} p_1^{m_1} p_2^{m_2} p_3^{m_3}
\]

where \( p_1, p_2, \) and \( p_3 = 1 - (p_1 + p_2) \) are the respective probabilities of the three
outcomes.

When applied to one sample of data, the median filter has three possible outcomes
whose probabilities are known:

1. the sample amplitude may be the median (probability \( p_1 \)),
2. the sample amplitude may be smaller than the median (probability \( p_2 \)), and
3. it may be larger than the median (probability \( p_3 \)).
\begin{align*}
p_1 &= P \left[ x \leq f_n \leq x + \Delta x \right] = p_f [x] \\
p_2 &= P \left[ f_n < x \right] = C_f [x] \\
p_3 &= P \left[ f_n > x \right] = 1 - C_f [x]
\end{align*}

where \( C_f [x] \) is the cumulative probability distribution of the continuous probability density function \( p_f [x] \):

\[
C_f [x] = \int_{-\infty}^{x} p_f [\alpha] \ d\alpha
\]

In this case, the distributions are continuous (rather than discrete), so the probability is the product of the probability density function \( p_{med} [x] \) and the infinitesimal element \( dx \). We substitute the known probabilities and the known number of occurrences of each into the Bernoulli formula for three outcomes:

\[
p_{med} [x] \ dx = \frac{N!}{\left( \frac{N-1}{2} \right)! \cdot \left( \frac{N-1}{2} \right)!} \cdot \frac{1}{1!} \cdot (C_f [x])^{\frac{N-1}{2}} \cdot (1 - C_f [x])^{\frac{N-1}{2}} \cdot p_f [x] \ dx
\]

If the window includes \( N = 3, 5, \) or \( 9 \) values, the following probability laws for the median result:

\[
N = 3 \implies p_{med} [x] \ dx = \frac{3!}{(1!)^2} \cdot (C_f [x])^{1} \cdot (1 - C_f [x])^{\frac{N-1}{2}} \cdot p_f [x] \ dx
\]

\[
N = 5 \implies p_{med} [x] \ dx = \frac{5!}{(2!)^2} \cdot (C_f [x])^{2} \cdot (1 - C_f [x])^{2} \cdot p_f [x] \ dx
\]

\[
N = 9 \implies p_{med} [x] \ dx = 630(C_f [x])^{4} \cdot (1 - C_f [x])^{4} \cdot p_f [x] \ dx
\]

**Example: Median Filter Applied to Uniformly Distributed Noise**

The statistical properties of the median will now be demonstrated for some simple examples of known probabilities. If the original pdf \( p_f [x] \) is uniform over the interval \([0, 1] \), then it may be written as a rectangle function:

\[
p_f [x] = RECT \left[ x - \frac{1}{2} \right]
\]
pdf of noise that is uniformly distributed over the interval [0, 1] and its associated cumulative probability distribution $F_c[x] = x \cdot \text{RECT}[x - \frac{1}{2}] + \text{STEP}[x - 1]$

The associated cumulative probability distribution may be written in several ways, including:

$$C_f[x] = \int_{-\infty}^{x} p_f(\alpha) d\alpha = x \cdot \text{RECT}[x - \frac{1}{2}] + \text{STEP}[x - 1]$$

so the product of the cumulative distribution and its complement is windowed by the rectangle to yield:

$$p_{\text{med}}[x] \ dx = \frac{N!}{((\frac{N-1}{2})!)^2} \left( x^{\frac{N-1}{2}} \cdot (1 - x)^{\frac{N+1}{2}} \right) \ \text{RECT} \left[ x + \frac{1}{2} \right] \ dx$$

The pdfs of the output of median filters for $N = 3, 5,$ and 9 are:

$$N = 3 \implies p_{\text{median}}[x] \ dx = 6 \left( x - x^2 \right) \ \text{RECT} \left[ x - \frac{1}{2} \right] \ dx$$

$$N = 5 \implies p_{\text{median}}[x] \ dx = 30 \left( x^4 - 2x^3 + x^2 \right) \ \text{RECT} \left[ x - \frac{1}{2} \right] \ dx$$

$$N = 9 \implies p_{\text{median}}[x] \ dx = 630 \left( x^8 - 4x^7 + 6x^6 - 4x^5 + x^4 \right) \ \text{RECT} \left[ x - \frac{1}{2} \right] \ dx$$

are compared to the pdfs of the output of the mean filters in the figure:
Comparison of pdfs of mean and median filter for uniform probability density function \( p_f[x] = \text{RECT}[x + \frac{1}{2}] \) for \( N = 3, 5, \) and 9. The pdf obtained from the mean filter is “taller” and “skinnier” than that from the median in all three cases, showing that mean filters reduce the variance more than the median filter. This advantage is offset by the edge-preserving property of the median filter.

Just like the mean filter, the maximum amplitude of \( p_{\text{median}}[x] \) increases and its “width” decreases as the number of input values in the median window increases (as \( N \uparrow \)). The calculated pdfs for the median and mean filters over \( N = 3 \) and \( N = 5 \) samples for input values from a uniform probability distribution are shown below to the same scale. Note that the output distributions from the mean filter are taller than for the median, which indicates that the median filter does a poorer job over averaging noise than the mean (Frieden determined that the SNR of the median filter is smaller than that of the mean by a factor of \( \log_e[2] \approx 0.69 \), so that there is a penalty in SNR of about 30% for the median filter relative to the averaging filter. Put another way, the standard deviation of the median of \( N \) samples decreases as \( \left( \sqrt{N} \cdot \log_e[2] \right)^{-1} \approx \frac{0.69}{\sqrt{N}} \) instead of as \( \frac{1}{\sqrt{N}} \). The lesser noise reduction of the median filter is offset by its ability to preserve the sharpness of edges.
Comparison of mean and median filter: (a) bitonal object \( f[m] \) defined over 1000 samples; (b) mean of \( f[m] \) over 25 samples, showing reduction in contrast with increasing frequency; (c) median of \( f[m] \) over 25 samples, which is identical to \( f[m] \); (d) \( f[m] + n[m] \), which is uniformly distributed over interval \([0, 1] \); (e) mean over 25 samples; (f) median over 25 samples. Note that the highest-frequency bars are (slightly) better preserved by the median filter.

Example: Median Filter Applied to Gaussian Noise

Probably the most important application of the median filter is to attenuate Gaussian noise (i.e., the gray values are selected from a normal distribution with zero mean) without blurring edges. The central limit theorem indicates that the statistical character of noise which has been generated by summing random variables from different distributions will be Gaussian in character. The probability distribution function is the Gaussian with mean value \( \mu \) and variance \( \sigma^2 \) normalized to unit area:

\[
p_f[x] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]
\]
The cumulative probability density of this noise is the integral of the Gaussian probability law, which is proportional to the error function:

\[ \text{erf}[x] \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \]

We can evaluate the cumulative density in terms of \( \text{erf}[x] \):

\[
C_f[x] = \int_{-\infty}^x p_c[x] \, dx \equiv \frac{1}{2} - \frac{1}{2\sqrt{2}\sigma} \text{erf}\left[\frac{x - \mu}{\sqrt{2}\sigma}\right] \quad \text{for } x \leq \mu
\]

\[
C_f[x] = \int_{-\infty}^x p_c[x] \, dx \equiv \frac{1}{2} + \frac{1}{2\sqrt{2}\sigma} \text{erf}\left[\frac{x - \mu}{\sqrt{2}\sigma}\right] \quad \text{for } x \geq \mu
\]

Therefore the probabilities of the different outcomes of the median filter are:

\[
p_{\text{med}}[x] \, dx = \frac{N!}{((N-1)/2)!)^2} \left( \frac{1}{2} + \frac{1}{2\sqrt{2}\sigma} \text{erf}\left[\frac{x - \mu}{\sqrt{2}\sigma}\right]\right)^{\frac{N-1}{2}} \cdot \left( \frac{1}{2} - \frac{1}{2\sqrt{2}\sigma} \text{erf}\left[\frac{x - \mu}{\sqrt{2}\sigma}\right]\right)^{\frac{N-1}{2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right] \, dx
\]

The error function is compiled and may be evaluated to plot the probability \( p_{\text{med}}[x] \)

The graphs illustrate the theoretical averaging effects of the mean and median filters on Gaussian noise. The graphs are plotted on the same scale and show the pdf of the original Gaussian noise (on the left) and the output resulting from mean and median
filtering over 3 pixels (center) and after mean and median filtering over 5 pixels (right). The calculated mean gray value and standard deviation for 2048 samples of filtered Gaussian noise yielded the following values:

\[
\begin{align*}
\mu_{in} &= 0.211 \\
\sigma_{in} &= 4.011 \\
\mu_3 - mean &= 0.211 \\
\sigma_3 - mean &= 2.355 \\
\mu_3 - median &= 0.225 \\
\sigma_3 - median &= 2.745
\end{align*}
\]

Effect of Window “Shape” on Median Filter

In the 2-D imaging case, the shape of the window over which the median is computed also affects the output image. For example, if the 2-D median is computed over a 5 × 5 window at the upper-right corner of a dark object on a bright background, the median will be the background value:

\[
\begin{array}{cccccc}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
+1 & +1 & +1 & 0 & 0 \\
+1 & +1 & +1 & 0 & 0 \\
+1 & +1 & +1 & 0 & 0 \\
\end{array}
\]

median of \( = 0 \)

The median calculated over a full square window (3 × 3, etc.) will convert bright pixels at outside corners of bright object to dark pixels, i.e., the corners will be clipped; it will also convert a dark background pixel at the inside corner of a bright object to a bright pixel. It will also eliminate lines less than half as wide as the window. Corner clipping may be prevented by computing the median over a window that only includes 9 values arrayed along horizontal and vertical lines:

\[
\begin{array}{cccc}
- & - & +1 & - \\
- & - & +1 & - \\
+1 & +1 & +1 & +1 \\
- & - & +1 & - \\
- & - & +1 & - \\
\end{array}
\]
If applied to the pixel in the corner, we obtain

\[
\begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
+1 & +1 & +1 & 0 & 0 \\
+1 & +1 & +1 & 0 & 0 \\
+1 & +1 & +1 & 0 & 0
\end{array}
\times
\begin{array}{cccc}
- & - & +1 & - & - \\
- & - & +1 & - & - \\
+1 & +1 & +1 & +1 & +1 \\
- & - & +1 & - & - \\
- & - & +1 & - & -
\end{array}
\]

\[
\begin{array}{cccc}
- & - & 0 & - & - \\
- & - & 0 & - & - \\
- & - & +1 & - & - \\
- & - & +1 & - & -
\end{array}
= \text{median of}\n\begin{array}{cccc}
+1 & +1 & +1 & 0 & 0 \\
- & - & +1 & - & - \\
- & - & +1 & - & -
\end{array}
= +1
\]

This pattern also is effective when applied to thin lines without eliminating them:

\[
\begin{array}{cccc}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
+1 & +1 & +1 & +1 & +1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}
\times
\begin{array}{cccc}
- & - & +1 & - & - \\
- & - & +1 & - & - \\
+1 & +1 & +1 & +1 & +1 \\
- & - & +1 & - & - \\
- & - & +1 & - & -
\end{array}
\]

\[
\begin{array}{cccc}
- & - & 0 & - & - \\
- & - & 0 & - & - \\
- & - & 0 & - & - \\
- & - & 0 & - & -
\end{array}
= \text{median of}\n\begin{array}{cccc}
+1 & +1 & +1 & +1 \\
- & - & +1 & - & - \\
- & - & +1 & - & -
\end{array}
= +1
\]

Other patterns of medians are also useful (Castleman, Digital Image Processing, Prentice-Hall, 1996, p. 249.)

### 16.3.2 Other Statistical Filters (Mode, Variance, Maximum, Minimum)

The statistical mode in the window (i.e., the most common gray level) is a useful operator on binary images corrupted by isolated noise pixels (“salt-and-pepper noise”).
The mode is found by computing a mini-histogram of pixels within the window and assigning the most common gray level to the center pixel. Rules must be defined if two or more gray levels are equally common, and particularly if all levels are populated by a single pixel. If two levels are equally populated, the gray level of center pixel is usually retained if it is one of those levels; otherwise one of the most common gray levels may be selected at random.

The variance filter $\sigma^2$ and standard deviation filter $\sigma$ replace the center pixel with the variance or standard deviation of the pixels in the window, respectively. The variance filtering operation is

$$g[x, y] = \sum_{\text{window}} (f[x, y] - \mu)^2$$

where $\mu$ is the mean value of pixels in the window. The output of a variance or standard deviation operation will be larger in areas where the image is busy and small where the image is smooth. The output of the $\sigma$-filter resembles that of the isotropic Laplacian, which computes the difference of the center pixel and the average of the eight nearest neighbors.

The Maximum or Minimum filter obviously replace the gray value in the center with the highest or lowest value in the window. The MAX filter will dilate bright objects, while the MIN filter erodes them. These provide the basis for the so-called morphological operators. A “dilation” (MAX) followed by an “erosion” (MIN) defines the morphological “CLOSE” operation, while the opposite (erosion followed by dilation) is an “OPEN” operation. The “CLOSE” operation fills gaps in lines and removes isolated dark pixels, while OPENING removes thin lines and isolated bright pixels. These nonlinear operations are useful for object size classification and distance measurements.

### 16.4 Adaptive Operators

In applications such as edge enhancement or segmentation, it is often useful to “change”, or “adapt” the operator based on conditions in the image. One example has already been considered: the nonlinear normalization used while convolving with a bipolar convolution kernel. For another example, it is possible to enhance differences in the direction of the local gradient (e.g. via a 1-D Laplacian) while averaging in the orthogonal direction. In other words, the operator used to enhance the edge information is determined by the output of the gradient operator. As another example, the size of an averaging neighborhood could be varied based on the statistics (e.g., the variance) of gray levels in the neighborhood.

In some sense, these adaptive operators resemble cascaded convolutions, but the resulting operation is not space invariant and may not be described by convolution with a single kernel. By judicious choice of algorithm, significant improvement of image quality may be obtained.
16.5 Convolution Revisited – Bandpass Filters

The parameters of a filter that determine its effect on the image are the size of the kernel and the algebraic sign of its coefficients. Kernels whose elements have the same algebraic sign are lowpass filters that compute spatial averages and attenuate the modulation of spatial structure in the image. The larger the kernel, the greater the attenuation. On the other hand, kernels that compute differences of neighboring gray levels will enhance the modulation of spatially varying structure while attenuating the brightness of constant areas. Note that the largest number of elements in a kernel with different algebraic signs is two; the spatial first derivative is an example.

We will now construct a hybrid of these two extreme cases that will attenuate the modulation of image structure that varies more slowly or rapidly than some selectable rate. In other words, the filter will pass a band of spatial frequencies and attenuate the rest of the spectrum; this is a bandpass filter. The bandpass filter will compute differences of spatial averages of gray level. For example, consider a 1-D image:

\[
f[n] = 1 + \sum_{i=0}^{2} \cos \left( \frac{2\pi n(2^i)}{128} \right) = 1 + \cos \left( \frac{2\pi n}{\infty} \right) + \cos \left( \frac{2\pi n}{128} \right) + \cos \left( \frac{2\pi n}{64} \right) + \cos \left( \frac{2\pi n}{32} \right)
\]

\[
= 2 + \cos \left( \frac{2\pi n}{128} \right) + \cos \left( \frac{2\pi n}{64} \right) + \cos \left( \frac{2\pi n}{32} \right)
\]

The spatial frequencies of the cosines are:

\[
\xi_0 = \frac{1}{\infty} = 0 \text{ cycles per sample} \implies X_0 = \infty
\]

\[
\xi_1 = \frac{1}{128} \simeq 7.8 \cdot 10^{-2} \text{ cycles per sample} \implies X_1 = 128 \text{ samples}
\]

\[
\xi_2 = \frac{1}{64} \text{ cycles per sample} \implies X_2 = 64 \text{ samples}
\]

\[
\xi_3 = \frac{1}{32} \text{ cycles per sample} \implies X_3 = 32 \text{ samples}
\]

This function is periodic over 128 samples, which is a common multiple of all of the finite-period cosines. The extreme amplitudes are +5 and +0.2466. Consider convolution of \( f[n] \) with several kernels; the first set are 3-pixel averagers whose weights sum to unity, therefore preserving the mean gray level of \( f[n] \):

\[
h_1[n] = \begin{bmatrix} 0 & +1 & 0 \end{bmatrix}
\]

\[
h_2[n] = \begin{bmatrix} +\frac{1}{4} & +\frac{1}{2} & +\frac{1}{4} \end{bmatrix}
\]

\[
h_3[n] = \begin{bmatrix} +\frac{1}{3} & +\frac{1}{3} & +\frac{1}{3} \end{bmatrix}
\]

Obviously, \( h_1[n] \) is the identity kernel, \( h_2 \) is a tapered averager that applies more weight to the center pixel, while \( h_3 \) is a uniform averager. Based on our experience with averaging filters, we know that \( g_1[n] = f[n] * h_1[n] \) must be identical to \( f[n] \), while the modulation of the output from \( h_2[n] \) will be reduced a bit in \( g_2 \) and somewhat
more in \( g_3 \). This expectation is confirmed by the computed maximum and minimum values:

\[
\begin{array}{cccc}
 f_{\text{max}} = 5 & (g_1)_{\text{max}} = 5 & (g_2)_{\text{max}} \simeq 4.987 & (g_3)_{\text{max}} \simeq 4.983 \\
 f_{\text{min}} \simeq 0.2466 & (g_1)_{\text{min}} \simeq 0.2466 & (g_2)_{\text{min}} \simeq 0.2564 & (g_3)_{\text{min}} \simeq 0.2596
\end{array}
\]

The mean gray values of these images are identical:

\[
\langle f \rangle = \langle g_1 \rangle = \langle g_2 \rangle = \langle g_3 \rangle = 2
\]

We can define a contrast factor based on these maximum and minimum values that is analogous to the modulation, except that the image is not sinusoidal:

\[
c_f \equiv \frac{f_{\text{max}} - f_{\text{min}}}{f_{\text{max}} + f_{\text{min}}}
\]

The corresponding factors are:

\[
\begin{array}{c|c|c|c|c}
 c_f = 0.906 & c_1 \simeq 0.906 & c_2 \simeq 0.9022 & c_3 \simeq 0.9019
\end{array}
\]

which confirms the expectation.

Now consider three 5-pixel averagers:

\[
\begin{align*}
 h_4[n] &= \begin{bmatrix} 0 & 0 & +1 & 0 & 0 \end{bmatrix} \\
 h_5[n] &= \begin{bmatrix} +\frac{1}{9} & +\frac{2}{9} & +\frac{3}{9} & +\frac{2}{9} & +\frac{1}{9} \end{bmatrix} \\
 h_6[n] &= \begin{bmatrix} +\frac{1}{5} & +\frac{1}{5} & +\frac{1}{5} & +\frac{1}{5} & +\frac{1}{5} \end{bmatrix}
\end{align*}
\]

Again, \( h_4 \) is the identity kernel that reproduces the modulation of the original image, \( h_2 \) is a tapered averager, and \( h_6 \) is a uniform averager. The computed maximum and minimum values for the images are:

\[
\begin{array}{cccc}
 f_{\text{max}} = 5 & (g_4)_{\text{max}} = 5 & (g_5)_{\text{max}} \simeq 4.967 & (g_6)_{\text{max}} \simeq 4.950 \\
 f_{\text{min}} \simeq 0.2466 & (g_4)_{\text{min}} \simeq 0.2466 & (g_5)_{\text{min}} \simeq +0.272 & (g_6)_{\text{min}} \simeq 0.285 \\
 \langle f \rangle = 2 & \langle g_4 \rangle = 2 & \langle g_5 \rangle = 2 & \langle g_6 \rangle = 2 \\
 c_f = 0.906 & c_4 = 0.906 & c_5 = 0.896 & c_6 = 0.891
\end{array}
\]

The average over a larger number of samples reduces the modulation further but does not affect the mean gray values.
Outputs from application of various averaging kernels to \( f[x] \). At this scale, the differences are not noticeable.

Now consider a kernel that is significantly wider:

\[
  h_7[n] = A_7 \cos \left( \frac{2\pi n}{64} \right) \cdot \text{RECT} \left( \frac{n}{32} \right)
\]

where the scale factor \( A_7 \) is usually used to normalize \( h_7[n] \) to unit area. The \( \text{RECT} \) function limits the support of the cosine to a finite width of 32 pixels, which is half the period of the cosine function. The kernel is wide and nonnegative; this is another example of a tapered averaging kernel that weights nearby pixels more heavily than more distant pixels. The corresponding uniform averaging kernel is:

\[
  h_8[n] = \frac{1}{32} \text{RECT} \left( \frac{n}{32} \right)
\]

To simplify comparison of the results of \( h_7 \) and \( h_8 \), we will set \( A_7 = \frac{1}{32} \) instead of a factor that ensures a unit area. The exact value of the scale factor will affect the output amplitudes and not the modulation. Based on the experience gained for \( h_1, h_2, \ldots, h_6 \), we expect that both \( h_7 \) and \( h_8 \) will diminish the modulation of spatially varying patterns, and that \( h_8 \) will have the larger effect. In fact, because the width of \( h_8 \) matches the period \( X_3 = 32 \), this cosine term will be attenuated to null amplitude. The effects on the amplitudes of the array are:

| \( f_{\text{max}} = 5 \) | \( (g_7)_{\text{max}} \approx 2.585 \) | \( (g_8)_{\text{max}} \approx 3.536 \) |
| \( f_{\text{min}} \approx 0.2466 \) | \( (g_7)_{\text{min}} \approx 0.593 \) | \( (g_8)_{\text{min}} \approx +1.205 \) |
| \( \langle f \rangle = 2 \) | \( \langle g_7 \rangle = 2 \) | \( \langle g_8 \rangle = 2 \) |
| \( c_f = 0.906 \) | \( c_7 \approx 0.627 \) | \( c_8 \approx 0.492 \) |

which again confirms the expectation that the uniform averager reduces the contrast to a greater degree than the tapered averager, but neither affects the mean gray value.
Outputs from filters of with 32, which are still averaging filters.

If the width of the $RECT$ function in the tapered averaging kernel $h_7$ is increased still further while the period of the constituent cosine function is retained, the resulting kernel includes some negative weights. For example:

$$h_9[n] = A_9 \cos \left[ \frac{2\pi n}{64} \right] RECT \left[ \frac{n}{48} \right]$$

The constant $A_9$ may be chosen so that the area of $h_9$ is unity ($A_9 \approx 0.06948$), or $A_9$ may be set to $\frac{1}{48}$, matching the normalization factor for the uniform averager:

$$h_{10}[n] = \frac{1}{48} RECT \left[ \frac{n}{48} \right]$$

which simplifies comparison of the resulting amplitudes. Kernel $h_9$ computes the same weighted average as $h_7$ in the neighborhood of the pixel, but then subtracts a weighted average of distant pixels from it; it computes differences of average amplitudes. The effects of these operations on the extrema are:

<table>
<thead>
<tr>
<th>$f_{\text{max}} = 5$</th>
<th>$(g_9)_{\text{max}} \approx 1.532$</th>
<th>$(g_{10})_{\text{max}} \approx 2.940$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\text{min}} \approx 0.2466$</td>
<td>$(g_9)_{\text{min}} \approx 0.118$</td>
<td>$(g_{10})_{\text{min}} \approx +1.420$</td>
</tr>
<tr>
<td>$\langle f \rangle = 2$</td>
<td>$\langle g_9 \rangle \approx 0.5997$</td>
<td>$\langle g_{10} \rangle = 2$</td>
</tr>
<tr>
<td>$c_f = 0.906$</td>
<td>$c_9 \approx 0.857$</td>
<td>$c_{10} \approx 0.349$</td>
</tr>
</tbody>
</table>

Kernel $h_{10}$ retained the mean value but further attenuated the contrast by pushing the amplitudes toward the mean. However, the difference-of-averages kernel $h_9$ actually increased the contrast and decreased the mean value.
Outputs from filters with impulse responses $h_9[n] = \frac{1}{48} RECT\left[\frac{n}{48}\right] \cdot \cos\left[2\pi \frac{n}{64}\right]$, which now has some negative amplitudes.

This trend may be continued by increasing the width of the RECT and using equal scale factors:

$h_{11}[n] = \frac{1}{64} \cos\left[2\pi \frac{n}{64}\right] RECT\left[\frac{n}{64}\right]$

$h_{12}[n] = \frac{1}{64} RECT\left[\frac{n}{64}\right]

Because the width of the RECT matches the period of the cosine in $h_1$, it may not be normalized to unit area. The extrema of these two processes are:

<table>
<thead>
<tr>
<th>$f_{\text{max}} = 5$</th>
<th>$(g_{11})_{\text{max}} \approx 0.712$</th>
<th>$(g_{12})_{\text{max}} = 2 + \frac{2}{\pi} \approx 2.637$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\text{min}} \approx 0.2466$</td>
<td>$(g_{11})_{\text{min}} \approx -0.511$</td>
<td>$(g_{12})_{\text{min}} \approx +1.364$</td>
</tr>
</tbody>
</table>

$\langle f \rangle = 2 \quad \langle g_{11} \rangle \approx 0 \quad \langle g_{12} \rangle = 2$

$c_f = 0.906 \quad c_{11} \approx 6.08 \quad c_{12} \approx 0.318$

The uniform averager $h_1$ continues to push the amplitudes toward the mean value of 2 and decreases the contrast, while the mean amplitude generated by the difference of averages kernel is now zero, which means that the minimum is less than zero.

Note that the output $g_{11}[n]$ looks like the kernel $h_{11}[n]$; in other words, the portion of $f[n]$ that was transmitted to $g_{11}[n]$ largely is a cosine of period 64. The distortion is due to cosines at other frequencies.
Outputs from convolutions with $h_{11}[n] = \frac{1}{64} \text{RECT} \left[ \frac{n}{64} \right] \cdot \cos \left[ 2\pi \frac{n}{64} \right]$ and $h_{12}[n] = \frac{1}{64} \text{RECT} \left[ \frac{n}{64} \right]$. The bipolar impulse response has zero area and “blocks” the constant part of the signal.

Now, consider filters whose widths are equal to the period of $f[n]$:

$$h_{13}[n] = \cos \left[ 2\pi \frac{n}{64} \right] \cdot \frac{1}{128} \text{RECT} \left[ \frac{n}{128} \right]$$
$$h_{14}[n] = \frac{1}{128} \text{RECT} \left[ \frac{n}{128} \right].$$

The figures of merit for the gray values of these arrays are:

<table>
<thead>
<tr>
<th></th>
<th>$f_{\text{max}}$</th>
<th>$(g_{13})_{\text{max}}$</th>
<th>$(g_{14})_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\text{min}} \approx 0.2466$</td>
<td>$(g_{13})_{\text{min}} \approx -0.5$</td>
<td>$(g_{14})_{\text{min}} = +2$</td>
<td></td>
</tr>
<tr>
<td>$\langle f \rangle = 2$</td>
<td>$\langle g_{13} \rangle = 0$</td>
<td>$\langle g_{14} \rangle = 2$</td>
<td></td>
</tr>
<tr>
<td>$c_f = 0.906$</td>
<td>$c_{13} = \infty$</td>
<td>$c_{14} = 0.0$</td>
<td></td>
</tr>
</tbody>
</table>

The output of the bipolar kernel $h_{13}$ is a sinusoid with period 64 and zero mean, while that of the averager $h_{14}$ is the constant average value of $f[n]$. Note that the contrast parameter $c_{13}$ is undefined because $f_{\text{min}} = -f_{\text{max}}$, while $c_{14} = 0$. 
Outputs resulting from convolution with \( h_{13} = \frac{1}{128} \text{RECT} \left[ \frac{n}{128} \right] \cdot \cos \left[ \frac{2\pi n}{64} \right] \) and \( h_{14} = \frac{1}{128} \text{RECT} \left[ \frac{n}{128} \right] \); the former is a bandpass filter and the latter is a lowpass filter.

To summarize, kernels that compute differences of averages are wide and bipolar, and typically yield bipolar outputs. As the width of a difference-of-averages kernel is increased, the output resembles the kernel itself to a greater degree, which is bipolar with zero mean. On the other hand, increasing the width of an averaging operator results in outputs that approach a constant amplitude (the average value of the input); this constant is a cosine with infinite period. The difference-of-averages operator rejects BOTH slowly and rapidly varying sinusoids, and preferentially passes a particular sinusoidal frequency or band of frequencies. Thus, differences of averages operators are called bandpass filters.

Kernels of bandpass filters are wide, bipolar, and resemble the signal to be detected.

### 16.6 Implementation of Filtering

#### 16.6.1 Nonlinear and Shift-Variant Filtering

Special-purpose hardware (array processors) is readily available for implementing linear operations. These contain several memory planes and can store multiple copies of the input. By shifting the addresses of the pixels in a plane and multiplying the values by a constant, the appropriate shifted and weighted image can be generated. These are then summed to obtain the filtered output. A complete convolution can be computed in a few clock cycles.

Other than the mean filter, the statistical filters are nonlinear, i.e., the gray value of the output pixel is obtained from those of the input pixel by some method other than multiplication by weights and summing. In other words, they are not convolutions and thus cannot be specified by a kernel. Similarly, operations that may be linear (output is a sum of weighted inputs) may use different weights at different locations in the image. Such operations are shift-variant and must be specified by...
different kernels at different locations. Nonlinear and shift-variant operations are computationally intensive and thus slower to perform unless special single-purpose hardware is used. However, as computer speeds increase and as prices fall, this is becoming less of a problem. Because of the flexibility of operations possible, nonlinear shift-variant filtering is a very active research area.

16.7 Neighborhood Operations on Multiple Images

16.7.1 Image Sequence Processing

It should now be obvious that we can combine the gray levels in neighborhoods of the input pixel in multiple images to obtain the output image $g[x, y]$. The multiple copies of $f[x, y]$ may have been spread over time (e.g., video), over wavelength (e.g., RGB images), or some other parameter. In these cases, the image and kernel are functions of three coordinates.

![Schematic of neighborhood operation on multiple images that differ in some other characteristic (time, wavelength, or focus depth).](image)

$g[x, y] = f[x, y, \lambda_1] \ast h[x, y] - f[x, y, \lambda_2] \ast h[x, y]$

16.7.2 Spectral + Spatial Neighborhood Operations

Additive noise is a common corrupter of digital images and may disrupt classification algorithms based on gray-level differences, e.g., in multispectral differencing to segment remote-sensing images. The noise can be attenuated by combining spatial averaging and spectral differencing, i.e.,
where

\[ h[x, y] = \frac{1}{9} \begin{bmatrix} +1 & +1 & +1 \\ +1 & +1 & +1 \\ +1 & +1 & +1 \end{bmatrix} \]

Since convolution and subtraction are linear, the order of operations can be interchanged:

\[ g[x, y] = (f[x, y, \lambda_1] - f[x, y, \lambda_2]) * h[x, y] \]

These can be combined into a single 3-D operation using a 3-D kernel \( h[x, y, \lambda] \)

\[ g[x, y] = f[x, y, \lambda_i] * h[x, y, \lambda_i] \]

where

\[ h[x, y, \lambda_i] = \frac{1}{9} \begin{bmatrix} +1 & +1 & +1 \\ +1 & +1 & +1 \\ +1 & +1 & +1 \end{bmatrix}, \begin{bmatrix} +1 & +1 & +1 \\ +1 & +1 & +1 \\ +1 & +1 & +1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \]
Chapter 17
Shape-Based Operations

An shape-based operation identifies or acts on groups of pixels that belong to the same object or image component. We have already seen how components may be identified on the basis of pixel gray level, color (multispectral gray level), or features such as edges, corners, texture, etc.. The output of an shape-based operation may not be an image at all, but rather a description of the objects, locations, etc. in the image.

Applications:

- image segmentation
- image description
- image compression

One application of both point and local neighborhood operations is image segmentation, i.e., classification of pixels that belong in one group. In the simple cases considered thus far, pixels were segmented by identifying clusters in a feature space (e.g., the multidimensional color histogram). Clustered pixels are assumed to belong to the same object class. This approach is appropriate in multispectral classification (e.g., ground-cover classification in remote sensing) or when regions may be distinguished by convolution with a particular kernel (e.g., patterned or textured regions). However, certain important problems may be more readily attacked by other means. For example, suppose that we want to build a machine vision system to read ideal binary text (i.e., black text on white background); individual characters or words must be segmented and identified. The histogram feature-space approach is not appropriate because all text pixels have the same gray level. Pixel classification (i.e., assigning membership of pixels to specific letters or words) must be based on a different criterion such as the shape of a group of adjoining pixels with similar gray levels.

This type of problem leads naturally to the concept of shape-based processing, where groups of pixels (objects) are processed as a whole. In the text processing example, pixels that belong to an individual text character are connected, i.e., adjacent pixels of the same gray level belong to the same character, while pixels with the same
gray level that are not connected belong to different characters. In this discussion, we will ignore problems with overdots (e.g., in lower case $i$ and $j$), which would be identified as separate characters.

Shape-based processing is also used for symbolic image description. Once pixels have been identified as connected, simpler shape descriptors may be derived by structural operations such as thinning to find the object skeleton, border-following, computation of moments of objects, and morphological transformations. The discussions of these operations will assume that the images are binary; the principles may be applied to gray-level imagery with some difficulty.

### 17.1 Pixel Connectivity

In a **binary** image, the component objects are clusters of foreground pixels (white) on a background (black). Two pixels with the same gray level that are adjacent or that are linked through an unbroken chain of foreground pixels are said to be connected and are assumed to belong to the same object. Unconnected foreground pixels must belong to different objects. The concepts of connectivity and adjacency must be well defined before a program may be written to identify and distinguish the foreground objects in the image.

A pixel on a rectangular grid are said to be four-connected or eight-connected when it has the same properties (gray level) as one of its nearest four or eight neighbors.

Definition of pixel connectivity: the four dark pixels are “four-connected” to pixel A; the eight dark pixels are “eight-connected” to pixel B.

Pixel A is 4-connected and Pixel B is 8-connected to dark pixels. Some questions may arise about whether some groups of connected pixels are connected to other groups. The picture below is composed of dark foreground objects on a bright background. How many dark objects are there? Regardless of the connectivity of the foreground and background, the dark pixels on the right compose a single foreground object which divides the background into two distinct regions. The number of foreground objects on the left depends on the definition of connectivity. If both foreground and background are considered to be 4-connected, the dark foreground pixels on the left actually compose four distinct objects and the background has two distinct components. Of course, if there are four distinct 4-connected foreground objects, then all background pixels should belong to a single object. Thus if the foreground is 4-connected, the background should be 8-connected. If the foreground is 8-connected, then there is a single foreground object on the left which divides the background into two components (inside and outside), and thus the background should be 4-connected.
17.2 Image Labeling

The membership of a pixel is specified by assigning a specific label to all pixels belonging to that component. Algorithms have been developed to automatically assign labels and the labeled clusters may then be trivially segmented by histogram thresholding. One simple technique requires two row-by-row scans of a binary image. When a foreground pixel is encountered during the first scan, its neighbors of each pixel are examined as shown:

The shaded pixel is to be labeled; if the object is assumed to be 4-connected, the already assigned labels of pixels B and C are checked. If either is also foreground, then the shaded pixel must belong to the same foreground object and receives the same label. If neither B nor C is foreground, then the shaded pixel is assigned a new label. If both B and C are foreground, but with different labels, then the shaded pixel connects two previously unconnected regions; it is assigned one of the labels and the equivalence of the two labels is noted in an equivalence table. During the
second scan, all equivalent labels are redefined to generate the final labeled image. If
the foreground is assumed to be 8-connected, the same procedure is followed, but the
label of pixel A is also considered when labels are assigned.

After labeling the components, image segmentation is quite trivial; different labels
may be used like different gray levels to segment the image by thresholding.

17.2.1 Example:

17.2.2 Border Coding

It is often useful to label the border pixels of objects in the image, which may be
defined in several ways. Recall that the connectivity of the foreground and background
must be complementary (4-C foreground ⇒ 8-C background). The border may be
defined as:

1. pixels belonging to the foreground object that are adjacent to background pixels,
or

2. pixels belonging to the background that are adjacent to foreground pixels, or
3. the crack between foreground and background.

The borders must define a closed curve, which be specified by its beginning (e.g., upper left corner) and the pixel-by-pixel direction around the object. A common notation for specifying the border code is to assign a direction code at each pixel using this numbering system:

```
 3 2 1
 4 * 0
 5 6 7
```

where the asterisk indicates the edge pixel being examined. This window is placed at the location of the upper-left-most edge pixel in the image of the edge. The number specifies the direction to the next pixel around the edge in a counterclockwise direction. The next edge pixel encountered around this loop is encoded by the number; the code is easily remembered; the \( m^{th} \) pixel is in the direction \( 45^\circ \cdot m \). The border code must be consistent with the connectivity of the region; if the border is defined by 4-connected foreground pixels, then only codes 0, 2, 4, and 6 are allowed and only 2 bits are required to store the code; if 8-connected, all 8 directions are allowed and 3 bits are needed. The border code is often called a chain code. If using definition [3], the four directions of the resulting crack code are specified in an identical fashion and requires 2 bits per segment.

Border coding is a useful tool in image recognition and compression, i.e., it reduces the number of bits required to store/transmit the image. Binary foreground objects may be completely specified by the border codes with fewer bits than storing the locations of each foreground pixel.

17.3 Border Operations

17.3.1 Contraction/erosion and expansion/dilation

Bright objects foreground in binary images may be shrunk by deleting (turning to black) the border pixels of the foreground (as defined in [1] above) or expanded by converting the border pixels of the background to white (foreground). These two operations (erosion and dilation) are the basic components of image morphology. This operations may be applied to a number of imaging problems, including noise removal, image compression, and image specification. This field of image processing has generated recent interest because of its applications to machine vision problems.

Most morphological operations are based on the two basic operations of erosion and dilation. In the simplest cases, erosion transforms all object pixels that are adjacent to background pixels to background. Dilation is the opposite; background pixels adjacent to object pixels are transformed to object pixels. In the more general case, a shape function \( p [x, y] \) is applied to objects in the image and determines which pixels would be converted. The shape function is often called the structuring element
or probe, and must specify the origin of coordinates, i.e., the pixel \( p[0, 0] \). The erosion of the image is the set of pixels defining those locations of \( p[0, 0] \) for which \( p[x, y] \) fits wholly within the image object. The dilation of the image is the set of pixels defined by the locations of the coordinate origin of the structure element where the element touches any object pixels in the image. The generalized erosion and dilation are identical to the simple definitions if the structure element is defined as the single pixel at the origin.

In the remainder of this discussion, we will denote the erosion of an image by the symbol “\( \bar{\Box} \)”, so that the erosion of \( f[x, y] \) by the “probe” function \( p[x, y] \) is:

\[
f[x, y] \bar{\Box} p[x, y],
\]

The dilation will be represented by the symbol “\( \bigcirc \)”, so that the dilation of \( f[x, y] \) by the “probe” function \( p[x, y] \) is:

\[
f[x, y] \bigcirc p[x, y].
\]

In general, the operation of dilation commutes but erosion does not, i.e.,

\[
f[x, y] \bigcirc p[x, y] = p[x, y] \bigcirc f[x, y]
\]

\[
f[x, y] \bar{\Box} p[x, y] \neq p[x, y] \bar{\Box} f[x, y]
\]

The morphological dilation and erosion operations are shift-invariant, because a translation of the object results in an identical translation of the erosion or dilation. However, they are not linear, i.e., the erosion of the sum of two images is not the sum of the erosions, and thus may not be computed by convolutions, but rather by determining if each pixel satisfies the requisite properties for the operation. For this reason, morphological operations tend to be computationally intensive, to the point where hours of CPU time may be required on general-purpose computers. However, they may be quickly computed in the binary case by evaluating the gray-scale cross-correlation of the binary image and the binary “probe” function (sometimes called the “structuring element”) The gray-scale crosscorrelation is thresholded; the dilation is obtained by thresholding just above the minimum value and erosion by thresholding at a level just below the maximum.
17.4 Cascaded Morphological Operations – “Opening” and “Closing”

The general erosion and dilation operators shrink and expand bright foreground objects. They are near-inverses of each other, but if an erosion removes an object consisting of an isolated foreground pixel, a subsequent dilation will not recreate the original object. Because they are not exact inverses, the cascade of erosion followed by dilation will yield a different result than a dilation followed by erosion. The former
is the morphological opening of the image because it will open up regions which are smaller than the structuring element. The opening operation smooths the boundary of bright binary objects while breaking apart any objects connected by thin lines. The cascade of dilation-erosion is the morphological closing of the image and will fill in areas of the foreground which are smaller than the structuring element. The closing operator fills up holes within or spaces between bright objects in the image.

\[
\text{Opening} = (f[x, y] \odot p[x, y]) \boxdot p[x, y]
\]

\[
\text{Closing} = (f[x, y] \oplus p[x, y]) \odot p[x, y]
\]

The opening and closing operations may be further cascaded or combined in various ways to perform useful operations. For example, the difference between the set of pixels that belong to the object and the set resulting from erosion with an averaging-like element will yield the boundary pixels of the object:

\[
\text{Boundary} = (f[x, y] \odot p[x, y]) - f[x, y] \quad \text{for} \quad p[x, y] = \begin{bmatrix}
+1 & +1 & +1 \\
+1 & +1 & +1 \\
+1 & +1 & +1 \\
\end{bmatrix}
\]

The so-called “hit-or-miss” operator is obtained by computing the erosion of the image and the dilation of the complementary image, and then finding the intersection of the two:

\[
\text{Hit-or-Miss} = (f[x, y] \odot p[x, y]) \cap ((1 - f[x, y]) \oplus p[x, y])
\]

\[
= (f[x, y] \odot p[x, y]) \cap \left(\hat{f}[x, y] \oplus p[x, y]\right)
\]

where \(\hat{f}\) denotes the complement image to \(f[x, y]\), i.e.,

\[
\hat{f}[x, y] = 1 - f[x, y]
\]

The concepts of morphological operations are presently being extended to gray-level (non-binary) images.

### 17.5 Applications of Morphological Operations

#### 17.5.1 Noise Removal

Additive noise in a binary image is often called salt-and-pepper; dark pixels are sprinkled in the white foreground and white pixels in the dark background. Isolated dark pixels may be removed by sequentially expanding and contracting borders of the
foreground; the complementary cycle will remove isolated white pixels.

\[
\begin{array}{cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

Original Binary Image

\[
\begin{array}{cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

After Erosion

\[
\begin{array}{cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

After Subsequent Dilation

\subsection{17.5.2 Medial Axis Transform}

When subjected to repeated erosion operations, the objects in the image will gradually shrink to the point where any further erosions will cut the object into discontinuous parts. The image obtained by assigning the number of erosions required to reach that point to each remaining pixel is the medial axis of the object, and the ensemble of those pixels is the skeleton of the object. These images are very useful for identifying components of the image and/or their orientation, i.e., image representation, and to reduce the number of data bits necessary to specify the representation, i.e., image
Shrinking and expansion operations are special cases of a more general field of image processing operations known as morphological operations, which specify the pixels that are inscribed by or circumscribed by other user-defined shapes (probes or structure elements).
17.6 Binary Morphological Operations

17.6.1 Horizontal Structuring Element

Bitonal morphological erosion, dilation, opening (dilation of erosion), and closing (erosion of dilation) for horizontal “structuring element” (probe function). Note that the horizontal “spaces” in the characters are filled in the closing.
17.6.2 Vertical Structuring Element

Bitonal morphological erosion, dilation, opening (dilation of erosion), and closing (erosion of dilation) for vertical “structuring element” (probe function). Note that the vertical “spaces” in the characters are filled in the closing.

Thin horizontal structures are removed by opening with vertical structure element. Vertical holes in object are filled by closing with vertical structure element.
Chapter 18

Geometric Operations

To this point, the image processing operations have computed the gray value (digital count) of the output image pixel based on the gray values of one or more input pixels; in other words, the operation “changed” the gray value to something new. Geometrical image processing operations are fundamentally different; instead of modifying the gray values of pixels, they redefine the pixel locations without changing their values (except to interpolate the values to the new pixel grid). In other words, geometrical operations change the spatial relationships between image pixels to correct distortions due to recording geometry, scale changes, rotations, perspective (keystoning), or due to curved or irregular object surfaces. In this section, we will define the procedures for specifying and implementing a range of geometrical operations.

In theory, it is possible (though exhausting!) to describe geometric transformations via a lookup table of input/output coordinates, i.e., the table would specify new output coordinates \([x', y']\) for each input location \([x, y]\). Equivalently, the coordinate lookup table could specify the input coordinates \([x, y]\) that map to a specific output pixel \([x', y']\). For an \(N \times M\) image, such a lookup table would contain \(N \cdot M\) ordered pairs (\(\Rightarrow\) 262,144 pairs for a 512 \(\times\) 512 image). The lookup table can specify any
arbitrary geometric transform, i.e., input pixels in the same neighborhood may move to locations that are very far apart in the output image. Besides using large blocks of computer memory, coordinate lookup tables are more general than usually necessary. In realistic applications, neighboring pixels in the input image will usually remain in close proximity in the output image, i.e., their coordinates will be transformed in similar manner. Such coordinate mappings are also called rubber-sheet transformations or image warping and may be specified by a set of parametric equations that specify the output coordinates for a given input position:

\[
x' = \alpha [x, y] \\
y' = \beta [x, y] \\
\Rightarrow O \{f [x, y]\} = f [x', y'] = f [\alpha [x, y], \beta [x, y]].
\]

The gray level \(f\) at the location \([x, y]\) is transferred to the new coordinates \([x', y']\) to create the output image \(f[x', y']\). This sequence of operations equivalent to defining a “new” image \(g[x, y]\) in the same coordinates \([x, y]\) but with different gray levels; hence:

\[
g[x, y] = f[\alpha [x, y], \beta [x, y]].
\]

To preserve the arrangement of pixel gray values in neighborhoods, the transformation of the continuous coordinates should be continuous, meaning that adjacent pixels in the input should be adjacent in the output. This means that the derivative of the coordinate transformation must be finite everywhere, so that the transformation is “smooth.” A power series of the input coordinates for positive powers satisfies this requirement:

\[
x' = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} a_{nm} x^n y^m = a_{00} + a_{10} x + a_{01} y + a_{11} xy + a_{20} x^2 + \cdots
\]

\[
y' = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} b_{nm} x^n y^m = b_{00} + b_{10} x + b_{01} y + b_{11} xy + b_{20} x^2 + \cdots
\]

In practice, the infinite series are truncated, thus limiting the range of possible transformations that may be described by the series. The order at which the series is truncated is thus determined by the transformation that must be allowed. In many cases, only four terms are necessary in each series, so that \(a_{nm} = b_{nm} = 0\) for \(n\) or \(m \geq 2:\)

\[
x' = a_{00} + a_{10} x + a_{01} y + a_{11} xy
\]

\[
y' = b_{00} + b_{10} x + b_{01} y + b_{11} xy
\]

Such a transformation is called \textit{bilinear}. There are eight unknown coefficients in the transformation, and thus eight independent equations are needed to find a solution. Knowledge of the coordinate transformation of the vertices of a quadrilateral is sufficient to find a solution for this transformation. In other words, knowledge of the
action on four locations

\[
\begin{align*}
[x_1, y_1] & \rightarrow [x'_1, y'_1] \\
[x_2, y_2] & \rightarrow [x'_2, y'_2] \\
[x_3, y_3] & \rightarrow [x'_3, y'_3] \\
[x_4, y_4] & \rightarrow [x'_4, y'_4]
\end{align*}
\]

will allow calculation of the eight coefficients. The problem may be cast in matrix notation where the known inputs \([x_i, y_i]\) and outputs \([x'_i, y'_i]\) are arranged as column vectors in the matrices \(X\) and \(X'\), respectively, and the unknown coefficients \(a_{ij}\) and \(b_{ij}\) are the rows of the matrix \(A\) in the expression:

\[
A \cdot X = X'
\]

\[
\begin{bmatrix}
a_{00} & a_{10} & a_{01} & a_{11} \\
b_{00} & b_{10} & b_{01} & b_{11}
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 1 & 1 \\
x_1 & x_2 & x_3 & x_4 \\
y_1 & y_2 & y_3 & y_4 \\
x_1y_1 & x_2y_2 & x_3y_3 & x_4y_4
\end{bmatrix}
= 
\begin{bmatrix}
x'_1 & x'_2 & x'_3 & x'_4 \\
y'_1 & y'_2 & y'_3 & y'_4 \\
x'_1y_1 & x'_2y_2 & x'_3y_3 & x'_4y_4
\end{bmatrix}
\]

If \(X\) is square (and if no three of the known points lie on the same straight line), then the inverse matrix \(X^{-1}\) exists; the coefficients \(a_{ij}\) and \(b_{ij}\) may be found via a straightforward calculation:

\[
(A \cdot X) X^{-1} = X' \cdot X^{-1}
\]

\[
\Rightarrow A = X' \cdot X^{-1}
\]

The four known vertices in the input and output images are sometimes called “control points”. More control points may be used in a bilinear fit, thus making the problem “overdetermined” (more equations than unknowns). A unique solution of an overdetermined problem may not exist if there is uncertainty (“noise”) in the data. Under such conditions, either a least-squares solution is generated or the control points are applied locally to find local transformations for different sections of the image. If the distortion cannot be adequately represented by a power series with eight coefficients, then more than four control points are required.
CHAPTER 18 GEOMETRIC OPERATIONS

18.1 Least-Squares Solution for Transformation

The procedure for computing the least-squares solution for the coefficients of a geometric transformation is quite easy in matrix notation. For example, consider a geometric transformation which adds a constant coordinate translation and magnifications in the orthogonal directions:

\[
\begin{align*}
  x' &= a_{00} + a_{10} x + a_{01} y \\
  y' &= b_{00} + b_{10} x + b_{01} y
\end{align*}
\]

where the coefficients \( a_{ij} \) and \( b_{ij} \) must be calculated. The system of equations has six unknown quantities and so requires six equations to obtain a solution. Since each control point (input-output coordinate pair) yields equations for both \( x \) and \( y \), three control points are needed. If more control points (say five) are available and consistent, the extras may be ignored and the the matrix inverse computed as before. If the positions of the control points are uncertain, the equations will be inconsistent and the matrix inverse will not exist. Under these conditions, the additional control points will improve the estimate of the transformation. The computation of the coefficients which minimizes the squared error in the transformation is called the least-squares solution, and is easily computed using matrix algebra as a pseudoinverse.

If we have five known control point pairs, the matrix transformation is composed of the 2 row by 3 column matrix \( A \), the 3 row by 5 column matrix \( X \) and the 2 row by 5 column matrix \( X' \):

\[
A \cdot X = X'
\]

\[
\begin{bmatrix}
  a_{00} & a_{10} & a_{01} \\
  b_{00} & b_{10} & b_{01}
\end{bmatrix}
\begin{bmatrix}
  1 & 1 & 1 & 1 & 1 \\
  x_1 & x_2 & x_3 & x_4 & x_5 \\
  y_1 & y_2 & y_3 & y_4 & y_5
\end{bmatrix}
= \begin{bmatrix}
  x'_1 & x'_2 & x'_3 & x'_4 & x'_5 \\
  y'_1 & y'_2 & y'_3 & y'_4 & y'_5
\end{bmatrix}
\]

If \( X \) were square (and invertible), we would compute \( X^{-1} \) to find \( A \) via:

\[
A = X' \cdot X^{-1}
\]
Because $X$ is NOT square in this case, its inverse $X^{-1}$ does not exist. However, we may evaluate the pseudoinverse of $X$ that constructs the solution for $A$ that minimizes the mean-squared error in the transformation. The process is actually quite simple via these steps for the general case where $X$ has $p$ rows and $q$ columns ($p = 3$ and $q = 5$ in the example above).

1. multiply both sides of the equation from the right by the transpose matrix $X^T$, which is obtained from $X$ by exchanging the rows and columns to obtain a matrix with $q$ rows and $p$ columns. The result is:

\[(A \cdot X) \cdot X^T = X^T \cdot X\]

2. The associativity of vector multiplication allows the second and third matrices on the left-hand side to be multiplied first:

\[(A \cdot X) \cdot X^T = A \cdot (X \cdot X^T)\]

3. The matrix $X \cdot X^T$ is $p \times p$ square. In the example above, the product of the two matrices is $3 \times 3$ square:

\[
X \cdot X^T = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
x_1 & x_2 & x_3 & x_4 & x_5 \\
y_1 & y_2 & y_3 & y_4 & y_5
\end{bmatrix}
\begin{bmatrix}
1 & x_1 & y_1 \\
1 & x_2 & y_2 \\
1 & x_3 & y_3 \\
1 & x_4 & y_4 \\
1 & x_5 & y_5
\end{bmatrix}
\]

\[
= \begin{bmatrix}
5 & \sum_{n=1}^{5} x_n & \sum_{n=1}^{5} y_n \\
\sum_{n=1}^{5} x_n & \sum_{n=1}^{5} (x_n)^2 & \sum_{n=1}^{5} x_ny_n \\
\sum_{n=1}^{5} y_n & \sum_{n=1}^{5} x_ny_n & \sum_{n=1}^{5} (y_n)^2
\end{bmatrix}
\]

Since $X \cdot X^T$ is square, its inverse $(X \cdot X^T)^{-1}$ may exist (we can guarantee its existence by requiring that no three of the control points lie on a straight line). Under this condition, then we can multiply the left-hand side from the right by this inverse; the result is the desired matrix of coefficients $A$:

\[A \cdot (X \cdot X^T) \cdot (X \cdot X^T)^{-1} = A\]

4. If we perform the same series of steps on the right-hand side, we obtain the
desired formula:

\[ A = (X' \cdot X'^T) \cdot (X \cdot X^T)^{-1} \]
\[ = X' \cdot (X'^T \cdot (X \cdot X^T)^{-1}) \]

\[
\begin{bmatrix}
  a_{00} & a_{10} & a_{01} & a_{11} \\
  b_{00} & b_{10} & b_{01} & b_{11}
\end{bmatrix}
\]

\[
\left( \begin{bmatrix}
  1 & x_1 & y_1 & x_1 y_1 \\
  1 & x_2 & y_2 & x_2 y_2 \\
  1 & x_3 & y_3 & x_3 y_3 \\
  1 & x_4 & y_4 & x_4 y_4 \\
  1 & x_5 & y_5 & x_5 y_5
\end{bmatrix} \cdot \begin{bmatrix}
  1 & 1 & 1 & 1 & 1 \\
  x_1 & x_2 & x_3 & x_4 & x_5 \\
  y_1 & y_2 & y_3 & y_4 & y_5 \\
  x_1 y_1 & x_2 y_2 & x_3 y_3 & x_4 y_4 & x_5 y_5
\end{bmatrix} \right) \cdot \begin{bmatrix}
  1 & x_1 & y_1 & x_1 y_1 \\
  1 & x_2 & y_2 & x_2 y_2 \\
  1 & x_3 & y_3 & x_3 y_3 \\
  1 & x_4 & y_4 & x_4 y_4 \\
  1 & x_5 & y_5 & x_5 y_5
\end{bmatrix}^{-1}
\]

18.2 Common Geometrical Operations

\[ x' = x, y' = y \implies g [x, y] = f[x', y'] = f [x, y] \implies \text{identity transformation, no change} \]
\[ x' = x + x_0, \ y' = y + y_0 \implies g [x, y] = f [x + a_{00}, y + b_{00}] \implies \text{translation by } [a_{00}, b_{00}] \]
\[ x' = a x, \ y' = b y \implies g [x, y] = f [a_{10} x, b_{01} y] \implies \text{spatial stretching or scaling} \]
\[ x' = x + a y, \ y' = y \implies g [x, y] = f [x + a y, y] \implies \text{skew} \]
\[ x' = x + a x y, \ y' = y \implies g [x, y] = f [x + a_{11} x y, y] \implies \text{perspective distortion} \]

\[ \begin{cases}
  x' = \alpha [x, y] = x \cos \theta - y \sin \theta \\
  y' = \beta [x, y] = x \sin \theta + y \cos \theta
\end{cases} \implies \text{rotation thru } \theta \]

\([x', y']\) are the coordinates of the input image that are mapped to \([x, y]\) in the output image.
18.3 Pixel Transfers

As already mentioned, a geometrical transformation may be implemented by specifying the where each pixel of the input image is located in the output grid, or by specifying the location of each output pixel on the input grid. Except for certain (and usually uninteresting) transformations, pixels of the input image will rarely map exactly to pixels of the output grid. It is therefore necessary to interpolate the gray value from pixels with noninteger coordinates (i.e., nongrid points) to pixels with integer coordinates that lie upon the grid. The method that transfers the input pixel coordinates and interpolates the gray value on the output grid is called pixel carryover by Castleman. It may also be called an input-to-output mapping. The algorithm that locates the output pixel upon the input grid and interpolates the gray value of the input pixels is called pixel filling or an output-to-input mapping.
18.4 Pixel Interpolation

Gray-value interpolation is based on the relative distances of the nearest neighbors to or from the geometrically transformed pixel. In the simplest case (nearest-neighbor or zeroth-order interpolation), all of the gray value is transferred to or from the nearest pixel. However, since the distances of the four neighbors must be evaluated to determine which is closest, it generally is quite easy to “upgrade” nearest-neighbor calculations to bilinear interpolation. This method divides the gray level of the non-integer pixel among its four nearest “integer” neighbors in inverse proportion to the distance; if the transferred pixel is equidistant from the four neighbors, then its gray value is divided equally, if it is much closer to one of the four grid points, most of its gray value is transferred to that pixel. For pixel filling, the gray value at the output pixel \( g[x', y'] \) is determined from the following equation, where \( x_n, y_n \) are the coordinates of the nearest pixel of the input grid to the transformed pixel, and \( d_n \) are the respective distances:

\[
g[x', y'] = \frac{\frac{1}{d_1} f_1[x_1, y_1] + \frac{1}{d_2} f_2[x_2, y_2] + \frac{1}{d_3} f_3[x_3, y_3] + \frac{1}{d_4} f_4[x_4, y_4]}{\frac{1}{d_1} + \frac{1}{d_2} + \frac{1}{d_3} + \frac{1}{d_4}}
\]

*Schematic of interpolation of pixel gray value. The input and output grids of pixels are shown as solid and dashed lines, respectively. The distances of the output pixel from the four nearest neighbors are labeled \( d_1 - d_4 \).*

The gray level of the transformed pixel can be divided among more of the neighboring pixels by using higher-order interpolation, e.g., cubic spline, etc.
18.4.1 Example: Image Rotation

The examples are small (64 × 64) images rotated by the degree increments specified. The images have been interpolated using the output-to-input mapping and bilinear interpolation. Note that artifacts are visible, particularly at 45°.
Chapter 19

Global Operations

If a pixel in the output image $g$ is a function of (almost) all of the pixels in $f[x, y]$, then $\mathcal{O}\{f[x, y]\}$ is a global operator. This category includes image coordinate transformations, of which the most important is the Fourier transform. These transformations derive new, usually equivalent, representations of images; for example, the Fourier transform maps from the familiar coordinate-space representation $f[x, y]$ to a new representation (a new image) whose brightness at each coordinate describes the quantity of a particular sinusoidal spatial frequency component present in $f[x, y]$. The sum of the component sinusoids is the original image. In other words, the Fourier transform generates the frequency-space representation $F[\xi, \eta]$ of the image $f[x, y]$. The coordinates of the image $[x, y]$ have dimensions of length (e.g., mm) while the coordinates of the frequency representation $[\xi, \eta]$ have units of inverse length (e.g., $\text{cycles/mm}$). Global gray-level properties of the image map to local properties in the Fourier transform, and vice versa. The frequency-space representation is useful for many applications, including segmentation, coding, noise removal, and feature classification. It also provides an avenue for performing other image operations, particularly convolution. Each output pixel is a function of the gray levels of all input pixels.

19.1 Relationship to Neighborhood Operations

The concept of a linear global operator is a simple extension of that of the linear local neighborhood operator. In that case, an output pixel was calculated by point-by-point multiplication of pixels in the input image by a set of weights (the kernel) and summing the products. The convolution at different pixels is computed by shifting the kernel. Recall that some accommodation must be made for cases where one or more elements of the kernel are off the edge of the image.

In the case of a global operator, the set of weights is as large as the image and constitutes a “mask function”, say $q[x, y]$. The output value obtained by applying a mask $q[x, y]$ to an input image $f[x, y]$ is:

$$g = \iint f[x, y] \ q[x, y] \ dx \ dy$$
In the discrete case, the integral becomes a summation:

$$g = \sum_n \sum_m f[n, m] \ q[n, m]$$

Note that a translation of the mask by one pixel in any direction shifts some of its elements over the edge of the image. If we assume that the output in such cases is undefined, only a single output pixel is calculated from one mask function $q[n, m]$. In general, different outputs result from different masks, i.e., we can define an output pixel by using different masks for each coordinate pair $[x', y']$:

$$g[k, \ell] = \sum_n \sum_m f[n, m] \ q[n, m; k, \ell]$$

Schematic of a global operation evaluated at two different output pixels $[x'_1, y'_1]$ and $[x'_2, y'_2]$. In each case, the input image $f[x, y]$ is multiplied by a “mask function” for the specific output pixel; the product values are summed to compute the output gray value $g$.

In general, the coordinates of $g$ are different from those of $f$, and often even have different dimensions (units). The action of the operator is obviously determined by the form of the mask function. The most common example is the Fourier transform, where the mask function is:

$$q[x, y; \xi, \eta] = \cos[2\pi (\xi x + \eta y)] - i \sin[2\pi (\xi x + \eta y)] = \exp[-2\pi i (\xi x + \eta y)]$$
19.2 Discrete Fourier Transform (DFT)

If the input signal has been sampled at discrete intervals (of width $\Delta x$, for example), the Fourier integral over $x$ reduces to a sum:

$$F[\xi] = \sum_{n=-\infty}^{+\infty} f[n \cdot \Delta x] \exp[-2\pi i \xi (n \cdot \Delta x)]$$

Recall that the Whittaker-Shannon sampling theorem states that a sinusoidal function must be sampled at a rate greater than two samples per period (Nyquist frequency) to avoid aliasing. Thus, the minimum period $X_{\text{min}}$ of a sampled sinusoidal function is two sample intervals ($2 \cdot \Delta x$ in the example above), which implies that the maximum spatial frequency in the sampled signal is:

$$\xi_{\text{max}} = \xi_{N_{\text{eq}}} = \frac{1}{X_{\text{min}}} = \frac{1}{2 \cdot \Delta x}$$

$\xi_{\text{max}}$ is measured in cycles per unit length (typically cycles per millimeter). Often the absolute scale of the digital image is not important, and the frequency is scaled to $\Delta x = 1$ pixel, i.e., the maximum spatial frequency is $\frac{1}{2}$ cycles per pixel. The range of meaningful spatial frequencies of the DFT is $\frac{1}{2 \Delta x} > |\xi|$.

If the input function $f[x]$ is limited to $N$ samples, the DFT becomes a finite sum:

$$F[\xi] \equiv \sum_{n=0}^{N-1} f[n \cdot \Delta x] \exp[-2\pi i \xi (n \cdot \Delta x)]$$

or

$$\sum_{n=-N/2}^{N/2-1} f[n \cdot \Delta x] \exp[-2\pi i \xi (n \cdot \Delta x)]$$

The DFT of a 1-D sequence of $N$ samples at regular intervals $\Delta x$ can be computed at any spatial frequency $\xi$. However, it is usual to calculate the DFT of a sequence of frequencies (e.g., a total $M$) separated by a constant interval $\Delta \xi$. Each sample of the DFT of a real sequence of $N$ pixels requires that $N$ values each of the cosine and sine be computed, followed by $2N$ multiplications and $2N$ sums, i.e., of the order of $N$ operations. The DFT at $M$ spatial frequencies requires of the order of $M \cdot N$ operations. Often, the DFT is computed at $N$ frequencies, thus requiring of the order of $N^2$ operations. This intensity of computation made calculation of the DFT a tedious and rarely performed task before digital computers. For example, a Fourier deconvolution of seismic traces for petroleum exploration was performed by Enders Robinson in 1951; it took the whole summer to do 32 traces by hand with a memoryless mechanical calculator. This task could now be done with the cheapest PC in less than a second. Even with mainframe digital computers into the 1960s, digital Fourier analysis was unusual because of the computation time. In 1965, J.W. Cooley and J.W. Tukey developed the Fast Fourier Transform algorithm, which substantially cut computation times and made digital Fourier analysis feasible.
19.3 1-D Fast Fourier Transform (FFT)

The FFT was developed to compute discrete Fourier spectra with fewer operations than the DFT by sacrificing some flexibility. The DFT may compute the amplitude of sinusoidal components at any frequency within the Nyquist window, i.e., the DFT maps discrete coordinates \( n \cdot \Delta x \) to a continuous set of frequencies \( \xi \) in the interval \([-\xi_{Nyquist}, \xi_{Nyquist}]\). The DFT may be computed at a single spatial frequency if desired. In the FFT, the amplitudes at \( N \) discrete equally spaced frequencies are computed in the interval \([-\xi_{Nyquist}, \xi_{Nyquist}]\) from \( N \) input samples. The frequency samples are indexed by the integer \( k \) and the interval between frequency samples is:

\[
\Delta \xi = \frac{2}{N} \cdot \xi_{Nyquist} = \frac{2}{N} \cdot \frac{\Delta x}{\Delta x} = \frac{1}{N \cdot \Delta x}
\]

\[
\Rightarrow \xi_k = k \cdot \Delta \xi = \frac{k}{N \cdot \Delta x}, \quad \left[ -\frac{N}{2} \leq k \leq \frac{N}{2} - 1 \right]
\]

\[
\Rightarrow N \cdot \Delta x \cdot \Delta \xi = 1
\]

If we substitute these specific frequencies into the DFT:

\[
F[k \cdot \Delta \xi] = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f[n \cdot \Delta x] \exp \left[ -2\pi i k \cdot \Delta \xi \cdot (n \cdot \Delta x) \right]
\]

\[
= \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f[n \cdot \Delta x] \exp \left[ -2\pi i k n \cdot \frac{\Delta x}{N \cdot \Delta x} \right]
\]

but \( \Delta \xi = \frac{1}{N \cdot \Delta x} \Rightarrow F[k \cdot \Delta \xi] = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f[n \cdot \Delta x] \exp \left[ -\frac{2\pi i k n}{N} \right]
\]

If \( \Delta x \) is assumed to be a dimensionless sample, then the Nyquist frequency is fixed at:

\[
\xi_{Nyquist} = \frac{1}{2 \text{ cycle sample}} = \frac{\pi}{\text{ radians sample}}
\]

Recall that the DFT assumes that the sample interval is \( \Delta x \) and computes a periodic spectrum with period \( \frac{1}{\Delta x} \). In the FFT, the spectrum is assumed to be sampled at intervals \( \Delta \xi = \frac{1}{N \cdot \Delta x} \), which implies in turn that the input function is periodic with period \( N \cdot \Delta x \). If \( N \) is a power of 2 (e.g., 128, 256, 512, \cdots), there are only \( N \) distinct values of the complex exponential \( \exp \left[ -\frac{2\pi i k n}{N} \right] \) to calculate. By using this fact, the number of required operations may be reduced and processing speeded up. The FFT of \( N \) samples requires of the order \( N \cdot \log_2 N \) operations vs. \( O\{N^2\} \) for the DFT.

Since both representations \( f[n \cdot \Delta x] \) and \( F[k \cdot \Delta \xi] \) are sampled and periodic, the
inverse FFT is a finite summation and is proportional to:

\[ f [n \cdot \Delta x] = C \sum_{k=-N/2}^{N/2-1} F [k \cdot \Delta \xi] \exp \left[ -\frac{2\pi ink}{N} \right] \exp \left[ +\frac{2\pi ink}{N} \right] \]

The proportionality constant \( c \) is required to ensure that \( \mathcal{F}_1 \{F[k]\} = f[n] \), and may be found by substituting the formula for the forward FFT for \( F[k \cdot \Delta \xi] \):

\[
\begin{align*}
    f [n \cdot \Delta x] &= C |\Delta x| \sum_{k=-N/2}^{N/2-1} \left( \sum_{m=-N/2}^{N/2-1} f [m \cdot \Delta x] \exp \left[ -\frac{2\pi imk}{N} \right] \exp \left[ +\frac{2\pi imk}{N} \right] \right) \\
    &= C \sum_{k=-N/2}^{N/2-1} \sum_{m=-N/2}^{N/2-1} f [m \cdot \Delta x] \exp \left[ -\frac{2\pi ik}{N} (n - m) \right] \\
    &= C \sum_{m=-N/2}^{N/2-1} f [m \cdot \Delta x] \sum_{k=-N/2}^{N/2-1} \exp \left[ -\frac{2\pi ik}{N} (n - m) \right] \\
    &= C \sum_{m=-N/2}^{N/2-1} f [m \cdot \Delta x] \cdot (N \cdot \delta [n - m]) \\
    &= C \cdot N \cdot f [n \cdot \Delta x] \\
    &= f [n \cdot \Delta x]
\end{align*}
\]

Thus \( C = N^{-1} \) and the inverse FFT may be defined as:

\[
f [n \cdot \Delta x] = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} F [k \cdot \Delta \xi] \exp \left[ +\frac{2\pi ink}{N} \right]
\]

The proportionality constant is a scale factor that is only significant when cascading forward and inverse transforms and may be applied in either direction. Many con-
CONVENTIONS (INCLUDING MINE) INCLUDE THE PROPORIONALITY CONSTANT IN THE FORWARD FFT:

$$F[k \cdot \Delta \xi] = F \left[ \frac{k}{N \cdot \Delta x} \right] = \frac{1}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f[n \cdot \Delta x] \exp \left[ -\frac{2\pi i nk}{N} \right]$$

$$f[n \cdot \Delta x] = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} F[k \cdot \Delta \xi] \exp \left[ +\frac{2\pi i nk}{N} \right]$$

$$N \cdot \Delta x \cdot \Delta \xi = 1$$

$$\xi_{Nyq} = N \cdot \frac{\Delta \xi}{2}$$

$$x_{\text{max}} = N \cdot \frac{\Delta x}{2}$$

19.4 Multidimensional DFTs

The concept of a 1-D Fourier transform can be easily extended to multidimensional continuous or discrete signals. The continuous 2-D transform is defined as:

$$\mathcal{F}_2 \{ f[x, y] \} \equiv F[\xi, \eta] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x, y] \exp[-2\pi i (\xi x + \eta y)] \, dx \, dy$$

For a uniformly sampled discrete 2-D function $f[n, m]$, the transform is a summation:

$$\mathcal{F}_2 \{ f[n \cdot \Delta x, m \cdot \Delta y] \} = \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} f[n, m] \exp[-2\pi i (\xi \cdot n \cdot \Delta x + \eta \cdot m \cdot \Delta y)]$$

$$= F[\xi, \eta]$$

The Fourier transform of a real-valued 2-D function is Hermitian (even real part and odd imaginary part. The Fourier transform of the image of a 2-D cosine is a pair of delta-function spikes at a distance from the origin proportional to the frequency of the cosine, as shown on the next page. The polar angle of the spikes relative to the origin indicates the direction of variation of the cosine, while the brightness of the spikes is proportional to the amplitude of the cosine. Notice that the Fourier transform of the sum of two cosine waves is the sum of the individual transforms (i.e., the Fourier transform is linear).

If the sampled input image $f[n, m]$ has $N \times N$ pixels, then there are only $N^2$ pieces of information in the input. Thus the DFT $F[\xi, \eta]$ also must contain at most only $N^2$ independent pieces of information. This allows the DFT to be sampled without loss of information:

$$F[\xi, \eta] \rightarrow F[k \cdot \Delta \xi, \ell \cdot \Delta \eta] \rightarrow F[k, \ell]$$
It is easy to show (see linear math course) that
\[
\Delta \xi = \frac{1}{N \cdot \Delta x} \\
\Delta \eta = \frac{1}{N \cdot \Delta y}
\]

The 2-D transform has the same properties mentioned before, including that global properties become local properties and vice versa. This is the primary reason why the Fourier transform is such a powerful tool for image processing and pattern recognition; \( F[\xi, \eta] \) is uniquely defined for each \( f[x, y] \), and the global properties of \( f[x, y] \) are concentrated as local properties of \( F[\xi, \eta] \). Therefore:

1. local modification of \( f[n, m] \) \( \rightarrow \) global modification of \( F[k, \ell] \)

2. global modification of \( f[n, m] \) \( \rightarrow \) local modification of \( F[k, \ell] \)

Local modification in the space domain is what we call “filtering” and is intimately related the local operation of convolution that we’ve already discussed. In fact, it is easy to prove that the Fourier transform of a convolution is the product of the Fourier transforms of the component functions. This result is called the filter theorem.

\[
\mathcal{F}_2 \{ f[n, m] * h[n, m] \} = \mathcal{F}_2 \{ f[n, m] \} \cdot \mathcal{F}_2 \{ h[n, m] \} \\
\implies f * h = \mathcal{F}_2^{-1} \{ F[k, \ell] \cdot H[k, \ell] \}
\]

We have already given the name impulse response or point spread function to \( h[x, y] \); the representation \( H[\xi, \eta] \) is called the transfer function of the system. The most common reason for computing Fourier transforms of digital signals or images is the to use this path for convolution. Examples are shown for both lowpass filtering and differentiation (highpass filtering).
Averaging of a bitonal “E” in the vertical direction with
\( h[n, m] = \delta_d[n] \cdot RECT[\frac{m}{\eta}] \); the transfer function is “skinny” in the \( \eta \)-direction, thus attenuating sinusoidal components that oscillate in the vertical direction.
Differentiation in the $x$-direction of a pair of “E”s applied in the Fourier domain. The impulse response is $h[n, m] = \delta_d[n + 1, m] - \delta_d[n, m]$; the MTF is large for $|k| >> 0$, so it enhances high-frequency sinusoids that oscillate in the horizontal direction. The output shows just the vertical edges of the “E”s.
19.5 Other Global Operations

Other families of mask functions may be used in the general coordinate transformation equation:

\[ g[u] = \int_{-\infty}^{+\infty} f[x] \ q[x; u] \ dx \quad (1-D \ \text{continuous functions}) \]

\[ g[k] = \sum_n f[n] \ q[n; k] \quad (1-D \ \text{discrete functions}) \]

\[ g[u, v] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f[x, y] \ q[x, y; u, v] \ dx \ dy \quad (2-D \ \text{continuous functions}) \]

\[ g[k, \ell] = \sum_n \sum_m f[n, m] \ q[n, m; k, \ell] \quad (2-D \ \text{discrete functions}) \]

Such a transformation is invertible if \( f[x, y] \) can be derived from \( g[u, v] \) via an expression of the form:

\[ f[x] = \int_{-\infty}^{+\infty} g[u] \ q'[x; u] \ du \quad (1-D \ \text{continuous functions}) \]

\[ f[n] = \sum_n g[k] \ q'[n; k] \quad (1-D \ \text{discrete functions}) \]

\[ f[x, y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g[u, v] \ q'[x, y; u, v] \ du \ dv \quad (2-D \ \text{continuous functions}) \]

\[ f[n, m] = \sum_k \sum_\ell g[k, \ell] \ q'[n, m; k, \ell] \quad (2-D \ \text{discrete functions}) , \]

where \( m' \) is the 4-D mask function for the inverse transform. For the purposes of this course it is not really essential to understand the conditions for the transformation to be invertible, so we will just say that the set of mask functions \( q[x, y; u, v] \) must be complete, i.e., any function \( f[x, y] \) must be representable as a sum (linear combination) of the mask functions.

19.6 Discrete Cosine Transform (DCT)

The discrete cosine transform (DCT) is a relative of the discrete Fourier transform that is often used in digital image compression (also called image coding). The term is used to denote any of several related algorithms, but we will consider only the even symmetric DCT. Consider a 2-D sampled image \( f[n, m] \) with \( N \times N \) pixels in the interval \( 0 \leq n, m \leq N - 1 \). We have already seen that the the DFT “replicates” the finite 2-D array to create an infinite array that is periodic over \( N \times N \) samples, as shown in the upper half of the figure. An equivalent operation that exhibits some significant advantages creates the infinite array of samples by appropriately reflecting or “flipping” the 2-D block of pixels so that border pixels touch only themselves in
the infinite array: For example, the DCT of the same 2-D array constructs a new function \( g[n,m] \) that is periodic over \( 2N \) samples in the interval \( 0 \leq n, m \leq 2N - 1 \). Consider the 1-D case first. The DCT is the \( N \)-pixel array \( f[n] \) \( (0 \leq n \leq N - 1) \) first creates the \( 2N \)-point DFT of \( g[n] \) via:

\[
g[n] = f[n] \quad \text{for} \quad 0 \leq n \leq N - 1 \\
g[n] = f[2N - 1 - n] \quad \text{for} \quad N \leq n \leq 2N - 1
\]

Thus the original pixel \( f[N - 1] \) touches “itself” in the replicated array, since \( g[N] = g[N - 1] = f[N - 1] \). This “new” \( 2N \)-pixel function \( g[n] \) is then replicated to create the infinite array. Note that the other “border” pixel \( f[0] \) is replicated at the other edge of the array and thus also butts up against itself. In this manner, the discontinuities at the edges of the image and “leakage” from the false transitions at the edge of the array are eliminated from the transform representation. This has the effect of concentrating the energy in the transform at fewer pixels so that more pixels in the transform will have small amplitudes. In this way, the histogram of the DCT will be more compact than that of the DFT. In addition, the DCT of a real-valued array can be made real valued.

The 2-D DCT is a simple extension of the 1-D version. Given an \( N \times N \) image \( f[n,m] \), a \( 2N \times 2N \) version \( g[n,m] \) is created by replication and reflection. The new image will be smoothly periodic over \( 2N \times 2N \) so that leakage will not be present.

The 1-D example will use an 8-pixel ramp function defined by:

\[
f[n] = n \quad \text{for} \quad 0 \leq n \leq 7
\]

The function \( f[n] \), the infinite array assumed by the DFT, and the real and imaginary parts of the 8-pixel DFT are shown in the figure:
CHAPTER 19 GLOBAL OPERATIONS

DFT Arrangement of Replicas of Pixel Blocks

Block of Pixels

DCT Arrangement of Replicas of Pixel Blocks
Illustration of “leakage” from discrete Fourier transform: (a) 8-pixel $f[n] \cong n$ for $0 \leq n \leq 7$; (b) replicated array assumed by the DFT, showing the transitions in gray level at the edges of the 8-pixel array; (c) real part of DFT $\Re \{F[k]\}$; (d) imaginary part of DFT $\Im \{F[k]\}$, showing significant amplitude at large spatial frequencies.

The $2M$-pixel synthetic array $g[n]$ is indexed over $n$ ($0 \leq n \leq 2M - 1$) and has the form:

$$g[n] = \begin{cases} 
  f[n] & \text{if } 0 \leq n \leq M - 1 \\
  f[2M - 1 - n] & \text{if } M \leq n \leq 2M - 1 
\end{cases}$$

In the case $M = 8$, the array $g[n]$ is defined:

$$g[n] = \begin{cases} 
  f[n] & \text{if } 0 \leq n \leq 7 \\
  f[15 - n] & \text{if } 8 \leq n \leq 15 
\end{cases}$$
The values of $g[n]$ for $8 \leq n \leq 15$ is a “reflected replica” of $f[n] :$

- $g[8] = f[7]$
- $g[9] = f[6]$
- $g[10] = f[5]$
- $g[12] = f[3]$
- $g[14] = f[1]$
- $g[15] = f[0]$

If the “new” array $g[n]$ is assumed to be periodic over $2M$ samples, its amplitude is defined for all $n$, e.g.,

$$g[n] = \begin{cases} 
    f[-1 - n] & \text{if} \quad -M \leq n \leq -1 \implies -16 \leq n \leq -1 \\
    f[n + 2M] & \text{if} \quad -2M \leq n \leq -M - 1 \implies -32 \leq n \leq -17 
\end{cases}$$

Note that the 16-sample block $g[n]$ is NOT symmetric about the origin of coordinates because $g[-1] = g[0]$; to be symmetric, $g[-c]$ would have to equal $g[c]$. For example, consider a 1-D example where $f[n]$ is an 8-pixel ramp as shown:
DFT of 16-pixel Function $g[n]$ constructed from 8-pixel “ramp”: (a) $g[n]$ for $2N = 16$; (b) replicas of $g[n]$, showing that edge pixels of $f[n]$ butt against themselves: (c) $\Re\{G[k]\}$; (d) $\Im\{G[k]\}$, showing the redundancy of the complex spectrum and the high-frequency terms due to the discontinuous transition at the edge.

The $2M$-pixel representation of $f[n]$ is the $g[n]$ just defined:

$$g[n] = \begin{cases} 
  f[n] & \text{for } 0 \leq n \leq 7 \\
  f[2M - 1 - n] & \text{for } 8 \leq n \leq 15 
\end{cases}$$

If this function were symmetric (even), then circular translation of the 16-point array by 8 pixels to generate $g[n - 8 \mod 16]$ also be an even function.

From the graph, it is apparent that the translated array is not symmetric about the origin; rather, it has been translated by $-\frac{1}{2}$ pixel from symmetry in the $2M$-pixel
array. Thus define a new 1-D array $c[n]$ that is shifted to the left by $\frac{1}{2}$ pixel:

$$c[n] = g\left[n - \frac{1}{2}\right]$$

This result may seem confusing at first; how can a sampled array be translated by $\frac{1}{2}$ pixel? For the answer, consider the continuous Fourier transform of a sampled array translated by $\frac{1}{2}$ unit:

$$\mathcal{F}_1\{c[n]\} \equiv C[\xi] = \mathcal{F}_1\left\{g\left[x - \frac{1}{2}\right]\right\} = \mathcal{F}_1\left\{g[x] * \delta\left[x - \frac{1}{2}\right]\right\}$$

$$= G[\xi] \cdot \exp\left[-2\pi i \xi \cdot \frac{1}{2}\right] C[\xi] = G[\xi] \cdot \exp[-i\pi \xi]$$

Thus the effect of translation by $\frac{1}{2}$ pixel in the space domain is multiplication of the Fourier transform by the specific linear-phase factor:

$$\exp[-i\pi \xi] = \cos[\pi \xi] - i \sin[\pi \xi].$$

The discrete form of this function is $C[k]$:

$$C[k] \equiv \exp\left[-2\pi i \frac{k}{N}\right] \text{ for } -N \leq k \leq N - 1$$
Phase correction \( C[k] \) applied to remove the imaginary part of the 2N-pixel DFT:

(a) \( \Re \{ C[k] \} = \cos \left( \frac{2\pi k}{N} \right) \) for \(-N \leq k \leq N - 1\); (b) \( \Im \{ C[k] \} = -\sin \left( \frac{2\pi k}{N} \right) \); (c) \( |C[k]| = 1[k]\); (d) \( \Phi \{ C[k] \} = -2\pi \frac{k}{N} \)

The 2M-point DFT of the symmetric discrete array (original array translated by \( \frac{1}{2} \) pixel) has the form:

\[
F_{2M} \{ c[n] \} = F_{2M} \left\{ g \left[ n - \frac{1}{2} \right] \right\} = F_{2M} \left\{ g[n] \ast \delta \left[ n - \frac{1}{2} \right] \right\} = G[k] \cdot \exp \left[ -i\pi \left( \frac{k}{2M} \right) \right] C[k] = G[k] \cdot \exp \left[ -i\pi \frac{k}{2M} \right]
\]

where the continuous spatial frequency \( \xi \) has been replaced by the sampled frequency \( \frac{k}{2M} \). This function \( C[k] \) is the DCT of \( f[n] \). Because the 2M-point translated function \( c[n] \) is real and even, so must be the 2M-point discrete spectrum \( C[k] \); therefore only \( M \) samples of the spectrum are independent. This array is the DCT of \( f[n] \).
$C[k] \cdot G[k]$ for $2N = 16$: (a) $\mathbb{R}\{C[k] \cdot G[k]\}$; (b) $\Im\{C[k] \cdot G[k]\} = 0[k]$.

The real part of the array is truncated to $N = 8$ to produce the DCT of $f[n]$:

**DCT of $f[n]$ is the real part of $G[k] \cdot C[k]$ truncated to 8 pixels.**

### 19.6.1 Steps in Forward DCT

To summarize, the steps in the computation of the 1-D DCT of an $M$-pixel block $f[n]$ are:

1. create a $2M$-pixel array $g[n]$ from the $M$-pixel array $f[n]$:
   
   $g[n] = f[n] : 0 \leq n \leq M - 1$
   
   $g[n] = f[2M - 1 - n] : M \leq n \leq 2M - 1$

2. compute the $2M$-point DFT of $g[n] = G[k]$

3. the $M$-pixel DCT $C[k] = \exp\left[-\frac{j\pi k}{2M}\right] \cdot G[k]$ for $0 \leq k \leq M - 1$
The entire process may be cast into the form of a single equation, though the algebra required to get there is a bit tedious,

\[ C[k] = \sum_{n=0}^{M-1} 2 f[n] \cos \left( \pi k \cdot \frac{2n + 1}{2M} \right) \quad \text{for} \quad 0 \leq k \leq M - 1 \]

19.6.2 Steps in Inverse DCT

The inverse DCT is generated by applying the procedures in the opposite order:

1. create a 2M-pixel array \( G[k] \) from the M-pixel DCT \( C[k] \):
   \[
   G[k] = \exp \left[ +\frac{i\pi k}{2M} \right] \cdot C[k] \quad \text{for} \quad 0 \leq k \leq M - 1
   \]
   \[
   G[k] = -\exp \left[ +\frac{i\pi k}{2M} \right] \cdot C[2M - k] \quad \text{for} \quad M + 1 \leq k \leq 2M - 1
   \]

2. compute the inverse 2M-pixel DFT of \( G[k] \rightarrow g[n] \)

3. \( f[n] = g[n] \) for \( 0 \leq n \leq M - 1 \)

The single expression for the inverse DCT is:

\[
f[n] = \frac{1}{M} \sum_{k=0}^{M-1} w[k] \cdot C[k] \cos \left( \pi k \cdot \frac{2n + 1}{2M} \right) \quad \text{for} \quad 0 \leq n \leq M - 1
\]

where \( w[k] = \frac{1}{2} \) for \( k = 0 \) and \( w[k] = 1 \) for \( 1 \leq k \leq M - 1 \).

19.7 Walsh-Hadamard Transform

A transform which has proven useful in image compression and pattern recognition grew out of matrix theory and was first described by Hadamard in 1893. A modification made by Walsh in 1923 is commonly used, and so the transformation is often named after both men. The W-H transform resembles a Fourier transform with a mask that has been thresholded so that there are only two values, \( \pm 1 \). We will first consider the W-H transform of a one-dimensional function \( f[n] \). The 1-D W-H transform of a two-pixel image is derived from two mask functions which may be represented as vectors:
\[ g[k] = \sum_{n=0}^{1} f[n] q[n; k] \quad \text{1-D discrete functions} \]

\[ q[0, 0] = +1, \quad q[0, 1] = +1 \quad \Rightarrow \quad q_0 = \begin{bmatrix} +1 \\ +1 \end{bmatrix} \]

\[ q[1, 0] = +1, \quad q[1, 1] = -1 \quad \Rightarrow \quad q_1 = \begin{bmatrix} +1 \\ -1 \end{bmatrix} \]

Thus the elements of the 2-pixel W-H transform are the sum and difference of the pixel gray values, which is identical to the 2-pixel Fourier transform. The vectors \( m_0 \) and \( m_1 \) are orthogonal (i.e., perpendicular) so that:

\[ q_i \cdot q_j \equiv \left( q_i \right)_0 \cdot \left( q_j \right)_0 + \left( q_i \right)_1 \cdot \left( q_j \right)_1 = \begin{cases} 0 & \text{if } i \neq j \\ 2 & \text{if } i = j \end{cases} \]

The column vectors that define the mask can be assembled into a 2 \( \times \) 2 orthogonal matrix:

\[ H_2 \equiv \begin{bmatrix} q_0 & q_1 \end{bmatrix} = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix} \]

\[ H_2^T = H_2 \]

\[ H_2 \cdot H_2^T = (H_2)^2 = \begin{bmatrix} +2 & 0 \\ 0 & +2 \end{bmatrix} = 2I \]

where \( I \) is the 2 \( \times \) 2 identity matrix. Note that \( H_2 \) is identical to its transpose \( H_2^T \), which is the criterion that defines an orthogonal matrix.

The inverse matrix \( H_2^{-1} \) is defined as the matrix that satisfies:

\[ H_2 H_2^{-1} = I \]

which shows that the inverse W-H transform is proportional to the transpose of \( H_2 \).
and thus to $H_2$ itself.

$$H_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}$$

$$H_2^{-1} = \frac{1}{2} \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}$$

The four-pixel W-H matrix is obtained by the direct product of the two-pixel W-H matrix with itself, i.e.,

$$H_4 = H_2 \times H_2 = \begin{bmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{bmatrix} = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix}$$

Thus when applied to a 4-pixel image with values $f(n)$ for $0 \leq n \leq 3$, the W-H transform is a 4-pixel image with gray values:


Note that the elements of $w[k]$ are sums and differences of the elements of $f[n]$, and thus no multiplication is required. Also the elements of the W-H transform representation will be integers if so are the input gray values. This is the source of one useful characteristic of the W-H transform: that the transform representation need not be requantized for display. Also note that the elements of the mask functions of the W-H transform are all purely real numbers so that the W-H transform of a real image is real.

The inverse 4-pixel W-H transform is easily confirmed to be:

$$f[0] = \frac{1}{4}(w[0] + w[1] + w[2] + w[3])$$

$$f[1] = \frac{1}{4}(w[0] - w[1] + w[2] - w[3])$$


Recall that the elements of the Fourier transform are ordered in terms of the spatial
frequency of the mask functions. In similar fashion, the W-H transform elements can be ordered in terms of the number of sign changes of the mask function, which is called the sequency of the term. The \textit{sequency-ordered} Walsh-Hadamard transformation matrix is:

\[
H'_4 = \begin{bmatrix}
+1 & +1 & +1 & +1 \\
+1 & +1 & -1 & -1 \\
+1 & -1 & -1 & +1 \\
+1 & -1 & +1 & -1
\end{bmatrix} \quad (\text{no sign changes})
\]

\[
\begin{bmatrix}
+1 & +1 \\
+1 & -1 \\
+1 & -1 \\
+1 & -1
\end{bmatrix} \quad (\text{one sign change})
\]

\[
\begin{bmatrix}
+1 & -1 \\
+1 & +1 \\
+1 & -1 \\
+1 & +1
\end{bmatrix} \quad (\text{two sign changes})
\]

\[
\begin{bmatrix}
+1 & -1 \\
+1 & -1 \\
+1 & -1 \\
+1 & -1
\end{bmatrix} \quad (\text{three sign changes})
\]

Note that the ordered W-H matrix is also orthogonal, so that the ordered inverse W-H transform is proportional to the ordered forward transform. In similar fashion, the unordered matrix $H_8$ can be obtained by computing the direct product of $H_4$ and $H_2$:

\[
H_8 = H_4 \times H_2 = \begin{bmatrix}
+H_4 & +H_4 \\
+H_4 & -H_4
\end{bmatrix}
\]

\[
\begin{bmatrix}
+H_2 & +H_2 & +H_2 & +H_2 \\
+H_2 & +H_2 & -H_2 & -H_2 \\
+H_2 & -H_2 & -H_2 & +H_2 \\
+H_2 & -H_2 & +H_2 & -H_2
\end{bmatrix}
\]

\[
= \begin{bmatrix}
+1 & +1 & +1 & +1 \\
+1 & -1 & +1 & -1 \\
+1 & +1 & -1 & -1 \\
+1 & -1 & -1 & +1
\end{bmatrix}
\]

The sequency-ordered matrix is:
\[ H'_s = (H'_s)^T = \begin{bmatrix}
+1 & +1 & +1 & +1 & +1 & +1 \\
+1 & +1 & +1 & -1 & -1 & -1 \\
+1 & +1 & -1 & -1 & -1 & +1 \\
+1 & +1 & -1 & +1 & -1 & -1 \\
+1 & -1 & -1 & +1 & +1 & +1 \\
+1 & -1 & -1 & +1 & +1 & +1 \\
+1 & -1 & +1 & -1 & +1 & +1 \\
+1 & -1 & +1 & -1 & +1 & +1
\end{bmatrix} \]

Basis functions of the Walsh-Hadamard transform arranged in order of increasing
“sequency,” which is analogous to the “frequency” of the sinusoidal basis functions of the Fourier transform. An 8-sample 1-D input function is decomposed into how much of that function can be written as each of these basis functions.

19.7.1 Interpretation of the W-H Transform

Note that the W-H matrices may be normalized so that the forward and inverse transforms are completely identical. This is done by dividing the elements of the matrix by $\sqrt{N}$ so that $H_N \cdot H_N = I$

The relative sizes of elements of the W-H representation will indicate the busyness of the input image; a smooth image will have larger values of the W-H transform for small values of $k$ while the W-H transform of a busy image will be larger for larger values of $k$. The gray value of an element of the transform may be interpreted as the similarity between the input image and the mask image. The elements of the W-H transform of realistic images whose gray levels are well-correlated (i.e., smooth) will tend to be large for small values of $k$. In other words, the energy of the transformed image will tend to be compressed into the pixels indexed by small $k$; pixels with large $k$ will have small values. This property is called energy compaction, and is useful in (and in fact, is the whole basis for) signal compression.
Examples of the 1-D Walsh-Hadamard transform evaluated over 64 pixels. As the input image gets “busier,” the maximum sequency of the W-H transform increases.

The set of mask functions for the 2-D W-H transform are products of the 1-D functions, i.e., the transform is separable. The $8 \times 8$ basis functions and the decomposition of two gray-scale images into their $8 \times 8$ block Walsh-Hadamard transforms are shown.
The 64 basis functions of the $8 \times 8$ Walsh-Hadamard transform: the 64 basis functions are shown on the left and are segmented by red dashed lines. The resulting $8 \times 8$ W-H transforms are on the right, and show that only one pixel is positive in each.

Two examples of $8 \times 8$ block Walsh-Hadamard transforms, showing that busier parts of the image produce more outputs with larger sequencies.
Chapter 20

Image Compression

20.1 References for Image Compression

Many WWW Sites on Compression – millions on Google.

- Key Papers in the Development of Information Theory, IEEE Press, 1974
- Raisbeck, Information Theory for Scientists and Engineers
- Ash, Information Theory, Dover
- Pierce, J.R., An Introduction to Information Theory, Signals, Systems, and Noise, Dover

20.2 Image Storage, Transmission, and Compression

All of you probably have an intuitive feeling for the fundamental difficulties of storing and transmitting information; the well-known phrase “one a picture is worth 1024 words” is an example. The subject has been quite important in imaging as far back as the early days of television development. It has become increasingly important as the digital juggernaut has taken hold. For comparison, consider the recent history of “imaging” of an audio music signal. The goal of the high fidelity industry for many years has been to record and play back ever more of the music signal by increasing the
bandwidth of the recorded signal to reproduce the sound as realistically as possible. In other words, the industry has attempted to increase the amount of recorded “information” to increase the “fidelity” of the reproduced signal. However, the recording industry has been revolutionized in the last few years (since before the turn of the millennium) by the influx of digital recording technology. The goal now is to discard as much information as possible (and thus “compress” the signal) without degrading the “message” conveyed by the sound. The goal is to increase the quantity of music recorded on a device with fixed data capacity, thus reducing the cost of storing and conveying the message. This is only possible because digitally recorded signals can be manipulated in ways that are not available for analog recordings.

The property of messages that allows them to be compressed is that some portion of the data is not essential to understanding of the message. For a simple example, consider the text message:

\[\text{Th qck brn fx jmpd ovr th lz dg}\]

which is perfectly readable (at least to most of us for whom English is their first language) but which requires only 75% of the letters in the complete text. The moral of this example is that vowels are (usually) unnecessary to the meaning of text messages. In general, vowels supply redundant information in the message and often may be discarded without loss of meaning. In an imaging sense, we might say that vowels supply low-frequency information (bulk properties of the words), while consonants carry the high-frequency information (information about the “edges”).

Redundancies in text messages or images may be identified and eliminated using different but analogous schemes. The data compression process often is called coding, because the representation is altered and may require a reference source (codebook) to interpret the message in terms of the original symbols. The name coding also implies a relationship between the processes of message compression and message cryptography, which are (in a sense) opposites. Compression is the process of removing nonessential information from the message to reduce the quantity of data; cryptography is the process of adding nonessential information to different messages to make them look indistinguishable.

The principles of image compression require an understanding of the fundamental concepts of information theory, which were laid down by Claude Shannon in the late 1940s, with later contributions by many authors, notably Norbert Wiener and A.I. Khinchin. The technologies (hardware and software) were very hot topics in the 1980s because of limitations in storage space (as difficult as it may be to believe now, a common size for a PC hard drive was 30 MBytes in the mid-late 1980s). The topic heated up again in the late 1990s for internet applications because of limitations in transmission bandwidth.

There are (at least) two classes of data redundancy: objective and subjective. Objectively redundant data can be removed without any loss of information. In other words, deleted objectively redundant data may be recovered without loss. Subjectively redundant data may be removed from a message or image without any “visible” loss of information, though the original image cannot be recovered without error. By
20.3 Image Compression

20.3.1 Histograms, Information Theory, and Imaging

A digital image is a matrix of pixels with each gray level described as a binary integer of \( m \) bits, for a total of \( M = 2^m \) levels. An image of 512 \( \times \) 512 pixels with 256 levels requires \( 512^2 \cdot 8 \) bits (256 KBytes) to store the image. The number of bits needed to store a full-color image is three times as large, as images in each of the three primary colors are required. Though the cost of digital memory capacity continues to decrease, and disk drives with capacities exceeding 300 GByte (300 \( \cdot \) 1024 \( ^3 \) \( \approx \) 3 \( \cdot \) 10\( ^{11} \) bytes) are becoming common, the sizes of digital images continue to increase as well. Six-megapixel cameras (3000 \( \times \) 2000) are now quite affordable, even for many amateurs. An output image from such a camera requires 2000 lines by 3000 pixels by 12 bits by 3 colors, or a total of 216 Mbits (\( \approx 25 \) Mbytes per image). To help scholars read old manuscripts, we regularly produce color images for display that are about 5000 \( \times \) 7500 color pixels, or 107 MBytes (approximately 6 such images may be stored on a standard CD-ROM). The requirements of this project have consistently confirmed Parkinson’s Law of computing: that data to be stored always exceeds available storage capacity. However, it is usually possible to satisfactorily store images of 512\( ^2 \) 8-bit (256 KBytes) with good visual quality while using much less data. In other words, real images “always” contain less information than the maximum possible; the difference is due to the redundancy of image data, i.e., the gray value of a pixel in a realistic image.
usually is not selected from a uniformly random distribution, but rather exhibits some type of correlation with the gray values of other (usually neighboring) pixels. An image or other message with redundant data may be compressed without loss of information by removing some (or even all) of the redundancy. Given the ratio of nonredundant data in a message or image to the total data, the redundancy $R$ in the message may be defined as:

$$R = 1 - \frac{\text{Nonredundant Data [bits]}}{\text{Total Data [bits]}}$$

$R = 0$ if all bits are required to transmit the message without loss.

Any system that is constrained by transmission and/or storage capacity will benefit from reducing (or even removing) data redundancy. For example, the total resolution (spatial, temporal, and color) of the human visual system ultimately is limited by the transmission capacity of the channel connecting the eye and brain, and removal of redundancies is a principal task of the vision system. This also leads to the concept of subjective redundancy, which means that some information may be discarded from an image without visible impact on the quality, e.g., oscillations in gray value with very short periods.

Obviously, to understand the principles of image compression by removing redundancy, the concept of information must be considered first. “Information” is the quantity of data in a message, i.e., how much data is required to describe the message, measured in bits (for storage) or in bits per second or in channel bandwidth (Hz) for transmission.

An image can be considered to be a message consisting of different gray values obtained from some set of well-defined possible values. The quantity of information in the message is a function of the probability of occurrence of the event described by the message, e.g., a message indicating that the maximum air temperature $T_{\text{max}} = 70^\circ \text{F}$ in Rochester on New Years’ Day contains more information than a message that $T_{\text{max}}[\text{January 1}] = 25^\circ \text{F}$. The reason is probably obvious; that $T_{\text{max}} = 70^\circ \text{F}$ occurs rarely (if ever) in Rochester in January (at least until global warming really kicks in!). In words, we can define the importance of the data in the message by noting that “the norm isn’t “news” but a rare occurrence is.”

In 1947, Shannon strictly defined the quantity called “information” to satisfy two basic requirements that make intuitive sense. In so doing, he began the study of information theory, which is a part of applied probability theory. Though Shannon’s original definition of information is very intuitive, it must be generalized to provide a more complete description of the concept.

The concept of information may be interpreted as the minimum number of questions with a binary answer set (the only possible outcomes are yes/no or 1/0) that must be answered before the correct message may be determined. As thus defined, information is synonymous with the removal of uncertainty. The uncertainty about the result of experiment $X$ increases with the number of possible outcomes $n$; if more distinct outcomes are possible, there is more information in a message stating which outcome occurs. The information $I$ about $X$ should be a monotonically increasing
20.3 IMAGE COMPRESSION

function of \( n \), and thus a monotonically decreasing function of the probability of a specific occurrence.

Shannon defined information by establishing intuitive criteria that the concept should obey and finding a mathematical expression that satisfies these requirements. As a first example, consider the simplest case of an experiment \( X \) which has \( n \) \textit{equally likely} possible outcomes (i.e., the probability of each outcome is \( n^{-1} \)). The information about the result of experiment \( X \) must be related to \( n \); if \( n \) is large, then the information about the specific outcome should be large as well. If \( X \) may be decomposed into two independent experiments \( Y \) (with \( n_1 \) equally likely possible outcomes) and \( Z \) (with \( n_2 \) equally likely outcomes), then:

\[
 n = n_1 \cdot n_2 \tag{1}
\]

and:

\[
 p[X] = \frac{1}{n} = p[Y] \cdot p[Z] = \frac{1}{n_1} \cdot \frac{1}{n_2} = \frac{1}{(n_1 \cdot n_2)} \tag{2}
\]

Thus the information \( I \) about the outcome of experiment \( X \) should be a function \( f \) that satisfies the requirement:

\[
 I[X] = f[n] = f[n_1 \cdot n_2] \tag{3}
\]

In addition, the information about a composite of independent experiments should be equivalent to the sum of the information of the component experiments. This establishes the second criterion for \( I[X] \):

\[
 I[X] = I[Y] + I[Z] = f[n_1] + f[n_2] \tag{4}
\]

Thus the information in a message describing the outcome of experiment \( X \) with \( n \) possible outcomes must satisfy:

\[
 I[X] \implies f[n] = f[n_1 \cdot n_2] = f[n_1] + f[n_2] \tag{5}
\]

One appropriate function that satisfies requirement 5 is the logarithm:

\[
 I[X] \implies f[n] = c \log_b (n) = -c \log_b \left( \frac{1}{n} \right) = -c \log_b (p[X]) \tag{6}
\]

where \( c \) is some constant (which can be assumed to be unity for now) and \( b \) is the base of the logarithm such that \( b > 1 \). In fact, Khinchin proved that the logarithm is the \textit{only} continuous function that satisfies the required properties for any finite number \( n \) of possible outcomes (symbols). This definition of information satisfies the additivity requirement for information, i.e.,

\[
 I[X] = \log_b [n] = \log_b [n_1 \cdot n_2] = \log_b [n_1] + \log_b [n_2] = I[Y] + I[Z] \tag{7}
\]

The units of information are \textit{base-}b \textit{digits}, e.g., decimal digits 0–9 for \( b = 10 \) and \textit{(bi)nary digits}, or \textit{bits}, for \( b = 2 \).
For a simple example, consider the quantity of information in a statement about
the result of a fair coin (so that \( p_H = p_T = 0.5 \)). The number of equally likely
outcomes is \( n = 2 \) and thus the information in the message about one “toss” is:

\[ I[X] = \log_2 2 = 1 \text{ bit} \]  

(8)

If the coin has two (indistinguishable!) heads, then there is only one (“equally likely”) outcome and \( p_H = 1, p_T = 0 \). The information content in a statement about the outcome of a toss of a two-headed coin is:

\[ I[X] = \log_2 1 = 0 \text{ bits} \]  

(9)

Shannon’s original definition of information in eq. 6 applies only to equiprobabilistic outcomes and must be generalized if the probabilities of the different outcomes are not equal. Consider the case of an unfair coin with \( p_H \neq p_T \) (both \( p_H > 0 \) and \( p_T > 0 \)). Recall that \( p_H \equiv \frac{N_H}{N} \) for large \( N \). This is an intermediate case between the certain outcome case (\( p_T = 0 \implies 0 \) bits of information per toss), and the fair coin (\( p_H = p_T = 0.5 \implies 1 \) bit of information per toss). Intuitively, a message that specifies the outcome when flipping such a coin will convey \( \alpha \) bits of information where \( 0 < \alpha < 1 \). The outcome of the unfair coin may be considered to be the result of a cascade of several experiments with equally likely outcomes, but where several outcomes are equivalent. For example, consider the case of an unfair coin where \( p_H = 0.75 \) and \( p_T = 0.25 \). The outcome of a single flip can be considered to have four equally likely outcomes \( A - D \), i.e.:

\[ p_A = p_B = p_C = p_D = 0.25, \]

where outcomes \( A, B, C \) are heads and outcome \( D \) is a tail. The quantity of information in a single experiment with four equally likely outcomes is:

\[ I[X] = \log_2 4 = -\log_2 \left( \frac{1}{4} \right) = 2 \text{ bits per toss} \]

However, since the probabilities of the two distinguishable outcomes are not equal, there must be excess information in statements about the identical outcomes that must be subtracted from the 2 bits. The excess information in the message about a head is \( \log_2 \left[ n_H \right] \) multiplied by the probability of a head \( p_H \). Similarly for a tail, the excess information is \( p_T \log n_T \), where \( p_T = \frac{n_T}{n} \) and \( n = n_H + n_T \).

\[ \text{Excess Information for head} = \frac{3}{4} \log_2 3 = \frac{3}{4} \cdot 1.585 = 1.189 \text{ bits} \]

\[ \text{Excess Information for tail} = \frac{1}{4} \log_2 1 = \frac{1}{4} \cdot 0 = 0 \text{ bits} \]
After substituting these results, the total information is:

\[ I [X] = \left( \log_2 [4] - \frac{3}{4} \log_2 [3] - \frac{1}{4} \log_2 [1] \right) \text{ bits} \approx 2 - 1.189 = 0.811 \text{ bits} \]

The general expression for information content in the case of the unfair coin with probabilities \( p_H \) and \( p_T \) is:

\[
I [X] = \log_2 [n] - (p_H \log_2 [n_H] + p_T \log_2 [n_T]) = (p_H + p_T) \log_2 [n] - p_H \log_2 [n_H] - p_T \log_2 [n_T]
\]

\[= -p_H (\log_2 [n_H] - \log_2 [n]) - p_T (\log_2 [n_T] - \log_2 [n])\]

\[= -p_H \log_2 \left( \frac{n_H}{n} \right) - p_T \log_2 \left( \frac{n_T}{n} \right)\]

\[= -p_H \log_2 [p_H] - p_T \log_2 [p_T]. \quad \text{(10)} \]

Since \( p_i \leq 1 \), \( \log p_i \leq 0 \) and \( I [X] \geq 0 \); the quantity of information is positive. For \( M \) possible outcomes with probabilities \( p_i \), this definition of information can be extended:

\[ I [X] = - \sum_{i=0}^{M-1} (p_i \log_2 [p_i]), \text{ subject to the constraint } \sum_{i=0}^{M-1} p_i = 1. \quad \text{(11)} \]

Any “impossible” outcome (with \( p_i = 0 \)) is ignored to eliminate the problem of calculating the logarithm of “0”.

If the \( M \) outcomes are equally likely, so that \( p_i = M^{-1} \), we have:

\[
I [X] = - \sum_{i=0}^{M-1} \left( \frac{1}{M} \log_2 \left[ \frac{1}{M} \right] \right) = -M \cdot \left( \frac{1}{M} \log_2 \left[ \frac{1}{M} \right] \right)
\]

\[= - \log_2 \left[ \frac{1}{M} \right] = + \log_2 [M] = - \log_2 [p_i], \quad \text{(12)} \]

as before. This demonstrates that the generalized definition of information is consistent with the earlier one.

For the case of the unfair coin with probabilities \( p_H = 0.75 \) and \( p_T = 0.25 \), the quantity of information in a statement about the coin toss is:

\[ I [X] = -0.75 \log_2 [0.75] - 0.25 \log_2 [0.25] \approx 0.811 \text{ bits} \quad \text{(13)} \]

Now, you may be wondering what a fractional number of bits of information actually “means?” It can be considered as the average uncertainty that is removed by the message of the outcome of the experiment \( X \). If we toss the fair coin 100 times, the resulting string of outputs (e.g., \( HTTHHHHTHT \cdots \)) may be transmitted with 100 bits. The outputs of 100 tosses of the two-headed coin requires 0 bits to transmit while the outcome of 100 tosses of the unfair coin may be specified by \( 100 \cdot 0.811 \approx 82 \) bits, if the proper coding scheme is used. A method for reducing the number of bits of such a message will be described shortly.
It is very important to note that Shannon’s definition of information assumes that the outcomes of a particular experiment are random selections from some particular probability distribution. If we flip a “256-sided” fair coin a large number of times (say $N = 512 \times 512 = 262,144$ “flips”), we can generate an “image” with $N$ pixels, each with 256 “outcomes” (gray levels) that will be approximately equally populated. If the gray level of some specific pixel is 4, then the gray levels of its neighbors are (by assumption) as likely to be 199 or 255 as 3 or 5. Such a system for generating data may be called a discrete memoryless source or DMS because an individual output from the system (pixel gray value) is independent of previous outputs. However, we know that adjacent pixels in pictorial images of real scenes usually belong to the same object and tend to have similar gray levels. Thus the gray values of adjacent pictures are usually correlated. The effect of correlated pixels on the information content of imagery will be considered shortly.

The measure of information (eq.11) often is called the entropy because its form is identical to the quantity that appears frequently in statistical thermodynamics. The entropy of a body of a gas depends on its volume, mass, and (most of all) its temperature. The energy of the gas also depends on these three properties. One of the steps in the Carnot cycle of the ideal heat engine allows a gas to expand within a thermally insulated confined space by pushing against a (slowly moving) piston. The insulation ensures that no heat flows into or out of the gas. The expansion requires that the gas lose some of its thermal energy through cooling; this is the principle of the refrigerator or air conditioner. The lost thermal energy performs work on the piston. Such a process in which no heat flows into or out of the system is called adiabatic and is reversible; the piston may do work on the gas, thus raising the temperature. The entropy of the system is unchanged by a reversible process.

In real life, physical interactions are not reversible. For example, no gas expansion is truly adiabatic because the process cannot be thermally insulated perfectly from its surroundings. Real gas expansions increase the entropy of the system. Consider a gas that is confined in a box that is divided in equal parts by a removable partition. A gas is confined in one half and a vacuum exists in the other. Until the partition is removed, the gas is capable of performing useful work through expansion by pushing on a piston. However, if the partition is removed suddenly, the gas expands over time to fill the entire container without performing any work. No thermal energy is lost. However, restoration of the system to its original state would require work to be done on the gas to push it back into the original volume; the system is not reversible. We say that the expansion of the gas increased the entropy of the system because the process was not reversible. In this way, the entropy is a measure of the capability of the system to perform useful work by changing thermal energy into mechanical energy. Equivalently, it is a measure of the disorder of the system. The original system was more ordered before removing the partition because we had more knowledge about the location of the molecules of the gas, thus the entropy was lower before the expansion. The entropy is a statistical description, we do not have a complete description of the location and velocity of each molecule either before or after removal of the partition.

The concept of entropy as a measure of disorder is the more applicable to the current problem. The energy of a set of coin flips (or the arrangement of gray levels
in an image) is determined by the statistics of the outcome; by the histogram of results. The entropy of the set of coin flips or image gray levels is determined by the uncertainty in their arrangement given knowledge of the statistics (the histogram). The less likely the result, the more information in a message that the result has occurred or will occur.

### 20.3.2 Information Content of Images

An image can be considered as the output of an ensemble of pixels (experiments) whose results are derived from a discrete set of possibilities (the histogram). It is the histogram of the image with $M$ gray levels which determines the quantity of information as defined by Shannon via the entropy equation:

$$I[X] = - \sum_{i=0}^{M-1} p_i \log_2 [p_i]$$  \hspace{1cm} (14)

For most realistic monochrome images digitized to 8 bits per pixel, the entropy is in the range of 4-6 bits per pixel. To further investigate the meaning of Shannon’s definition of information, consider that an image with $N$ pixels and $M$ gray levels can be produced from $N$ experiments with each having $M$ possible outcomes for a total of $M^N$ possible distinct images. In the binary image case, the gray values may be produced by $N$ coin flips. For a $1 \times 1$ image, $N = 1$ and the number of possible images is $2^1 = 2$ (0 and 1), and there are also two possible histograms (1,0) and (0,1). There are $2^2 = 4$ possible two-pixel binary images (11, 10, 01, and 00) and three possible histograms (2,0), (1,1), and (0,2). Note that there are two images with histogram (1,1). For $N = 3$ pixels, there are $2^3 = 8$ distinct images (111, 110, 101, 011, 100, 010, 001, 000) and four possible histograms (3,0), (2,1), (1,2), and (0,3). There is one image each with histogram (3,0) or (0,3), and three images each with histogram (2,1) and (1,2). The progression of the number of possible binary images divided into a number of distinct histograms specified by the set of binomial coefficients:

$$N[N_0, N_1] \equiv \frac{N!}{N_0! \cdot N_1!}$$

where $N$ is the number of pixels in the image, and $N_0, N_1$ are the number of pixels with level 0 and 1, respectively. Note that the constraint $N_0 + N_1 = N$ must be satisfied. The array of the binomial coefficients defines Pascal’s triangle as shown below, where the row number represents the number of pixels in the image, the sum of the numbers in the row is the number of possible images, the number of groups in each row is the number of possible histograms, and the number in each group is the number of images that has a particular histogram:
Number of binary images with different histograms for $N=1,2,3,4,5,6$

For example, there are $1 + 4 + 6 + 4 + 1 = 16$ images composed of four binary pixels with five different histograms. One image has histogram $(4,0)$, one with $(0,4)$, four each with histograms $(3,1)$ and $(1,3)$, and six images with histogram $(2,2)$. Note that the number of images for a specific histogram increases as the population of gray values becomes more equally distributed, i.e., toward the center of the row in Pascal’s triangle. As we will soon demonstrate, the information content of an image with a specified histogram is largest when the histogram populations are equal, i.e., for the histogram with the largest number of possible images and thus the greatest uncertainty of image content.

For images with $M$ gray levels ($M > 2$), calculation of the number of possible images and histograms is somewhat more complicated. The multilevel analog of the binomial coefficient used in Pascal’s triangle is the multinomial coefficient:

$$N [N_0, N_1, N_2, N_3, \ldots, N_M] \equiv \frac{N!}{N_0! N_1! N_2! \ldots N_{M-1}!}$$

where $N$ is the number of pixels and $N_0, N_1, N_2, \ldots, N_M$ are the populations of each gray level subject to the constraint $N_0 + N_1 + N_2 + \ldots + N_M = N$. For $N = 8$ pixels and $M = 4$ possible gray levels (0-3; 2 bits of quantization), the number of possible images is $4^8 = 65,536 = 64K$. The number of distinct 8-pixel images with histogram $[N_0, N_1, N_2, N_3]$ is:

$$N [N_0, N_1, N_2, N_3] = \frac{8!}{N_0! N_1! N_2! N_3!},$$

where $N_0 + N_1 + N_2 + N_3 = 8$. For example, if the histogram is known to be $(4,4,0,0)$, the number of possible 8-pixel 2-bit images is:

$$N [4, 4, 0, 0] = \frac{8!}{4! 4! 0! 0!} = 70.$$

Other specific histograms yield $N [3, 3, 1, 1] = 1120$ possible images, $N [3, 2, 2, 1] = 1680$ images, and $N [2, 2, 2, 2] = 2520$ images. Again, note that the number of distinct
images increases as the histogram becomes “flatter”. Also recall that Shannon’s definition of information is determined by the image histogram. Because there are more distinct images with a flat histogram than with a clustered histogram, the “amount” of information should be larger for an image with a flat histogram. An image with a flat histogram indicates that there is maximum uncertainty and maximum information content, and thus requires the most bits of data for complete specification.

Now consider the number of possible images with specific histograms for more “reasonable” image formats. A common size for a monochrome digital image is \( N \times N = 512^2 \) pixels and \( M = 256 \) levels, so that the number of distinguishable images is:

\[
M^{N \times N} = 256^{512^2} = 2^{(8 \cdot 2^{29})} = 2^{2,097,152} = (10^{\log_{10}[2]})^{2,097,152} \approx (10^{0.30103})^{2,097,152} = 10^{0.30103 \cdot 2,097,152} \sim 10^{631,306.66}
\]

The size of this number may be gauged against the estimates that there are \( 10^{78} \) atoms and \( 10^{88} \) cubic millimeters in the universe. Of these MANY images, only a single one has the histogram with all pixels clustered at gray level 0. There are \( 512^2 \) possible images with one pixel at gray level 1 and the rest (262,143) at gray level zero. The number of images with two pixels at gray level 1 and 262,142 at 0 is:

\[
N[262142, 2, 0, \cdots, 0] = \frac{(262144)!}{(262142! \cdot 2! \cdot 0! \cdots 0!)} \approx 3.44 \cdot 10^{10}
\]

If we continue this progression and add more populated gray levels to “flatten” the histogram, the number of distinguishable images increases very rapidly. Such an image with a perfectly flat histogram has \( 512^2/256 = 1024 \) pixels in each level; the number of such images is:

\[
N[1024_0, 1024_1, \cdots, 1024_{255}] = \frac{(512^2)!}{(1024!)^{256}} \approx 10^{630,821}.
\]

The approximation was derived via Stirling’s formula for \( N! \) where \( N \) is large:

\[
\lim_{N \to \infty} [N!] \approx \sqrt{2\pi N} \cdot N^N \cdot e^{-N}
\]

We should check this formula; substitute \( N = 10 \) to find that

\[
10! \approx 3.5987 \times 10^6
\]

whereas the actual result is \( 10! = 3.6288 \times 10^6 \), so the error is only about a factor of \( 10^{-3} \).

If 254 levels are populated with 1028 pixels each and one level has 1032 pixels (a “slightly clustered” histogram), the number of images is:

\[
N[1028_0, 1028_1, 1028_2, \cdots, 1028_{253}, 1032_{254}, 0_{255}] = 512^2\frac{!}{(1028!)^{254} \cdot 1032! \cdot 0!} \approx 10^{630,377},
\]
which is smaller than the number of images with a flat histogram (by a factor of $10^{444}$, which is still pretty large!). Note that the number of possible images again is maximized when the histogram is flat; the uncertainty, the information content, and thus the number of bits to specify the image are maximized when the histogram is flat.

### 20.3.3 Maximizing Image Information

We just showed by example that an image with a flat histogram has the maximum possible information content. The rigorous proof of this assertion is an optimization problem.

Given that the probability distribution of gray levels for $f[n, m]$ is proportional to the image histogram $H[f]$.

$$p_i = \frac{H[f]}{N}$$

where $N$ is the number of pixels in the image. Since the maximum possible population of a gray level is the total number of pixels (i.e., $N$), $0 \leq p_i \leq 1$ as required. The problem is to find the set of gray-level probabilities $\{p_i\}$ for $M$ levels that maximizes information content:

$$I[f] = - \sum_{i=0}^{M-1} p_i \log_b [p_i],$$

subject to the constraint:

$$\sum_{i=0}^{M-1} p_i = 1.$$ 

To maximize $I$, we set its total derivative equal to zero:

$$dI = \frac{\partial I}{\partial p_0} dp_0 + \frac{\partial I}{\partial p_1} dp_1 + \frac{\partial I}{\partial p_2} dp_2 + \cdots + \frac{\partial I}{\partial p_{M-1}} dp_{M-1} = 0.$$ 

subject to the constraint that the probabilities sum to a constant:

$$d \left( \sum_{i=0}^{M-1} p_i \right) = 0.$$ 

This optimization problem is easily solved by Lagrangian multipliers. If we maximize a linear combination of $I[f]$ and $\sum p_i$, we will automatically maximize $I[f]$. Construct a function $L[f]$ to be maximized:

$$L[f] = I[f] - \lambda \sum_{i=0}^{M-1} p_i = - \sum_{i=0}^{M-1} p_i \log_b [p_i] - \sum_{i=0}^{M-1} \lambda p_i = - \sum_{i=0}^{M-1} p_i [\log_b [p_i] + \lambda].$$

where $\lambda$, the Lagrangian multiplier, is a constant to be determined. To maximize $L$,
we set its total derivative equal to zero:

\[ dL = \sum_{i=0}^{M-1} \frac{\partial L}{\partial p_i} dp_i = 0 \]

Because the differential probabilities \( p_i \) are arbitrary, a necessary and sufficient condition for \( dL \) to vanish is to have the component derivatives \( \frac{\partial L}{\partial p_i} \) vanish individually:

\[
\frac{\partial L}{\partial p_i} = \frac{dp}{dp_i} \cdot (\log_b [p_i] + \lambda) + p_i \cdot \frac{\partial}{\partial p_i} (\log_b [p_i] + \lambda)
\]

\[
= (\log_b [p_i] + \lambda) + p_i \cdot \frac{1}{p_i} = 1 + \log_b [p_i] + \lambda = 0, \text{ for all } i
\]

\[
\log_b [p_i] = - (1 + \lambda)
\]

\[ \implies p_i = e^{-(1+\lambda)} \], where \( \lambda \) is constant.

Note that the probability \( p_i \) for the occurrence of the \( i^{th} \) level is constant, thus the probabilities of all outcomes must be equal when \( L [f] \) (and hence \( I [f] \)) is maximized. The constraint allows us to calculate the numerical value of the Lagrangian multiplier:

\[ \sum_{i=0}^{M-1} p_i = 1 = \sum_{i=0}^{M-1} b^{-(1+\lambda)} = M b^{-(1+\lambda)} = M p_i = M p \implies \left( p = \frac{1}{M} \right) \]

In words, the information content \( I [f] \) of an image \( f[n,m] \) is maximized when all gray levels are equally populated, \( i.e., \) when the histogram is flat. Slow variations in gray level from pixel to pixel are most visible in an image with a flat histogram. This is (of course) the motivation for “histogram equalization” in digital image processing. We should note that histograms of quantized images cannot be “equalized” in a strict sense unless the gray values of some pixels are changed to fill underpopulated bins. This can be performed only as an approximation by considering the gray values of neighboring pixels.

### 20.3.4 Information Content (Entropy) of Natural Images

The redundancy \( R \) in a message was already defined as a dimensionless parameter based on the ratio of nonredundant data to the total data:

\[ R = 1 - \frac{\text{Nonredundant Data}}{\text{Total Data}} \]

Because the human visual system is constrained by transmission or storage capacity, it is useful to reduce/remove data redundancy beforehand. For example, the total resolution (spatial, temporal, and color) ultimately is limited by the transmission capacity of the channel connecting the eye and brain, and removal of such redundancies is a principal task of the vision system. Examples of HVS mechanisms that remove redundancy include lateral inhibition of the neural net, opponent-color processing,
and the nonlinear response of the vision receptors.

In *Predictability and redundancy of natural images* (JOSA A, Vol. 4, 2395, 1987) D. Kersten estimated the redundancy of some natural images by presenting to observers eight $128^2$ 4-bit images (16,384 pixels, each with 16 gray levels stretched over the available dynamic range). The 4-bit quantization ensured that the images had significant redundancy and precluded the presence of significant additive noise in the quantized values (i.e., additive noise would be typically quantized to zero). The images ranged in content from “simple” to “busy” and included natural scenes. The pixels subtended a rectangle of size $10 \times 7$ arcminutes to the observer. A predetermined fraction of the pixels were deleted and replaced with one of the 16 gray levels chosen at random. For example, the original image was in fact a random image, then it would be impossible to determine which pixels were altered; such an image would have no redundancy. If the original image were pictorial (e.g., a human face), then most of the replaced pixels will be obvious; such an image has significant redundancy. Observers guessed at the gray value of the replaced pixel until the correct value was assigned. The larger the number of guesses, the more uncertain the gray level of the pixel. The histogram of the number of guesses was used to determine the entropy $I[f]$. The number of guesses at each replaced pixel for the random image should be large (mean value of 8), and small for the pictorial image. Therefore, the redundancy is:

$$R[f] = 1 - \frac{I[f]}{I_{\text{max}}} = 1 - \frac{I[f]}{4 \text{ bits}}$$

where $I[f]$ is the entropy of image $f$ and $I_{\text{max}}$ is the number of quantization bits. Kersten’s results indicated that the eight images have redundancies in the interval:

$$0.46 \text{ ("busy" picture)} \leq R \leq 0.74 \text{ (picture of face)}$$

The corresponding content of nonredundant information is:

$$2.16 \text{ bits per pixel} \geq I[f] \geq 1.04 \text{ bits per pixel}$$

Based on the previous discussion, the number of $128^2$ 4-bit images is:

$$\text{Number} = 2^{4 \cdot 128^2} = 2^{65,536} \cong 10^{19,728}$$

But the number of natural images would lie in the range:

$$2^{2.16 \cdot 128^2} \cong 2^{35,389} \cong 10^{10,653} \geq \text{Number} \geq 2^{1.04 \cdot 128^2} \cong 2^{17,0397} \cong 10^{5,129}$$

which is STILL many times larger than the estimated number of atoms in the universe!
20.4 Lossless Compression

20.4.1 Run-Length Encoding

Run-length encoding is a very simple method for compression of sequential data (and particularly for binary data) that takes advantage of the fact that consecutive single “tokens” (gray values) are often identical in many data sets. Run length encoding inserts a special token each time a chain of more than two equal input tokens are found. This special input advises the decoder to insert the following token n times into his output stream. For example, consider the sequence of 3-bit data:

```
7 7 7 2 6 6 6 6 2 2 5 5 5 5 5 5 5 5 5
```

The encoded RLE image would be:

```
7 3 2 1 6 5 2 2 5 9
```

where the first digit in the encoded image is the gray value of the first pixel, the second digit is the number of occurrences, etc. The sequence is reduced from 20 3-bit numbers (60 bits) to 10 digits (though the long run of “5” at the end requires a 4-bit number to encode). This means that this sequence might be encoded into 5 3-bit numbers and 5 4-bit numbers, for a total of 35 bits. If the system were limited to three bits, the sequence of 9 examples of level “5” would be split into one sequence of 7 and one of 2:

```
7 3 2 1 6 5 2 2 5 7 5 2
```

for a total of 10 3-bit numbers, or 30 bits.

The compression in RLE occurs when the image exhibits strings of the same value. If the image is “noisy”, then the image after RLE coding will likely require more bits than the uncompressed image.

The common “bitmap” image format (.BMP) uses run-length encoding.

20.4.2 Huffman Code

We have already seen that the information content of an image is a function of the image histogram; an image with a clustered histogram contains less information than an image with a flat histogram because the former contains statistically redundant information. This implies the existence of methods to encode the information with fewer bits while still allowing perfect recovery of the image. The concept of such a code is quite simple: the most common gray levels are assigned to a code word that requires few bits to store/transmit, while levels which occur infrequently are encoded with many bits; the *average number* of bits per pixel is reduced by attempting to equalize the number of bits per gray level. A procedure for obtaining a code from the histogram was specified by David A. Huffman in 1951 while he was a 25-year-old
graduate student at MIT. Huffman developed the coding scheme as part of a final assignment in a course on information theory given by Prof. Robert A. Fano. After working unsuccessfully on the problem for months, the solution came to Huffman as he was tossing his notebooks into the trash (related in *Scientific American*, September 1991, pp. 54-58). He presented his method to Fano, who “Is that all there is to it?” Huffman’s coding scheme is now ubiquitous. (This story must be a metaphor for something.) (Huffman, DA., *Proc. IRE* **40**, 1098-1101, 1952).

The Huffman code assumes that the gray values are selected at random from some known probability distribution. In other words, it assumes that the gray values of adjacent pixels are unrelated (which is, of course, not true for meaningful pictorial images, though perhaps true for images transformed to a different coordinate system). A source of uncorrelated random numbers is called “memoryless”.

The entropy of the Huffman-coded image always is within 1 bit per pixel of the information content defined by Shannon. The Huffman code removes bits from the message by discarding *objective redundancy*. It is *lossless* and *unambiguous*. This first quality means that the original image may be reconstructed without error from knowledge of the coded image and the code book. The quality of no ambiguity ensures that only one set of gray levels may be decoded from an ungarbled string of binary digits encoded by the Huffman procedure.

The Huffman code is perhaps best described by example. The pixels in an image with 8 gray levels are usually defined specified by a binary code that requires 3 bits per pixel. For attempted clarity, the gray levels will be indicated by alphabetic characters:

\[
\begin{array}{ll}
\text{Decimal.} & \text{Binary} \rightarrow \text{Alphabetic} \\
0. & = 000_2 \rightarrow A \\
1. & = 001_2 \rightarrow B \\
2. & = 010_2 \rightarrow C \\
3. & = 011_2 \rightarrow D \\
4. & = 100_2 \rightarrow E \\
5. & = 101_2 \rightarrow F \\
6. & = 110_2 \rightarrow G \\
7. & = 111_2 \rightarrow H \\
\end{array}
\]

Consider a 3-bit 100-pixel image with levels distributed as in the following histogram:

\[ H [A, B, C, D, E, F, G, H] = [0, 9, 12, 40, 30, 5, 4, 0] \]

The probability of each gray level therefore is:

\[ p[A] = 0, \ p[B] = .09, \ p[C] = 12, \ p[D] = .40, \ p[E] = 30, \ p[F] = .05, \ p[G] = .04, \ p[H] = 0 \]
The information content in this image obtained from Shannon’s entropy formula:

\[ I = -0.4 \log_2 [0.4] - 0.3 \log_2 [0.3] - \cdots \]
\[ \cong 0.529 + 0.521 + \cdots + 0 \]
\[ \cong 2.131 \text{ bits/pixel} \]

which is considerably less than the 3-bit quantization. To derive the Huffman code, arrange the occupied levels in descending order of probability. One bit is used to distinguish the least likely pair of levels using either the convention that the binary digit 1 is assigned to the more probable level and 0 to the less probable, or vice versa. This is the least significant bit in the Huffman code, as it distinguishes the most rarely occurring gray levels. The probabilities of the two rarest levels are then summed (giving a total probability of 0.09 in this case) and the list of gray values is rearranged in descending order. In this example, the sum of the probabilities of the rarest levels \( F \) and \( G \) is equal to the probability of the next level \( B \), and so reordering is not required. The new pair of least-likely levels in the rearranged list are distinguished by again assigning a binary digit using the same convention as before. The last two probabilities are summed, the list is reordered, and the process continues until bits are assigned to the last pair of probabilities. The last binary digit derived using this procedure is the most significant bit. The schematic of the entire sequence is shown in the figure:

Calculation of Huffman code for 3-bit image with probabilities as listed in the text. The gray levels are arranged in descending order of probability. One bit is used to distinguish the rarest levels, the probabilities are summed, the list is rearranged in descending order, and the process continues.

The Huffman code for a specific gray level is the sequence of binary digits assigned from right to left for that level, i.e., from most significant to least significant bit. The code for level \( B \) is obtained by following the sequence from the right: the path for level \( B \) is in the upper branch on the far right (code = 1), and the lower branch at two more junctions (codes = 0 and 0) so that the code for level \( B \) is 10112. The codes for the other levels are found similarly and the resulting code book is listed above. The number of bits to encode all pixels having each gray level may be calculated, and their sum is the average number of bits per pixel required to encode the entire image:
To demonstrate the unambiguous nature of the code, consider the sequence of 14 pixels “CDDEEBEDFGCDEC”, which encodes to a sequence of 35 bits (2.5 bits per pixel)

\[
CDDEEBEDFGCDEC = 10000111110111101011010011100
\]

The message is decoded by examining the order of bits. Since the first bit is not a “0”, it cannot be the most common gray value “D”. Since the second bit is not a “1”, the first pixel cannot be an “E”. The third bit is “0”, and therefore the first pixel must be “C”

\[
\begin{align*}
100 & \quad 0111110111101011010010011100  \\
& \quad 10001111110111101011010010011100
\end{align*}
\]

Now repeat: the fourth bit is “0”, and only “D” has this first bit:

\[
\begin{align*}
100 & \quad 0111110111101011010100100011100  \\
& \quad 1001011110111101110101100100011100
\end{align*}
\]

Ditto, “D” is the third character:

\[
\begin{align*}
100 & \quad 0111110111101011010100100011100  \\
& \quad 1001011110111101110101100100011100
\end{align*}
\]

The fourth pixel begins with “11”, and therefore is “E”:

\[
\begin{align*}
100 & \quad 0111110111101011010100100011100  \\
& \quad 1001011110111101110101100100011100
\end{align*}
\]

Ditto for the fifth”

\[
\begin{align*}
100 & \quad 0111110111101011010100100011100  \\
& \quad 1001011110111101110101100100011100
\end{align*}
\]

The sequence continues until all bits have been decoded:
This example also may be used to illustrate the pitfall of Huffman coding. If a bit is garbled ("flipped"), then it is likely that many (if not all) of the following characters in the message will not be decodable, because the redundancy in the message has been removed. Consider the example where the sixth bit is flipped from a “1” to a “0”:

\[
1000\overline{0}111011101010110100100011100
\]

In this case, the decoded message would be:

\[
1000\overline{0}111011101010110100100011100
\]

instead of:

\[
100001110111010101010100100011100
\]

The “flipped” bit resulted in the incorrect decoding of the sample “B” as two samples “DE”.

This code exceeds the theoretical limit of Shannon information by only 0.04 bits/pixel, and reduces the number of bits required to store the image to (2.17/3) bits, or 72% of that required for the uncompressed original. The efficiency of the Huffman code is defined as the ratio of the average number of bits per pixel for the code to the theoretical limit determined by Shannon’s definition. In this case, the compression efficiency is:

\[
\eta = \frac{2.131}{2.17} = 0.982
\]

For real images, lossless compression via a Huffman code on a pixel-by-pixel basis can achieve compression efficiencies in the range of 0.67 ≤ η ≤ 0.25, a modest improvement. Note that the code book must be available to the receiver to allow reconstruction of the image. This extra data is known as the overhead of the compression scheme and has not been included in the calculations of compression efficiency. The length of the Huffman codeword of a gray level whose histogram probability is p is \(-\log_2 [p]\), e.g., if the probability is 0.125, the ideal codeword length is 3 bits. If \log_2 [p] is significantly different from an integer (e.g., \(p = 0.09 \Rightarrow -\log_2 [p] = 3.47\) bits), the coding efficiency will suffer.

A significant shortcoming of a Huffman code is that it cannot adapt to locally varying image statistics, or equivalently, a particular Huffman code will not be optimum for a variety of images. In fact, inappropriate application of a Huffman code can actually lead to an increase in the storage requirements over the original, e.g., if the image contains many levels which were rare in the original source image.
20.4.3 Information in Correlated Images – Markov Model

We’ve just seen how the Huffman code takes advantage of clustering of the histogram to reduce the number of bits required to store/transmit an image whose gray levels are not equally populated, and assuming that the gray values of pixels are obtained from a discrete memoryless source, i.e., the gray value at a pixel is a number selected at random from the probability distribution (the image histogram). Obviously, this last assumption is false for most (if not all) realistic images consisting of objects whose component pixels have similar properties (e.g., gray level, color, or texture). These correlations provide a context for a pixel and add additional redundancy, which may be exploited to achieve additional compression. Redundancy may be considered as creating clusters in particular histograms generated from the original images. Examples of redundancy include similarities of pixel gray level in local neighborhoods (interpixel redundancy) in a single image, that may be exploited by constructing codes for groups of pixels of a single image (vector coding) or by coding linear combinations of blocks of pixels, as in the JPEG standard. Similarities in color (spectral redundancy) generate clusters in the multispectral histogram, thus reducing data-transmission/storage requirements in color images. This clustering is used by the NTSC video transmission standard and the Kodak PhotoCD™. Correlations of corresponding pixels across image frames in a motion picture (temporal redundancy) allows significant additional compression and is exploited in the MPEG standards. In addition, images meant for human viewing may be compressed by removing image content that is not visible to the eye; these subjective redundancies or superfluous information are present the spectral, spatial, and temporal dimensions, and are utilized for additional compression in all of the consumer compression standards, such as JPEG, MPEG, and PhotoCD™.

The statistical properties of a correlated image are more complicated than those of a discrete memoryless (random) source, and one of the difficulties in developing efficient standard algorithms for image compression is the creation of a mathematical model of these various redundancies. The simplest model of interpixel redundancies is the Markov source, where the probability that the pixel located at coordinate \( n \) has gray level \( f \) is a function of the gray level at some number of neighboring pixels. The number of these neighboring pixels is the order of the Markov source; the higher the order, the more correlated are the gray values in a neighborhood and the more predictable the image content from previous levels. Thus we should be able to define a measure of information for a Markov source that is less than the normal entropy. For a correlated source model, such as a Markov source of order 1, we can define the first-order, or conditional, entropy of the image with \( M \) gray levels as:

\[
I [f_k|f_{k-1}] = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} p_{i,j} \log_2 [p_{i,j}|p_j]
\]

where \( f_k \) and \( f_{k-1} \) are the gray levels of the \( k^{th} \) and \((k-1)^{st}\) pixels and \( I [f_k|f_{k-1}] \) is the information in the \( k^{th} \) pixel \( f_k \) given that the \((k-1)^{st}\) pixel has gray value \( f_{k-1} \). The conditional probability of \( p_{i,j} \) given \( p_j \) is denoted \([p_{i,j}|p_j] \). It may be clearer
to think of the conditional entropy as just the information content of the set of 2-D gray-level “vectors” $\mathbf{f} = [f_1, f_2]$

$$I[\mathbf{f}] = - \sum_{\mathbf{f}} p[\mathbf{f}] \log_2 [p[\mathbf{f}]]$$

where the sum is over all the 2-D pixel vectors $\mathbf{f}$. Note that the total number of gray-level states of the 2-D histogram is the square of the number of gray levels; the vector histogram of an image with $M$ levels has $M^2$ bins.

The first-order entropy is the average information content of the experiment at the $k^{th}$ pixel given that the $k-1^{st}$ pixel gray value is known. If the gray values are correlated, the first-order entropy per pixel may be much less than the Shannon entropy (pixels considered independently); in other words, the vector histogram of the image is clustered. Note that the Shannon entropy may be considered as the zeroth-order, or scalar entropy of the image. If the pixel gray levels are independent (random DMS source), then the first-order entropy will be $\log_2 [M^2] = 2 \cdot \log_2 [M]$ bits per pixel because every pair of gray levels will be equally likely; the 2-D vector histogram is thus flat, i.e., all 2-D vectors (whose components are gray levels of adjacent pixels) will be equally likely. The image from a first-order Markov source may have a flat histogram, but the 2-D histogram of neighboring pixels may be clustered; 2-D vector coding of such an image will exhibit significant data compression.

### 20.4.4 “Vector” Coding (Compression)

Images generated by a Markov source often may be compressed by examining the histogram of groups of pixels to look for correlations. As a simple example of the difference between a DMS and a Markov source, consider first a bilevel DMS with a priori probabilities $p[0] = 0.2$ and $p[1] = 0.8$. In words, the output of the source is more likely a 1 (white) than a zero. A FAX output is a bilevel image that might have this probability function. The Shannon entropy of this source is:

$$I = -0.2 \log_2 [0.2] - 0.8 \log_2 [0.8] = 0.7219 \text{ bits per pixel}$$

Now consider the probabilities of the four cases for two pixels generated by the DMS. Since the pixels are independent, the probability that the current pixel is a “1” given that the previous pixel is a “1” is just the product of the probabilities – there is no influence of the previous pixel on the choice of the current pixel:

$$f[n] = [1, 1] : \quad p[1, 1] = p[1] \cdot p[1] = 0.8 \cdot 0.8 = 0.64$$
Similarly, the probability that the current pixel is a “0” given that the previous pixel is a “0” is:

\[
\begin{align*}
  f[n] = 0 & : p\begin{bmatrix}0 & 0\end{bmatrix} = p[0] \cdot p[0] = 0.2 \cdot 0.2 = 0.04 \\
  f[n] = 1 & : p\begin{bmatrix}1 & 0\end{bmatrix} = p[1] \cdot p[0] = 0.8 \cdot 0.2 = 0.16 \\
  f[n] = 0 & : p\begin{bmatrix}0 & 1\end{bmatrix} = p[0] \cdot p[1] = 0.8 \cdot 0.2 = 0.16
\end{align*}
\]

Note that the sum of these conditional probabilities is still unity, as required. The entropy of the conditional probabilities is:

\[
I = -0.64 \log_2 [0.64] - 0.04 \log_2 [0.44] - 0.16 \log_2 [0.16] - 0.16 \log_2 [0.16] = 1.4438 \text{ bits per element}
\]

Since there are two pixels per element, the entropy of the pixels taken two at a time is just:

\[
I = 1.4438 \text{ bits per pair} / 2 \text{ pixels per pair} = 0.7219 \text{ bits per pixel}
\]

which is identical to the scalar entropy of the DMS. In words, the information in the pixels from the DMS taken two at a time is identical to that from the DMS taken one at a time; there is no additional compression because there is no interpixel correlation.

In a realistic imaging situation, we might expect that black and white pixels will be grouped together in the message. Thus the probabilities \(p\begin{bmatrix}1 & 1\end{bmatrix}\) and \(p\begin{bmatrix}0 & 0\end{bmatrix}\) would be larger in a real image that for a discrete memoryless source. To ensure that the sum of the probabilities is unity, \(p\begin{bmatrix}0 & 1\end{bmatrix}\) and \(p\begin{bmatrix}1 & 0\end{bmatrix}\) would be expected to decrease, and also to be the same, since we would expect the same number of transitions from black to white as from white to black. A possible table of probabilities from a first-order Markov source with \(p[0] = 0.2\) and \(p[1] = 0.8\) would be:

\[
\begin{align*}
  p\begin{bmatrix}1 & 1\end{bmatrix} &= 0.8 \\
  p\begin{bmatrix}0 & 0\end{bmatrix} &= 0.16 \\
  p\begin{bmatrix}0 & 1\end{bmatrix} &= p\begin{bmatrix}1 & 0\end{bmatrix} = 0.02
\end{align*}
\]

The resulting entropy of the pixels taken two at a time is:

\[
I = -0.80 \log_2 [0.80] - 0.16 \log_2 [0.16] - 0.02 \log_2 [0.02] - 0.02 \log_2 [0.02] = 0.9063 \text{ bits per element} = 0.4632 \text{ bits per pixel}
\]

There is a significant reduction in the entropy of 0.7291 bits per pixel for this example of a first-order Markov source. The additional compression is due to interpixel
correlation.

The concept may be extended to higher-order entropies. If the source is second-order Markov, the gray value of a pixel is influenced by those of two adjacent pixels. In this case, the set of 3-D vectors defined by triplets of gray values would be clustered, and 3-D vector coding will further reduce the number of bits.

As an example of the effect of spatial correlations on image compression, consider first the 4 × 2-bit image shown:

\[
\begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 1 & 2 & 2 \\
0 & 1 & 2 & 3 \\
1 & 2 & 2 & 3 \\
\end{pmatrix}
\]

\[
f[n, m] =
\]

\[
H[f] = [5, 4, 5, 2], \quad p[f] = \begin{bmatrix}
\frac{5}{16} & \frac{4}{16} & \frac{5}{16} & \frac{2}{16}
\end{bmatrix}
\]

\[
I[f] = - \left( \frac{5}{16} \log_2 \left( \frac{5}{16} \right) + \frac{1}{4} \log_2 \left( \frac{1}{4} \right) + \frac{5}{16} \log_2 \left( \frac{5}{16} \right) + \frac{1}{8} \log_2 \left( \frac{1}{8} \right) \right)
\]

\[
= 1.924 \text{ bits/pixel}
\]

Because the histogram of this image is approximately "flat," little (if any) benefit would be obtained by using a Huffman code. But note that adjacent pairs of pixels exhibit some correlation; if a pixel is dark (0 or 1), it is more likely that its neighbor to the right is dark than bright. This is demonstrated by constructing the two-dimensional histogram of the pairs of left-right pixels of \(f[n, m]:\)

\[
H[L, R] =
\]

<table>
<thead>
<tr>
<th>Gray Values</th>
<th>LEFT</th>
<th>SIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RIGHT 1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SIDE 2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that each pixel appears in only one pair; there is no double-counting. The sum of the elements of the 2-D histogram is 8, and thus there are 16 pixels (8 pairs) in \(f[n, m].\) Also note that there is a total of 16 possible elements in the 2-D histogram, four times as many as in the 1-D histogram; the 2-D histogram of a 2-bit image is a 4-bit image. Even so, if the 2-D histogram of pixel pairs is clustered, it may be possible to encode the representation in fewer bits. The entropy of this specific 2-D
histogram is:

\[
I[H[L, R]] = -\frac{3}{8} \log_2 \left[ \frac{3}{8} \right] + 3 \cdot \left( -\frac{1}{8} \log_2 \left[ \frac{1}{8} \right] \right) - \frac{2}{8} \log_2 \left[ \frac{2}{8} \right] = 2.156 \text{ bits/element}
\]

Because there are two pixels per element, the information content per pixel in the 2-D histogram is:

\[
\frac{2.156 \text{ bits}}{2 \text{ pixel}} = 1.078 \text{ bits/pixel} < 1.924 \text{ bits/pixel}
\]

The reduction in entropy of the image is due to the correlations between neighboring gray values. As before, the elements of the 2-D histogram may be coded by a Huffman procedure. This approach is called vector coding or block coding, because each pixel pair describes a two-dimensional vector of gray levels. In analogous fashion, the original Huffman approach to encode individual pixels is sometimes called scalar coding.

Of course, if it is feasible to encode larger blocks, which is equivalent to constructing gray-level vectors with more dimensions. For example, consider the 4 × 4 3-bit image:

\[
\begin{array}{cccc}
0 & 1 & 0 & 1 \\
2 & 3 & 2 & 3 \\
0 & 1 & 0 & 1 \\
2 & 3 & 2 & 3 \\
\end{array}
\]

If we encode 2 × 2 blocks of four pixels, the four resulting four-dimensional vectors are identical. Because the 4-D histogram has only one populated bin, the information content of the vector image is 0 bits per block, or 0 bits per pixel. Of course, we must also transmit the gray-level formation of the block (i.e., the codebook that specifies the gray values assigned to each code). This represents the necessary overhead of the code. In this case, the number of required bits is determined by the codebook alone.

Note that the effectiveness of a vector code depends strongly on gray-level correlations of the image. A vector code that is effective for one image (e.g., a human face) may be ridiculously ineffective for a different type of image (e.g., an aerial image of Kuwait). Conversely, a vector code that is appropriate for a particular type of image will likely be so for images of objects in the same class. Also note that if the vector histogram is flat (i.e., approximately equal populations in each bin), then there will be no reduction in storage requirements obtained by using the Huffman code.

**Example – Entropy of the English Alphabet**


J.R. Pierce, *An Introduction to Information Theory*


The simplest messages in the English language may be written with 26 letters (one
20.4 LOSSLESS COMPRESSION

Case) and the space. If these 27 characters were equally probable, the information content in a message would be:

\[
I = \sum_{i=1}^{27} - \left( \frac{1}{27} \right) \log_2 \left( \frac{1}{27} \right) = - \log_2 \left( \frac{1}{27} \right) = + \log_2 [27]
\]

\[
\approx 4.755 \text{ bits per character}
\]

A sample of typical text with equal probabilities is:

XFOML RXKHRJFFJUJ ZLPWCWKCYJFFJ EYVKCQSGHYD QPAAMKB

Of course, the probabilities of English characters are not equal. The histogram of the characters may be determined from a statistical study of words. Abramson gives the following table of character occurrences:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Probability</th>
<th>Symbol</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0642</td>
<td>N</td>
<td>0.0574</td>
</tr>
<tr>
<td>B</td>
<td>0.0127</td>
<td>O</td>
<td>0.0632</td>
</tr>
<tr>
<td>C</td>
<td>0.0218</td>
<td>P</td>
<td>0.0152</td>
</tr>
<tr>
<td>D</td>
<td>0.0317</td>
<td>Q</td>
<td>0.0008</td>
</tr>
<tr>
<td>E</td>
<td>0.1031</td>
<td>R</td>
<td>0.0484</td>
</tr>
<tr>
<td>F</td>
<td>0.0208</td>
<td>S</td>
<td>0.0514</td>
</tr>
<tr>
<td>G</td>
<td>0.0152</td>
<td>T</td>
<td>0.0796</td>
</tr>
<tr>
<td>H</td>
<td>0.0467</td>
<td>U</td>
<td>0.0228</td>
</tr>
<tr>
<td>I</td>
<td>0.0575</td>
<td>V</td>
<td>0.0083</td>
</tr>
<tr>
<td>J</td>
<td>0.0008</td>
<td>W</td>
<td>0.0175</td>
</tr>
<tr>
<td>K</td>
<td>0.0049</td>
<td>X</td>
<td>0.0013</td>
</tr>
<tr>
<td>L</td>
<td>0.0321</td>
<td>Y</td>
<td>0.0164</td>
</tr>
<tr>
<td>M</td>
<td>0.0198</td>
<td>Z</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(space)</td>
<td>0.1859</td>
</tr>
</tbody>
</table>
Using this “1-D” histogram of probabilities (i.e., assuming that the characters occur “independently” but are taken from this histogram of probabilities), the entropy of an English-language message would be somewhat less than for equally likely occurrences:

$$I = \sum_{i=1}^{27} (-p_i \log_2 [p_i]) \approx 4.080 \text{ bits per character}$$

A sample of text that might be selected from this probability distribution is:

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA L

Just as obviously, we know that English characters do not occur “independently”; the probability of occurrences of a character depends on the particular characters in the preceding sequence. The next most realistic description is based on the frequency of occurrence of pairs (“digrams”) of characters, and thus on the 2-D histogram of $27^2 = 729$ pairs. We know that some combinations (e.g., “QX”, “ZJ”) occur very rarely if at all, and so these “bins” of the 2-D histogram will be unoccupied. Therefore, the 2-D histogram of digrams is “clustered” and thus we expect some characters to be “predictable”. Therefore the information content of those characters will be decreased. The histogram of character frequencies may be computed from statistical digram frequency tables that were constructed by cryptographers. Shannon computed the resulting entropy to be:

$$I(27 \text{ characters as pairs}) = \sum_{i=1}^{27} \sum_{j=1}^{27} (-p[i,j] \log_2 [p[i,j]]) \approx 3.56 \text{ bits per character}$$
Typical text selected from the 2-D histogram is:

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
ACHIN D ILONASIVE TUCOOWE AT

Note that this text appears to be (perhaps) slightly more “intelligible” than that selected from the independent histogram – some characters almost form words! (e.g., inctore, deamy)

To continue the idea, consider the computation of entropy based on “trigrams”, or triplets of characters. Shannon computed the entropy:

\[
I(27 \text{ characters as triplets}) = \sum_{i=1}^{27} \sum_{j=1}^{27} \sum_{k=1}^{27} (-p[i|j|k] \log_2 [p[i|j|k]]) \approx 3.3 \text{ bits per character}
\]

A typical sample of such text is:

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME
OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Note that as more characters are included in a group for statistical computation of the entropy, the more the typical text resembles English. Shannon continued the process and estimated the upper and lower bounds of the entropy per character for groups of English letters up to 15 characters and also for 100 characters. The entropy per character approaches a limit for more than 9 or so letters in the interval \( 1 \leq I \leq 2 \) bits per character, but drops to the range \( 0.6 \leq I \leq 1.3 \) bits for groups of 100. This means that characters in long strings of English characters are correlated; the 100-dimensional histogram of groups of 100 characters exhibits clustering.

Of course, messages could be constructed from tables of word frequencies instead of character frequencies. A message based on first-order word frequencies (one word at a time) is:

REPRESENTING AND SPEEDILY IS AN GOOD APTOR COME CAN DIFFERENT
NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME
TO FURNISHES THE LINE MESSAGE HAD BE THESE

and a message using second-order frequencies is:

“THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT
THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR
THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR
AN UNEXPECTED”
20.4.5 Other Flavors of Huffman Coding

Modified Huffman Codes

In many cases of both message and image compression, the codebook is quite large but includes many symbols with very small probabilities. It often is useful to combine the very unlikely codes into a single symbol ELSE, which is encoded by the Huffman process and is transmitted along with the actual binary code for the character or gray level. A variation of these scheme is used for digital facsimile.

Adaptive Huffman Codes

Another modification of Huffman coding allows the process to adapt to varying statistics, but the algorithms are complicated to implement.

20.4.6 Arithmetic Coding

IBM Jour. Res. & Dev., 32(6), November 1988
Gonzalez and Woods, pp. 348-9
Rabbani and Jones, §3.5

Like the Huffman code, the result of an arithmetic code is a sequence of variable-length code symbols, but the symbols are not assigned to pixels (or blocks thereof) which were quantized to fixed numbers of bits. In other words, the arithmetic code for a group of gray levels is not restricted to an integer number of bits. For example, consider encoding a bitonal image (such as a FAX), where one level (probably white) is much more likely than the other. Because there are only two levels, it is not possible to improve on one bit per pixel, even using a Huffman code. Instead, the arithmetic code is a tree code, where one codeword is assigned to each string of input pixels of some fixed length. The generated code word for this string is a representation of an interval on the real line within the range \([0,1)\) whose length is proportional to the likelihood of the string. If the string to be coded is lengthened, then the corresponding subinterval becomes shorter and requires more bits to distinguish it from its neighbors. A slight change in the sequence can result in a significantly different codeword. The algorithm for generating an arithmetic code is most easily described by example. Consider a sequence of 1-bit characters where the frequencies of 0 and 1 are 3/4 and 1/4, respectively. One example of such a sequence is 0010. A unit interval is divided into subintervals based on the order of occurrences of characters (quantized pixels) in the sequence. If the next character in the message is 0, the bottom 3/4 of the interval will be selected; if 1, the upper 1/4 will be selected. At the start the full interval is:

\[0 \leq x_0 < 1\]

The first character in the message is “0” with known frequency of 3/4; the unit interval is shrunk to the lower 3/4:

\[0 \leq x_1 < \frac{3}{4}, \ |x_1| = \frac{3}{4}\]
The second character also is “0”, and the interval is subdivided to the lower 3/4 again:

\[0 \leq x_2 < \frac{9}{16}, \quad |x_2| = \frac{9}{16}\]

The third character is “1” with frequency \(\frac{1}{4}\), so the next subinterval is the upper \(\frac{1}{4}\) of \(x_2\):

\[
\frac{9}{16} - \frac{1}{4} \cdot \frac{9}{16} = \frac{27}{64} \leq x_3 < \frac{9}{16}, \quad |x_3| = \frac{9}{64}
\]

\[(0_\triangle 011011)_2 \leq (0_\triangle 1)_2 < (0_\triangle 10000111)_2\]

where the symbol “\(\triangle\)” is the “binary point” (analogous to the “decimal point”; it separates the bits for positive and negative powers of 2). The last character in the message is “0”, so the interval is subdivided to the lower 3/4:

\[
\frac{27}{64} = 0.421875 \leq x_4 < \frac{27}{64} + \left(\frac{3}{4} \cdot \frac{9}{64}\right) = \frac{135}{256} = 0.5273475, \quad |x_4| = \frac{27}{256}
\]

Any point in the subinterval \(x_4\) can be used as the code for the string, because that \(x_4\) can only be obtained by that specific sequence of \(m\) input characters.

Because the probabilities are multiples of powers of two in this specific example, the length of the subinterval is easily represented as a binary fraction. Though this is not always true, it is useful to continue the analysis to show how the code may be represented. Recall that fractional numbers may be represented in binary notation by breaking up into inverse powers of 2, so the bit closest to the binary point represents \(2^{-1}\), the second bit represents \(2^{-2}, \ldots\). The endpoints of the interval \(x_4\) in this example are:

\[
\frac{27}{64} = 0 + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{0}{128} + \frac{0}{256} = (0_\triangle 01101100)_2 \rightarrow (0_\triangle 0110111)_2
\]

\[
\frac{135}{256} = 0 + \frac{0}{4} + \frac{0}{8} + \frac{0}{16} + \frac{0}{32} + \frac{1}{64} + \frac{1}{128} + \frac{1}{256} = (0_\triangle 1000011)_2
\]

\[(0_\triangle 0110111)_2 \leq x_4 < (0_\triangle 10000111)_2\]

The arithmetic code for the sequence “0010” is produced by selecting the representation of ANY binary fractional number in the interval \(x_4\), though the shortest binary fractional number will give the shortest code. In this example, the binary fractional number \((0_\triangle 1)_2\) could be selected because it lies between the endpoints:

\[(0_\triangle 0110111)_2 \leq (0_\triangle 1)_2 < (0_\triangle 10000111)_2\]

The binary sequence to the right of the binary point is the encoded sequence, which in this case is the one-bit number 1; four letters have been encoded with a single bit.

As a second example, consider encoding of the sequence “1000” with the same a
priori probabilities. The first subinterval is the upper fourth of the unit interval:

\[
\frac{3}{4} \leq x_1 < 1, |x_1| = \frac{1}{4}
\]

The second subinterval is the lower 3/4 of \(x_1\):

\[
\frac{3}{4} \leq x_2 < \frac{3}{4} + \frac{3}{4} \cdot \frac{1}{4} = \frac{15}{16}, |x_2| = \frac{3}{16}
\]

The third subinterval is the lower 3/4 of \(x_2\):

\[
\frac{3}{4} \leq x_3 < \frac{3}{4} + \left(\frac{3}{4}\right)^2 \cdot \frac{1}{4} = \frac{57}{64}, |x_3| = \frac{9}{64}
\]

The final subinterval is the lower 3/4 of \(x_4\):

\[
\frac{3}{4} \leq x_4 < \frac{3}{4} + \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4} = \frac{219}{256}, |x_4| = \frac{9}{256}
\]

The binary codes for the endpoints are:

\[
\frac{1}{2} + \frac{1}{4} = (0\triangle 11000000)_2 \leq x_4 < \frac{1}{2} + \frac{1}{4} + \frac{0}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{1}{128} + \frac{1}{256} = (0\triangle 110111111)_2
\]

The code for the subinterval is the lower limit \((0\triangle 11)_2\), thus encoding the four-bit sequence with two bits.

Endpoints for and lengths of other possible sequences are:

“0000” \(\rightarrow [0, (0.75)^4] = [0, 0.31640625) \rightarrow |x_4| = (0.75)^4
\]

“1111” \(\rightarrow [1 - (0.25)^4, 1) = [0.99609375, 1) \rightarrow |x_4| = (0.25)^4 = 0.0039 \ldots
\]

“1010” \(\rightarrow [0.890625, 0.92578125) \rightarrow |x_4| = 0.03515625
\]

“1011” \(\rightarrow [0.92578125, 0.9375) \rightarrow |x_4| = 0.01171875
\]

Notice that strings of more frequent characters yield longer subintervals, while strings of rare characters result in short subintervals. If the histogram of input characters is flat, then the intervals derived by the arithmetic code to representing a string of characters will be of equal length and will require equal numbers of bits to represent them. If the histogram is clustered so that some characters are more likely to occur, then strings of frequently occurring characters will be mapped to longer subintervals within \([0,1)\), and may be represented by short codes which indicate a location in the interval.

To decode the sequences, the a priori probabilities must be known. The first two
examples will be demonstrated, where the codes were binary strings. First, the binary points are added on the left of the code to yield:

\[ x = (0\triangle1)_2 = \frac{1}{2} \]

for the first example. The decision point which determines the first character is located at \( t_1 = \frac{3}{4} \), so that \( x < t_1 \) and specifying that the first character must have probability \( \frac{3}{4} \), i.e., it is 0. The decision point for the second character is the lower \( 3/4 \) of the interval determined by \( t_1 \) i.e., \( t_2 = \frac{9}{16} \). Again \( x \leq t_2 \), so the second character also must be 0. The decision point for the third character is located at:

\[ t_3 = \frac{3}{4} \cdot t_2 = \frac{27}{64} < x = \frac{1}{2} \]

Because the coded point is larger than \( t_3 \), the third character is a 1 with probability equal to 0.25. The final decision point is divides the upper quarter of the interval in the proportion 3:1:

\[ t_4 = \frac{135}{256} > x \]

So the fourth character also is 0.

In the second example, the code is

\[ x = (0\triangle11)_2 = \frac{3}{4} \]

The decision point which determines the first character is located at \( t_1 = \frac{3}{4} \), so that \( x = t \). By the convention specified for the unit interval, a point at the threshold is in the upper subinterval. This specifies that the first character has probability \( \frac{1}{4} \) and is 1. The decision point for the second character divides the first subinterval in proportion 3:1, so that:

\[ t_2 = \frac{3}{4} + \left(\frac{3}{4} \cdot \frac{1}{4}\right) = \frac{15}{16} \]

Because \( x < t_2 \), the second character is 0. The decision point for the third character is located at:

\[ t_3 = \frac{3}{4} + \left(\left(\frac{3}{4}\right)^2 \cdot \frac{1}{4}\right) = \frac{57}{64} \]

Because \( x < t_3 \), the third character is 0. The fourth threshold divides the fourth interval in the same ratio, and is located at:

\[ t_4 = \frac{3}{4} + \left(\left(\frac{3}{4}\right)^3 \cdot \frac{1}{4}\right) = \frac{219}{256} \]

The coded point \( x < t_4 \), so the fourth character is 0.

An arithmetic code may be generated by redrawing the chart of the Huffman coding process to make a Huffman decision tree, where each bit represents the result.
of a binary decision:

A decision which selects a code symbol from a set of symbols is decomposed into a sequence of binary decisions. The codeword may be considered to be the pointer to the gray level being coded; the binary codeword defines the steps taken to reach the symbol; the more probable the sequence, the wider the interval of the pointer, and the shorter the decision sequence to obtain that codeword. The advantage of arithmetic coding is that it can be adaptive, i.e., the code can adjust to changes in the relative probabilities of symbols. The adaptive arithmetic coder assumes some a priori probability distribution of characters and updates the distribution after sending (or receiving) each new character.

This is the basis for the Q-coder, which was developed by IBM to encode binary sequences using a 12-bit register. The coder derives a robust estimate of the probability distribution as it reads the source symbols. Because of its adaptability, the Q-coder is effective for nonstationary sources.

20.4.7 Dictionary-Based Compression

Huffman and arithmetic coding use the statistics of the message/image (or a model of the statistics) to construct shorter codes for frequently occurring characters/gray levels. In other words, single gray levels (or blocks of gray levels) are represented by strings of varying length. A different kind of compression encodes variable-length strings as single characters, or tokens. In a sense, this is the type of coding used for postal ZIP codes – 5-digit numbers are used to represent individual post offices whose names and addresses require strings of different lengths. Another example is the common cryptograph where a word may be coded by its coordinates in the dictionary (page number and position on the page).

Dictionary-based compression replaced frequently occurring characters/strings/gray levels with symbols, often called tokens. This may be done statically, where the entire message is scanned for recurring strings to construct the dictionary, which will be optimum for a specific message and which must be communicated to the decoder. Adaptive codes may be constructed to build up the dictionary as the message is scanned, and the dictionary can adapt to varying statistics of the message characters. In this case, the size of the dictionary must be enlarged to allow new symbols to be
entered as the statistics change. A common application is the compression of data before sending it to magnetic tape/disk. One nice side effect is an increase in the effective transfer rate of the tape. The QIC (Quarter-Inch Cartridge) tape standard, a common backup medium for computer disks, uses a dictionary-based compression scheme which intermixes plain text and dictionary symbols by adding a one-bit symbol to distinguish the two.

Though these dictionary-based compression schemes were introduced for storing text files, they may be used for storing image or graphics files as well; they just assume that the binary files are, in fact, ASCII data. The GIF image compression scheme developed by Compuserve is a dictionary-based algorithm that was designed for graphics files and images.

**Lempel-Ziv-Welch (LZW) Coding**

During the discussion of Huffman coding, it became obvious that it is desirable to encode an image while it is being read, without knowledge of either the histogram of the final image or the correlations among neighboring pixels. A simple method for this type of real-time encoding was developed by Lempel and Ziv (*A universal algorithm for sequential data compression*, IEEE Trans. Info. Thry. 23, 337-343, 1977 and *Compression of individual sequences via variable-rate coding*, IEEE Trans. Info. Thry. 24, 530-536, 1978), with a later extension by Welch (*A Technique for high-performance data compression*, IEEE Computer 17, 8, 1984). Compression based these works are referred to as LZ77, LZ78, and LZW, respectively. LZ77 allowed a 4 KByte dictionary of symbols for the codebook; matches within the text with data strings already seen are encoded as fixed-length pointers. LZ77 uses a large sliding text window (several thousand characters) that scans over the text, viewing a large block of recently coded text and a small block of text to be coded. Text in the look-ahead buffer is compared to the dictionary to find matches. This may take much time during the compression step, but decompression is not so constrained. The length of the longest possible match is limited by the size of the look-ahead buffer.

LZ78 abandoned the concept of the text buffer and built up the dictionary of character strings in increments of one character. The encoded strings may be very long, thus allowing significant compression if strings repeat themselves frequently. The LZW process is used in several of the common PC shareware utilities for compressing data files (*e.g.* PKARC, PKZIP, PAK). The ARC scheme was introduced in 1985 and dominated compression of MS-DOS files for several years, in no small part due the fact that it was available as shareware. LZW generates a fairly efficient code as pixel gray levels are read, without prior knowledge of the image statistics. For an m-bit image (\(2^m\) gray levels), the LZW implementation is:

1. Select the number of bits \(k\) in the codeword, for a total of \(2^k\) codewords. The number \(k\) must be greater than \(m\) (preferably, \(k \gg m\)).

2. Initialize the codes by setting the first \(2^m\) codes to the available gray levels — this ensures that legitimate codes exist for all pixels, even after the code table is filled.
3. Read the gray level of the first pixel; it is the first element in the string “S” and is a member of the code table (from step 2).

4. If all pixels have been read, then output the code for “S” (k bits) and quit.

5. If pixels are left, then read the gray level $P$ of the next pixel and append to “S” to create string “SP”.

6. If “SP” is in the table of codewords, then set “S” = “SP” and go to Step 4, otherwise continue.

7. Append the codeword for “S” (k bits) to the coded image.

8. If there are unused codewords left, then add “SP” to the table.

9. Reset the string “S” = “P” and go to step 5.

Consider this simple example of a 36-pixel 2-bit image ($m = 2$, 4 levels $\alpha, \beta, \gamma, \delta$), which requires 72 bits to transmit as is:

$$f[n] = \alpha\alpha\beta\alpha\gamma\alpha\alpha\delta\alpha\alpha\beta\alpha\alpha\gamma\alpha\alpha\beta\alpha\alpha\alpha\alpha\alpha\alpha\beta\alpha\alpha\gamma$$

Assume that $k = 4$, so that the codebook contains $2^4 = 16$ symbols. From step 2, the first four codes are $A \equiv \alpha$, $B \equiv \beta$, $C \equiv \gamma$, and $D \equiv \delta$. The coding process proceeds in this manner:

1. $S_1 = \alpha$, in codebook as “A”
   $P_1 = \alpha \rightarrow S_1P_1 = \alpha\alpha$, not in codebook,
   $S_2 \rightarrow P_1 = \alpha$

2. $P_2 = \beta$,
   $S_2P_2 = \alpha\alpha$, already exists in codebook as symbol “E”
   $S_3 \rightarrow S_2P_2 = \alpha\alpha$

3. $P_3 = \gamma$, $S_3P_3 = \alpha\alpha\gamma$, not in codebook
   $S_4 \rightarrow S_3P_3 = \alpha\alpha\gamma$

4. $P_4 = \alpha$, $S_4P_4 = \beta\alpha$, not in codebook
   $S_5 \rightarrow P_4 = \alpha$

5. $P_5 = \gamma$, $S_5P_5 = \alpha\alpha\gamma$ exists in codebook as “E”
   $S_6 \rightarrow \alpha\alpha$

6. $P_6 = \gamma$, $S_6P_6 = \alpha\alpha\gamma$, not in codebook
   $S_7 \rightarrow \gamma$
After all pixels have been interrogated (and if I made no mistakes), the codebook is:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(\alpha)</td>
</tr>
<tr>
<td>B</td>
<td>(\beta)</td>
</tr>
<tr>
<td>C</td>
<td>(\gamma)</td>
</tr>
<tr>
<td>D</td>
<td>(\delta)</td>
</tr>
<tr>
<td>E</td>
<td>(\alpha\alpha)</td>
</tr>
<tr>
<td>F</td>
<td>(\alpha\alpha\beta)</td>
</tr>
<tr>
<td>G</td>
<td>(\beta\alpha)</td>
</tr>
<tr>
<td>H</td>
<td>(\alpha\alpha\gamma)</td>
</tr>
<tr>
<td>I</td>
<td>(\gamma\alpha)</td>
</tr>
<tr>
<td>J</td>
<td>(\alpha\alpha\delta)</td>
</tr>
<tr>
<td>K</td>
<td>(\delta\alpha)</td>
</tr>
<tr>
<td>L</td>
<td>(\alpha\alpha\alpha)</td>
</tr>
<tr>
<td>M</td>
<td>(\alpha\beta)</td>
</tr>
<tr>
<td>N</td>
<td>(\beta\alpha\alpha)</td>
</tr>
<tr>
<td>O</td>
<td>(\alpha\alpha\gamma\alpha)</td>
</tr>
<tr>
<td>P</td>
<td>(\alpha\alpha\beta\alpha)</td>
</tr>
</tbody>
</table>

The entire coded message is “AEBECDEAGHFPLEFH”. Note that new elements are added to the codebook quite rapidly, and the later elements typically represent longer and longer sequences. When the codebook is full, the last element (or the least used) can be deleted, and the process can proceed with the available elements. The coded image requires 17 characters at 4 bits per character, for a total of 68 bits, which is not much of a reduction when compared to the original quantity of 72 bits.

If a 3-bit code had been used, the last character in the codebook would be \(H\), and the coded message would be AEBECDEAGECFEEEFH, for a total of 18 3-bit characters, or 54 bits. Note that the total number of bits for the 3-bit code is less than for the 4-bit code (which is not the usual case). This illustrates the sensitivity of the process to the local image statistics.

As mentioned above, the LZW algorithm was used in most PC file compressors (archiving programs such as “PKARC” and “PKZIP”) back in the days of small disk drives. In those applications, the files consisted of 1-bit ASCII characters. LZW with a 13-bit codeword was known as “squashing”, while 12-bit LZW was “crunching”.
20.5 Transform Coding

The various lossless compression schemes (Huffman, arithmetic, and LZW coding) are useful for compressing images with clustered histograms (scalar or vector). However, the compression ratios are fixed by the statistics of the images. In many cases, “lossy” coding is useful where nonredundant information is discarded. If the loss of this information is not objectionable to the viewer, then the reduction in storage requirements may well be beneficial. Probably the most familiar example of transform compression is JPEG encoding.

One convenient method for implementing lossy coding is to first construct a new representation of the original image via an invertible image transformation. Such a transformation may be a gray-scale remapping (lookup tables) for monochrome images, a color-space transformation for color images, or a shift-invariant or shift-variant spatial transformation. These transformations “reorganize” the gray values of the image and thus change the correlation properties of the image. It is possible to compress the reorganized gray values using one (or more) of the schemes already considered.

In the discussion of image processing operations, we have seen that images may be recast into a different (and possibly equivalent) form by an image transformation. The general form for the transformation of a 1-D vector is:

\[ F[\ell] = \sum_n f[n] m[\ell,n] \]

where \( m[\ell,n] \) is the 2-D “reference” or “mask” function of the transform. A familiar such function is the Fourier transform, where \( m[\ell,n] = \exp[-2\pi i \ell N] \). If the transform is invertible, then there exists a mask function \( M[n,\ell] \) such that:

\[ f[n] = \sum_\ell F[\ell] M[n,\ell] \]

The transformation is a space-invariant convolution if the mask function \( m[\ell,n] \) has the form of a convolution kernel: \( m[\ell-n] = h[\ell-n] \). In the study of linear systems, it is demonstrated that convolution with an impulse response \( h[n] \) is invertible if the transfer function \( H[k] \) is nonzero everywhere. A space-variant transformation is invertible if the set of mask functions \( m[\ell,n] \) is complete. In either case, the gray level \( F \) at pixel \( \ell_1 \) of the transformed image is a measure of the similarity of the input image \( f[n] \) and the specific mask \( m[\ell_1,n] \). As such, the transformed image pixels \( F[\ell_1] \) are measures of the correlation of \( f[n] \) and \( m[\ell_1,n] \); the greater the similarity between image and mask, the larger the amplitude of \( F \) at \( \ell_1 \).

The goal of transform coding is to generate an image via an invertible transform whose histogram is less flat (more clustered) than that of the original image, thus having less entropy and requiring fewer bits for storage/transmission. Many authors describe the operation as compacting the image information into a small number of coefficients (corresponding to pixels of the transformed image), though I prefer the
picture of the transform as generating an image with a more clustered histogram. Though transform coding often clusters the histogram, the lunch is not free; we pay by increasing the number of possible gray levels to be encoded and thus the number of bits per recorded pixel if the image is to be recovered without loss. In the example of encoding the image derivative considered below, the transformed image \( F[\ell] \) may occupy up to twice as many levels as the input. If the transform \( F[\ell] \) is requantized to the same number of levels before storage/transmission, some information is lost and \( f[n] \) cannot be recovered perfectly. This is one example of *lossy* coding where the compression efficiency can be quite high. If the statistics of the source can be quantified (e.g., discrete memoryless source, Markov source, etc.), it is possible to quantify the effect of reducing the number of encoded levels on the fidelity of the final image. These studies form a branch of information theory which is known as *rate distortion theory*.

Block diagram of transform compression. The original image \( f[n,m] \) is converted to a different coordinate system via the invertible transformation \( T\{f[n,m]\} = F[k,l] \), which is then quantized and Huffman coded. The quantization step ensures that the recovered image is generally an estimate \( \hat{f}[n,m] \)

To introduce the concept of compression via image transformations, consider a 64-pixel, 6-bit, 1-D image \( f[n] \) of a linear ramp in the interval \(-32 \leq n \leq 31\), where \( n \) is the pixel address:

\[
f[n] = n + 32; \quad 32 \leq n \leq 31, \quad 0 \leq f[n] \leq 63
\]
This histogram of this image is flat with a population of one pixel per bin, and therefore the information content of the image is 6 bits per pixel.

1-D 6-bit “ramp” image \( f[n] = n + 32 \) for \(-32 \leq n \leq 31\). The histogram is “flat” with one count per gray level. The information content is 6 bits per pixel.

Because the slope of \( f[n] \) is constant, the image of its derivative is constant also. Recall that discrete differentiation can be implemented as a convolution:

\[
\frac{df}{dn} = f[n] * h[n]
\]

\[
h[n] = \begin{bmatrix} +1 & -1 & 0 \end{bmatrix}
\]

For this derivative operation, the discrete transfer function (discrete Fourier transform of the impulse response \( h[n] \)) is:

\[
H[k] = |k|
\]

and is nonzero at all samples \( k \) except at the origin (i.e., at zero frequency — the constant part or average value). The derivative of a 1-D image is invertible if the average (or initial) value is known as a boundary condition. In this example, the histogram of the derivative has only two occupied gray levels, and the entropy of the derivative image is 0.116 bits/pixel. This transformed image can be encoded by one-bit characters with a compression efficiency of \( \eta = \frac{6}{0.116} \approx 52 \), which means that this coding scheme produced a bit rate that is very much smaller than the Shannon limit. How is this possible? Because we encoded a quality of groups of characters rather than the individuals.
Derivative of the ramp image: $f'[n] = f[n] - f[n-1]$; the “gray value” of the first pixel is “0” and all of the rest are “1”. The histogram has 63 pixels at 1 and 1 pixel at 0, for an information content of 0.116 bits per pixel.

Obviously, derivative images can have negative gray values (though it did not in this case). In fact, if the dynamic range of $f[n]$ is 64 levels in the interval $[0,63]$ (6 bits per pixel), the theoretical dynamic range of the derivative image is 127 levels in the interval $[-63,63]$ (not quite 7 bits per pixel). So although the histogram of the derivative image may have less entropy if the pixels of the original image are well correlated, an additional bit per pixel must be stored if the image is to be recovered without loss. If some error in the uncompressed image is tolerable, sparsely populated levels in the transform may be substituted with a similar gray level, or levels with small amplitudes may be quantized to 0. These can significantly reduce the information content. The latter process is called coring by some in the image compression community.

The process of encoding the derivative image rather than the original is the basis for both run-length encoding and differential pulse code modulation (DPCM). These are examples of predictive coding, where a reduction in entropy is obtained by transmitting the difference between the actual gray value at a pixel and a prediction obtained by some rule. In run-length encoding of images with large uniform regions (e.g., binary text images for FAX machines), the transmitted data are the number of consecutive pixels with the same gray level before the next switch. If the strings of 0s and 1s are long, run-length encoding can reduce the data stream significantly. In DPCM, the gray value at a pixel is predicted from some linear combination of previous gray levels; the error $\epsilon$ between the actual gray level and the prediction is quantized and encoded. The predictors of the pixel gray value $f[x,y]$ may include one or several previous pixels, including $f[x-1,y], f[x-1,y-1], f[x,y-1], f[x+1,y-1]$, as shown below:

<table>
<thead>
<tr>
<th>$f[x-1,y-1]$</th>
<th>$f[x-1,y]$</th>
<th>$f[x,y]$</th>
<th>$f[x,y-1]$</th>
<th>$f[x,y]$</th>
<th>$f[x+1,y-1]$</th>
<th>$f[x+1,y]$</th>
<th>$f[x+1,y+1]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f[x-1,y]$</td>
<td>$f[x,y]$</td>
<td>$f[x,y]$</td>
<td>$f[x,y]$</td>
<td>$f[x,y]$</td>
<td>$f[x,y]$</td>
<td>$f[x,y]$</td>
<td>$f[x,y]$</td>
</tr>
</tbody>
</table>
The predictor may be expressed as a linear combination of these levels:

\[ f[x, y] = \sum_{i,j} a_{ij} f[x - i, y - j] \]

where the \( a_{ij} \) are the weights of the linear combination. Among the predictors commonly used are:

1. 1-D first-order (based on one pixel): \( f[x_0, y_0] = f[x_0 - 1, y_0] \)
2. 2-D second-order (based on two pixels): \( f[x_0, y_0] = \frac{1}{2} (f[x_0 - 1, y_0] + f[x_0, y_0 - 1]) \)
3. 2-D third-order (based on three pixels):

\[ f[x_0, y_0] = \frac{1}{4} (3 f[x_0 - 1, y_0] - 2 f[x_0 - 1, y_0 - 1] + 3 f[x_0, y_0 - 1]) \]

In the first case, the difference between the actual and predicted gray values is just the discrete first derivative:

\[ \epsilon[x_0, y_0] = f[x_0, y_0] - f[x_0 - 1, y_0] = f[x_0, y_0] - f[x_0 - 1, y_0] = \left. \frac{\partial f}{\partial x} \right|_{x=x_0} \]

The 2-D predictor usually improves compression efficiency significantly over 1-D predictors for real images.

In adaptive DPCM, the mathematical form of the predictor may vary based on the image structure. The compression in DPCM results because the prediction error is quantized to fewer levels than the original image data. Note that the final image may exhibit grainy or contouring errors if the minimum quantization level is coarse and the gray levels vary slowly across the image.

### 20.5.1 Color-Space Transformations

\[
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
0.596 & -0.274 & -0.322 \\
0.211 & -0.523 & 0.312
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
= 
\begin{bmatrix}
Y \\
I \\
Q
\end{bmatrix}
\]

The “luminance” \( Y \) is a weighted sum of the color-value triplet \([R, G, B]\). Note that the weights sum to one, which means that a gray pixel with the values \( R = G = B \) will have the same luminance value.

\[ Y = 0.299R + 0.587G + 0.114B \]

\( I \) and \( Q \) are the “chrominance” values and are weighted differences of the color-value triplet, where the weights sum to zero. \( I \) is a weighted sum of green and blue subtracted from a weighted red; in other words, it may be thought of as Red - Cyan. The \( Q \) channel is weighted green subtracted from a weighted combination of red and blue, or Magenta - Green.
20.5.2 Space-Variant Transformations

Most authors consider only space-variant operations as transforms for image coding; for 1-D images, the form of the space-variant operation is:

\[ F[k] = \sum_{n} f[n] \ p[n; k] \]

while for 2-D images, the form of the space-variant operation is:

\[ F[k, \ell] = \sum_{n,m} f[n, m] \ p[n, m; k, \ell] \]

In words, the 2-D transform is the product of a 4-D matrix and the 2-D input. The most common such transform in imaging is the discrete Fourier transform (DFT); the mask \( m[n, k] \) for the 1-D DFT is the set of 1-D complex-valued sinusoids with spatial frequency \( k/N \):

\[
p[n; k] = \exp \left[ -\frac{i2\pi nk}{N} \right] = \cos \left( \frac{2\pi nk}{N} \right) - i \sin \left( \frac{2\pi nk}{N} \right)
\]

The mask \( p[n, m; k, \ell] \) for 2-D images is the set of 2-D complex-valued sinusoids; they vary in a sinusoidal fashion along one direction and are constant in the orthogonal direction. For an \( N \times N \) input image, the mathematical expression for the mask function is:

\[
p[n, m; k, \ell] = \exp \left[ -\frac{2\pi i (nk + m\ell)}{N} \right] = \cos \left[ 2\pi \left( nk + \frac{m\ell}{N} \right) \right] - i \sin \left[ 2\pi \left( nk + \frac{m\ell}{N} \right) \right]
\]

The spatial frequencies of the sinusoidal mask indexed by \( k \) and \( \ell \) are respectively \( \xi = \frac{k}{N} \) and \( \eta = \frac{\ell}{N} \). The gray level of each pixel in the transformed image \( F[k, \ell] \) describes the degree of similarity between \( f[n, m] \) and that specific 2-D sinusoid. Recall that the amplitudes of all pixels in a 1-D sinusoid can be completely specified by three numbers: the magnitude, spatial frequency, and phase. In other words, the gray value of a particular sample of a sinusoid is determined completely by any other pixel if the parameters of the sinusoid are known; a perfect interpixel correlation exists among the amplitudes. The DFT operation compresses the image information into the number of bits required to represent those three parameters. Therefore, an image composed of a small number of sinusoidal components can be compressed to a very significant degree by converting to its Fourier representation. We begin by considering a simple 1-D case; the image to be compressed has 64 pixels and is...
a 1-D sinusoid of period 32 pixels, as shown below. The image has been sampled but not quantized, i.e., all real numbers between 0 and 31 are allowed gray levels. Because the Shannon entropy (information content) is defined as a sum over discrete probabilities (gray levels), it strictly can not apply to images with real-valued gray levels; the image must be quantized to calculate the entropy. The histogram of the sinusoid after quantizing to 64 bins is shown; note that it is approximately flat:

\[ f[n] = 16 + 15 \cos \left( \frac{2\pi n}{32} \right), 0 \leq f[n] \leq 31 \]

One definition of the 1-D DFT is:

\[ F[\xi] = F[k \cdot \Delta\xi] \implies F[k] = \sum_{n=-N}^{N-1} f[n] \exp \left[ -\frac{2\pi i nk}{N} \right] \]

Note that the transform often includes a scale factor of \( N^{-1} \) that is not included here. In words, the transform is the sum of gray values of the product of the input image and the mask function, which is real-valued in the interval \([-1, +1]\). In general, the transform \( F[k] \) generates noninteger values that typically lie outside the dynamic range of \( f[n] \). The amplitude of the transform at \( k = 0 \) is the sum of the gray values of the image. In the example under consideration which has a mean gray value of 16, the DC value of the DFT is:

\[ F[k = 0] = 64 \text{ pixels} \cdot 16 \text{ (mean gray)} = 4096 \]

The entire discrete spectrum \( F[k] \) has 61 samples with value zero, two with value 480 (located at \( k = \pm 2 \) so that \( \xi = \pm \frac{k}{N} = \pm \frac{1}{32} \)) and one with value 4096 at \( k = 0 \). The histogram of \( F[k] \) generally needs an infinite number of bins to account for the continuously valued range of \( F[k] \), but is certainly much less flat than the histogram of \( f[n] \); thus there is less entropy in \( F[\xi] \).

The sinusoid after quantization to 32 gray levels (5 bits per pixel) and the resulting histogram are shown below. The original image entropy is 3.890 bits/pixel. Since the image is quantized, it no longer is a sampled pure sinusoid and the Fourier transform includes extra artifacts. The histogram of the quantized transform indicates that the information content of the transform is only 0.316 bits/pixel.

The 1-D discrete Fourier transform of an \( N \)-pixel real-valued array is an \( N \)-pixel complex-valued array. Because each complex-valued pixel is represented as a pair of real-valued numbers, the DFT generates twice as much data as the input array. However, the DFT of a real-valued array is redundant, meaning that values in the resulting DFT array are repeated. The real part of the DFT of a real-valued function is even and the imaginary part is odd. Therefore, half of the data in the DFT of a real-valued array may be discarded without loss.
20.5.3 Block DFT Coding via DCT

Because the statistical properties of typical images vary across the image (i.e., the statistics are shift variant), it is common to apply the selected transform to local blocks of pixels (often of size $8 \times 8$ or $16 \times 16$) that are coded individually using a spatial transform (to reduce gray-level redundancy), followed by Huffman coding. The discrete Fourier transform (DFT) is a possible such transform, but its advantage of generating an equivalent image which exhibits a clustered histogram often is offset somewhat by its assumption that an input array of size $N \times M$ pixels actually is infinite array that is periodic over $N \times M$ blocks of pixels. In words, the DFT assumes that the gray value of a pixel off the edge on one side of the array is identical to the gray value on the edge of the other side of the array. Of course, it is common for pixels at opposite edges of the array to have different gray levels (for sky at the top and for ground at the bottom, for example). The gray-level transitions at the boundaries of the periods of the array generate artifacts in the DFT known as leakage, or false frequency components. Though these false frequency terms are necessary to generate the true sharp edge boundaries of the block, the additional samples of the DFT with non-zero amplitudes increase redundancy (and thus the entropy) of the transformed image. Therefore, the potential efficiency of compression using the DFT often suffers.

An additional problem arises because the DFT of a real-valued discrete input image is a complex-valued Hermitian array. The symmetry of the transform array (even real part and odd imaginary part) ensures that only half of each part need be stored, but the histograms of both parts must be accounted when computing the entropy of the transform coefficients.

To reduce the impact of the sharp transitions at the edges of the blocks, as well as to obtain a transform that is real-valued for a real-valued input image, the discrete cosine transform (DCT) may be used instead of the DFT. The DCT has become very important in the image compression community, being the basis transformation for the JPEG and MPEG compression standards. The DCT of an $M \times M$ block may be viewed as the DFT of a synthetic $2M \times 2M$ block that is created by replicating the original $M \times M$ block after folding about the vertical and horizontal edges:
The original $4 \times 4$ block of image data is replicated 4 times to generate an $8 \times 8$ block of data via the DFT format and an $8 \times 8$ DCT block by appropriate reversals. The transitions at the edges of the $4 \times 4$ DCT blocks do not exhibit the “sharp” edges in the $4 \times 4$ DFT blocks.

The resulting $2M \times 2M$ block exhibits smaller discontinuities at the edges. The symmetries of the Fourier transform for a real-valued image ensure that the original $M \times M$ block may be reconstructed from the DCT of the $2M \times 2M$ block.

Consider the computation of the DCT for a 1-D $M$-pixel block $f[n]$ ($0 \leq n \leq M - 1$). The $2M$-pixel synthetic array $g[n]$ is indexed over $n$ ($0 \leq n \leq 2M - 1$) and has the form:

$$g[n] = f[n] \text{ for } 0 \leq n \leq M - 1$$

$$g[n] = f[2M - 1 - n] \text{ for } M \leq n \leq 2M - 1.$$ 

In the case $M = 8$, the array $g[n]$ is defined:

$$g[n] = \begin{cases} f[n] & \text{for } 0 \leq n \leq 7 \\ f[15 - n] & \text{for } 8 \leq n \leq 15 \end{cases}$$
The values of $g[n]$ for $8 \leq n \leq 15$ is a “reflected replica” of $f[n]$:  

\begin{align*}
  g[8] &= f[7] \\
  g[14] &= f[1] \\
  g[15] &= f[0]
\end{align*}

If the “new” array $g[n]$ is assumed to be periodic over $2M$ samples, its amplitude is defined for all $n$, e.g.,  

\begin{align*}
  g[n] &= f[-1 - n] \text{ for } -M \leq n \leq -1 \implies -16 \leq n \leq -1 \\
  g[n] &= f[n + 2M] \text{ for } -2M \leq n \leq -M - 1 \implies -32 \leq n \leq -17
\end{align*}

Note that the 16-sample block $g[n]$ is NOT symmetric about the origin of coordinates because $g[-1] = g[0]$; to be symmetric, $g[-\ell]$ would have to equal $g[+\ell]$. For example, consider a 1-D example where $f[n]$ is an 8-pixel ramp as shown:

The $2M$-pixel representation of $f[n]$ is the $g[n]$ just defined:

\[
  g[n] = \begin{cases} 
    f[n] & \text{for } 0 \leq n \leq 7 \\
    f[2M - 1 - n] & \text{for } 8 \leq n \leq 15 
  \end{cases}
\]

If this function were symmetric (even), then circular translation of the 16-point array by 8 pixels to generate $g[n - 8 \mod 16]$ also be an even function.

From the graph, it is apparent that the translated array is not symmetric about the origin; rather, it has been translated by $-\frac{1}{2}$ pixel from symmetry in the $2M$-pixel array. Thus define a new 1-D array $c[n]$ that is shifted to the left by $\frac{1}{2}$ pixel:

\[
  c[n] = g\left[n - \frac{1}{2}\right]
\]

This result may seem confusing at first; how can a sampled array be translated by $\frac{1}{2}$ pixel? For the answer, consider the continuous Fourier transform of a sampled array translated by $\frac{1}{2}$ unit:

\[
  \mathcal{F}_1 \{c[n]\} \equiv C[\xi] = \mathcal{F}_1 \left\{ g\left[x - \frac{1}{2}\right] \right\} = \mathcal{F}_1 \left\{ g\left[x\right] \ast \delta\left[x - \frac{1}{2}\right] \right\} = G[\xi] \ast \exp\left[-2\pi i \xi \cdot \frac{1}{2}\right] = G[\xi] \ast \exp\left[-i\pi \xi \right]
\]

\[
  C[\xi] = G[\xi] \cdot \exp\left[-i\pi \cdot \frac{1}{2}\right]
\]
Thus the effect of translation by $\frac{1}{2}$ pixel on the transform is multiplication by the specific linear-phase factor:

$$\exp[-i\pi\xi] = \cos[\pi\xi] - i \sin[\pi\xi].$$

The 2M-point DFT of the symmetric discrete array (original array translated by $\frac{1}{2}$ pixel) has the form:

$$F_{2M}\{c[n]\} = F_{2M}\left\{g\left[n - \frac{1}{2}\right]\right\} = F_{2M}\left\{g[n] \ast \delta\left[n - \frac{1}{2}\right]\right\}
= G[k] \cdot \exp\left[-i\pi\left(\frac{k}{2M}\right)\right] C[k] = G[k] \cdot \exp\left[-i\pi\frac{k}{2M}\right]
= G[k] \cdot \left(\cos\left[\frac{\pi k}{2M}\right] - i \sin\left[\frac{\pi k}{2M}\right]\right)$$

where the continuous spatial frequency $\xi$ has been replaced by the sampled frequency $\frac{k}{2M}$. This function $C[k]$ is the DCT of $f[n]$. Because the 2M-pixel translated function $c[n]$ is real and even, so must be the 2M-pixel discrete spectrum $C[k]$; therefore only $M$ samples of the spectrum are independent. This array is the DCT of $f[n]$.

**Steps in Forward DCT**

To summarize, the steps in the computation of the 1-D DCT of an $M$-pixel block $f[n]$ are:

1. create a 2M-pixel array $g[n]$ from the $M$-pixel array $f[n]$:

   $$g[n] = f[n] : 0 \leq n \leq M - 1$$
   $$g[n] = f[2M - 1 - n] : M \leq n \leq 2M - 1$$

2. compute the 2M-point DFT of $g[n] = G[k]$

3. the $M$-pixel DCT $C[k] = \exp[-i\pi\frac{k}{2M}] \cdot G[k]$ for $0 \leq k \leq M - 1$

The entire process may be cast into the form of a single equation, though the algebra required to get there is a bit tedious,

$$C[k] = \sum_{n=0}^{M-1} 2 f[n] \cos\left(\pi k \cdot \frac{2n + 1}{2M}\right) \quad \text{for} \quad 0 \leq k \leq M - 1$$

**Steps in Inverse DCT**

The inverse DCT is generated by applying the procedures in the opposite order:
20.5 TRANSFORM CODING

1. create a $2M$-pixel array $G[k]$ from the $M$-pixel DCT $C[k]$: 
\[ G[k] = \exp\left[ +\frac{i\pi k}{2M} \right] \cdot C[k] \text{ for } 0 \leq k \leq M - 1 \]
\[ G[k] = -\exp\left[ +\frac{i\pi k}{2M} \right] \cdot C[2M - k] \text{ for } M + 1 \leq k \leq 2M - 1 \]

2. compute the inverse $2M$-pixel DFT of $G[k] \to g[n]$

3. $f[n] = g[n]$ for $0 \leq n \leq M - 1$

The single expression for the inverse DCT is:
\[ f[n] = \frac{1}{M} \sum_{k=0}^{M-1} w[k] \cdot C[k] \cos\left(\pi k \cdot \frac{2n + 1}{2M}\right) \text{ for } 0 \leq n \leq M - 1 \]
where $w[k] = \frac{1}{2}$ for $k = 0$ and $w[k] = 1$ for $1 \leq k \leq M - 1$.

**Forward DCT OF 2-D Array**

The corresponding process to compute the 2-D DCT $C[k, \ell]$ of an $M \times M$-pixel block $f[n, m]$ is:

1. create the $2M \times 2M$-pixel $g[n, m]$: 
\[ g[n, m] = f[n, m] : 0 \leq n, m \leq M - 1 \]
\[ g[n, m] = f[2M - 1 - n, m] : M - 1 \leq n \leq 2M - 1, 0 \leq m \leq M - 1 \]
\[ g[n, m] = f[n, 2M - 1 - m] : M - 1 \leq n \leq 2M - 1, 0 \leq m \leq M - 1 \]
\[ g[n, m] = f[2M - 1 - n, 2M - 1 - m] : M - 1 \leq n, m \leq 2M - 1 \leq k, \ell \leq M - 1 \]

2. compute the $2M$-point DFT of $g[n, m] \to G[k, \ell]$

3. $C[k, \ell] = \exp\left[ -\frac{i\pi(k+\ell)}{2M} \right] \cdot G[k, \ell] \text{ for } 0 \leq k, \ell \leq M - 1$
CHAPTER 20 IMAGE COMPRESSION

Inverse DCT OF 2-D Array

1. create a $2M \times 2M$ array $G[k, \ell]$ from the DCT $C[k, \ell]$:

   \[
   G[k, \ell] = \exp\left[\frac{i\pi (k + \ell)}{2M}\right] \cdot C[k, \ell] : 0 \leq k, \ell \leq M - 1
   \]

   \[
   G[k, \ell] = -\exp\left[\frac{i\pi (k + \ell)}{2M}\right] \cdot C[2M - n, m] : M - 1 \leq n \leq 2M - 1, 0 \leq m \leq M - 1
   \]

   \[
   G[k, \ell] = -\exp\left[\frac{i\pi (k + \ell)}{2M}\right] \cdot C[n, 2M - m] : M - 1 \leq n \leq 2M - 1, 0 \leq m \leq M - 1
   \]

   \[
   G[k, \ell] = \exp\left[\frac{i\pi (k + \ell)}{2M}\right] \cdot C[2M - n, 2M - m] : M - 1 \leq k, \ell \leq 2M - 1
   \]

2. compute the inverse $2M$-pixel DFT of $G[k, \ell] \rightarrow g[n, m]$

3. $f[n, m] = g[n, m]$ for $0 \leq n, m \leq M - 1$

The form of the forward DCT may be written more simply as:

\[
F[k, \ell] = 4 \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} w[n] w[m] f[n, m] \cos \left(\frac{(2n + 1) k \pi}{2N}\right) \cos \left(\frac{(2n + 1) \ell \pi}{2N}\right)
\]

where:

\[
w[j] \equiv \begin{cases} 
\frac{1}{\sqrt{2}} & \text{if } j = 0 \\
1 & \text{if } j = 1, \ldots, N - 1
\end{cases}
\]

and the corresponding form of the inverse 2D DCT is:

\[
f[n, m] = \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} w[k] w[\ell] F[k, \ell] \cos \left(\frac{(2k + 1) k \pi}{2N}\right) \cos \left(\frac{(2\ell + 1) \ell \pi}{2N}\right)
\]

The action of the 2-D $8 \times 8$ block DCT will be detailed for blocks in the $64 \times 64$ 5-bit image LIBERTY.
The image and the $8 \times 8$ block DCT are shown. Note the bright pixel in each $8 \times 8$ block; its amplitude is proportional to the average gray value of that block. The numerical values of the block and the DCT will be shown for two cases. In the first, the block is taken from the upper right where all gray values are white:

$$
\begin{bmatrix}
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
31 & 31 & 31 & 31 & 31 & 31 & 31 & 31 \\
\end{bmatrix}
$$
The corresponding $8 \times 8$ DCT block is:

\[
\begin{bmatrix}
248 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

The amplitude of the upper-left pixel in the DCT is the zero-frequency (DC) term; its amplitude is eight times the average gray value of the block. The other terms in the DCT are proportional to the amplitude of oscillating sinusoidal components and often are called the AC terms; in this constant block, the oscillating terms have null amplitude because all gray values in the block are equal.

The gray values of a second $8 \times 8$ block located near the center of the image are:

\[
\begin{bmatrix}
9 & 7 & 21 & 12 & 11 & 13 & 15 & 13 \\
31 & 8 & 9 & 13 & 9 & 12 & 12 & 11 \\
31 & 22 & 7 & 10 & 13 & 11 & 11 & 10 \\
23 & 8 & 10 & 7 & 10 & 13 & 10 & 11 \\
11 & 31 & 31 & 26 & 9 & 8 & 9 & 10 \\
31 & 31 & 31 & 14 & 14 & 20 & 9 & 8 \\
31 & 31 & 28 & 10 & 29 & 31 & 31 & 30 \\
31 & 31 & 21 & 18 & 31 & 31 & 30 & 29 \\
\end{bmatrix}
\]

with an average gray value of 17.95.
The amplitudes of the DCT of the block near the center of the image are approximately (rounded to one decimal place):

\[
\begin{bmatrix}
143.6 & 14.6 & 9.7 & 5.7 & -3.3 & -1.1 & 4.1 & 2.9 \\
-30.9 & -0.1 & -0.9 & -0.7 & 5.0 & 6.5 & 1.1 & -1.0 \\
13.9 & -9.3 & 1.3 & 3.7 & 0.3 & -0.4 & 0.5 & 3.0 \\
-1.5 & 3.8 & -5.7 & -8.6 & -6.1 & 0.3 & 2.9 & 0.5 \\
-4.3 & -4.7 & -5.2 & -6.1 & -1.2 & 0.5 & -0.9 & -0.6 \\
0.3 & -5.8 & 2.2 & 3.2 & -1.1 & -1.8 & -2.0 & 3.7 \\
-1.5 & 4.7 & -1.0 & 0.5 & -1.1 & -1.6 & -0.7 & -1.0 \\
5.9 & -0.1 & -1.8 & -5.4 & -2.4 & -2.4 & -4.2 & -0.8 \\
\end{bmatrix}
\]

Again, the amplitude of the sample in the upper-left corner is eight times the average gray value of 17.95. The other 63 coefficients (the AC terms) are bipolar. A negative AC coefficient means that the particular cosine term oscillates out of phase by \(\pi\) radians. The rate of oscillation of an AC component increases with distance from the DC term, and the direction of oscillation is determined by the orientation relative to the DC term at the origin; cosines that oscillate horizontally are specified by the amplitudes along the first row and those that oscillate vertically by the amplitudes in the first column. Note that the amplitudes of higher-frequency AC components (away from the upper-left corner) tend to be smaller than the low-frequency terms (towards the upper left). This is the usual result for realistic imagery, and is utilized to obtain additional compression in the JPEG standard.

The DCT is useful as a transform for image compression because:

1. (a) it is real valued,

(b) the amplitudes are proportional to the amplitudes of sinusoids with different oscillation rates, and

(c) the amplitudes of higher-frequency terms are smaller for realistic imagery.

Object B: \(f[n,m] = \cos[2\pi(n\xi' + m\eta')]\), \(\xi' = \frac{1}{4}\sin[\frac{\pi}{8}] = \sqrt{2}/8 \approx 0.1768\), oscillates two times in diagonal direction, period \(X' \approx 0.1768^{-1} = 5.6561\), \(\eta' = \frac{1}{4}\cos[\frac{\pi}{4}]\)

The “low-frequency” content of the signal concentrates the DCT amplitude near the origin, but the fact that the array is only pseudoperiodic over 8 samples produces larger amplitudes at more pixels in the DCT array.

Examples of Individual 8 × 8 blocks and their associated DCT arrays
Object A: $f[n, m] = \cos[2\pi \left(0n + \frac{m}{4}\right)]$, oscillates “vertically” two times in 8 samples, $\xi = 0, \eta = \frac{1}{4}$ average gray value $= 0$

Note that the DCT array is zero except in the first vertical column (horizontal frequency $= 0$ cycles per pixel)

In the next example, we compute the DCT of the individual $8 \times 8$ “basis images”. Each of these images produces an $8 \times 8$ DCT that has all pixels at zero except one. The object composed of these “blocks” is shown on the left and the DCT on the right. The low-frequency terms appear in the upper left and the highest-frequency terms to the right and bottom. This observation provides the basis for the JPEG encoding standard that is considered next.

20.6 JPEG Image Compression of Static Images

A standard has been developed by the “Joint Photographic Experts Group” and has become very common since the early 1990s. In its original form, it was based on the DCT, to the point where the method is now sometimes called “JPEG (DCT)”. This standard is based on the properties of human vision where the sensitivity of the eye generally decreases with increasing spatial frequency.
20.6 JPEG IMAGE COMPRESSION OF STATIC IMAGES

8-bit grayscale image used to illustrate JPEG image compression.

20.6.1 Example 1: “Smooth” Region

JPEG Encoding

64 × 64 block of pixels in illustration

Magnified view of 64 × 64 block of pixels
“Smooth” Region of image of “Parrots”

Gray values of $8 \times 8$ block

\[
\begin{array}{cccccccccccccccc}
110 & 111 & 111 & 110 & 111 & 109 & 112 & 117 \\
112 & 115 & 111 & 110 & 108 & 111 & 129 & 139 \\
111 & 112 & 107 & 109 & 120 & 140 & 142 & 142 \\
108 & 107 & 119 & 126 & 142 & 150 & 145 & 143 \\
111 & 131 & 137 & 141 & 152 & 148 & 147 & 140 \\
138 & 138 & 141 & 148 & 144 & 140 & 147 & 148 \\
141 & 143 & 150 & 144 & 139 & 149 & 151 & 148 \\
143 & 136 & 137 & 142 & 148 & 148 & 138 & 130
\end{array}
\]

The midgray value of 128 is subtracted from the values in the $8 \times 8$ block to produce
bipolar data

\[
f[n, m] - 128 =
\begin{array}{cccccccc}
-18 & -17 & -17 & -18 & -17 & -19 & -16 & -11 \\
-16 & -13 & -17 & -18 & -20 & -17 & 1 & 11 \\
-17 & -16 & -21 & -19 & -8 & 12 & 14 & 14 \\
-20 & -21 & -9 & -2 & 14 & 22 & 17 & 15 \\
-17 & 3 & 9 & 13 & 24 & 20 & 19 & 12 \\
10 & 10 & 13 & 20 & 16 & 12 & 19 & 20 \\
13 & 15 & 22 & 16 & 11 & 21 & 23 & 20 \\
15 & 8 & 9 & 14 & 20 & 20 & 10 & 2
\end{array}
\]

The DCTs of the 8 × 8 blocks are evaluated (data rounded to one decimal place):

\[
DCT \ F[k, \ell] \approx
\begin{array}{cccccccc}
24.4 & -49.6 & -2.4 & 2.1 & -4.6 & -0.4 & -0.2 & 0.8 \\
-92.0 & -18.0 & 18.8 & -5.6 & 1.2 & -3.9 & -1.9 & -0.2 \\
-23.7 & 34.8 & 13.7 & -2.5 & 4.2 & 3.3 & 1.1 & -1.0 \\
8.5 & 18.8 & -5.9 & -16.1 & 3.0 & 0.3 & -4.9 & 1.8 \\
-8.4 & 0.4 & -24.8 & 9.2 & 1.6 & -4.7 & 1.2 & 3.1 \\
1.8 & -5.6 & 0.5 & 4.5 & -2.6 & 2.8 & 7.3 & -0.6 \\
-4.7 & 3.0 & -0.2 & 9.2 & 3.8 & -0.1 & -0.2 & -2.0 \\
2.6 & -0.5 & -2.3 & 0.8 & -9.8 & -3.8 & -1.0 & -1.2
\end{array}
\]

If the block were pure white (gray value 255), then the DCT would produce a single nonzero amplitude of +1016 at the DC term and zero elsewhere; if the block were black, the DC coefficient of the DCT would be −1024.

The DCT of the 64 8 × 8 blocks can be displayed in pictorial form:
DCT values are divided by the values in a normalization matrix that accounts for the contrast sensitivity function of the eye. Smaller numbers imply more weighting applied. Largest numbers apply to the high-frequency terms positioned toward the lower right corner. Note that the DC component (less 128) is divided by a larger number than the neighboring low-frequency AC components.

\[
Q[u, v] = \begin{bmatrix}
16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\
12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\
14 & 13 & 16 & 24 & 50 & 57 & 69 & 56 \\
14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\
18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\
24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\
49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\
72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \\
\end{bmatrix}
\]

\[
F[k, \ell] = \frac{DCT}{N[k, \ell]} = \begin{bmatrix}
1.5 & -4.5 & -0.2 & 0.1 & -0.2 & 0.0 & 0.0 & 0.0 \\
-7.7 & -1.5 & 1.3 & -0.3 & 0.0 & -0.1 & 0.0 & 0.0 \\
-1.7 & 2.7 & 0.9 & -0.1 & 0.1 & 0.1 & 0.0 & 0.0 \\
0.6 & 1.1 & -0.3 & -0.6 & 0.1 & 0.0 & -0.1 & 0.0 \\
-0.5 & 0.0 & -0.7 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.1 & -0.2 & 0.0 & 0.1 & 0.0 & 0.0 & 0.1 & 0.0 \\
-0.1 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & -0.1 & 0.0 & 0.0 & 0.0 \\
\end{bmatrix}
\]

Normalized 8 × 8 block displayed as an image.
Quantization rounds the values in the block to the nearest integer; note the large number of null coefficients:

\[
Q \left\{ \frac{F[k,\ell]}{\overline{N}[k,\ell]} \right\} = \begin{bmatrix}
2 & -5 & 0 & 0 & 0 & 0 & 0 & 0 \\
-8 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\
-2 & 3 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

The 2-D \(8 \times 8\) is scanned in a “zig-zag” format to convert the 2-D block to a 1-D sequence of coefficients. Because of the layout of the scan, the largest coefficients should appear first for most “real” images.

The zig-zag path traveled during Huffman coding of the quantized DCT coefficients.
The DC coefficients of blocks are encoded separately using a differential coding scheme. The AC coefficients are encoded based on the entropy (Huffman code) with three pieces of information:

1. the run length of consecutive zeros that preceded the current element in the zigzag sequence.
2. the number of bits to follow in the amplitude number
3. the amplitude of the coefficient

<table>
<thead>
<tr>
<th>Bit Count</th>
<th>Amplitudes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1,+1</td>
</tr>
<tr>
<td>2</td>
<td>-3,-2,+2,+3</td>
</tr>
<tr>
<td>3</td>
<td>-7 to -4, +4 to +7</td>
</tr>
<tr>
<td>4</td>
<td>-15 to -8, +8 to +15</td>
</tr>
<tr>
<td>5</td>
<td>-31 to -16, +16 to +31</td>
</tr>
<tr>
<td>6</td>
<td>-63 to -32, +32 to +63</td>
</tr>
<tr>
<td>7</td>
<td>-127 to -64, +64 to +127</td>
</tr>
<tr>
<td>8</td>
<td>-255 to -128, +129 to +255</td>
</tr>
<tr>
<td>9</td>
<td>-511 to -256, +256 to +511</td>
</tr>
<tr>
<td>10</td>
<td>-1023 to -512, +512 to +1023</td>
</tr>
</tbody>
</table>

2 (DC term coded separately)  

\[ -5 \quad -8 \quad -2 \quad 1 \quad 3 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad -1 \quad 0 \times 39 \]

This string is encoded using a predetermined Huffman code based on the number of zeros in the string to the next nonzero coefficient and the numerical value of the coefficient. Short strings of zeros are encoded with shorter sequences:

\[ (0, 3) \quad (0, 4) \quad (0, 2) \quad (0, 1) \quad (0, 2) \quad (0, 1) \quad (2, 1) \quad (0, 1) \quad (10, 1) \quad (0, 1) \quad \text{EOB} \]

The Huffman code has the partial form:

\[ 100+0010 \quad 1011+0+100 \quad 01+0+ \quad \cdots \]
Recover from JPEG coding:

First we reconstruct the approximate DCT from the Huffman code and multiplication by the normalization matrix:

\[
\begin{bmatrix}
32 & -55 & 0 & 0 & 0 & 0 & 0 & 0 \\
-96 & -24 & 14 & 0 & 0 & 0 & 0 & 0 \\
-28 & 39 & 16 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

The inverse DCT produces an 8 × 8 block of bitonal “gray values:"

\[
\begin{bmatrix}
-22.6 & -16.3 & -11.2 & -13.0 & -18.5 & -19.6 & -13.6 & -6.6 \\
-12.7 & -18.4 & -24.6 & -25.7 & -19.5 & -8.7 & 1.8 & 8.1 \\
-14.2 & -22.1 & -27.2 & -20.4 & -4.3 & 10.0 & 15.5 & 15.1 \\
-23.4 & -19.1 & -8.8 & -6.1 & 19.3 & 23.9 & 19.5 & 13.5 \\
-15.1 & -2.8 & 12.9 & 22.6 & 24.1 & 21.9 & 20.9 & 21.4 \\
8.8 & 15.4 & 20.2 & 17.7 & 12.1 & 12.9 & 22.5 & 32.3 \\
20.1 & 18.8 & 16.0 & 12.7 & 11.2 & 13.4 & 18.5 & 22.8 \\
14.8 & 11.8 & 11.4 & 16.8 & 22.8 & 21.4 & 11.4 & 1.6 \\
\end{bmatrix}
\]

These values are rounded and the constant 128 is added back to obtain the recovered
block, which is compared to the original values:

\[
\hat{f}[n,m] = 
\begin{bmatrix}
105 & 112 & 117 & 115 & 110 & 108 & 114 & 121 \\
115 & 110 & 103 & 102 & 108 & 119 & 130 & 136 \\
114 & 106 & 101 & 108 & 124 & 138 & 143 & 143 \\
105 & 115 & 110 & 103 & 125 & 138 & 147 & 152 \\
110 & 109 & 119 & 122 & 147 & 148 & 148 & 142 \\
111 & 125 & 141 & 151 & 152 & 150 & 149 & 149 \\
137 & 143 & 148 & 146 & 140 & 141 & 150 & 160 \\
148 & 147 & 144 & 141 & 139 & 141 & 147 & 151 \\
142 & 140 & 139 & 145 & 151 & 149 & 139 & 130 \\
110 & 111 & 111 & 110 & 111 & 109 & 112 & 117 \\
112 & 115 & 111 & 110 & 108 & 111 & 129 & 139 \\
111 & 112 & 107 & 109 & 120 & 140 & 142 & 142 \\
\end{bmatrix}
\]

\[
f[n,m] = 
\begin{bmatrix}
108 & 107 & 119 & 126 & 142 & 150 & 145 & 143 \\
111 & 131 & 137 & 141 & 152 & 148 & 147 & 140 \\
138 & 138 & 141 & 148 & 144 & 140 & 147 & 148 \\
141 & 143 & 150 & 144 & 139 & 149 & 151 & 148 \\
143 & 136 & 137 & 142 & 148 & 148 & 138 & 130 \\
\end{bmatrix}
\]
The error in the block is obtained by subtracting the recovered image from the original:

\[
\varepsilon[n, m] = f[n, m] - \hat{f}[n, m] =
\begin{bmatrix}
+3 & -2 & 0 & +4 & -5 & -2 & -3 & +1 \\
-2 & +6 & -4 & -10 & 0 & -2 & -2 & -9 \\
+1 & -5 & -7 & +2 & +4 & -1 & -3 & -12 \\
-7 & -4 & +6 & +3 & 0 & +8 & +4 & -3 \\
+1 & -4 & -2 & -3 & -1 & -1 & 0
\end{bmatrix}
\]

### 20.6.2 Example 2: “Busy” Image

Consider the central 8 × 8 block of the “Liberty” image, with gray values:

\[
f[n, m] =
\begin{bmatrix}
160 & 184 & 96 & 160 & 192 & 136 & 168 & 152 \\
176 & 152 & 160 & 144 & 160 & 88 & 80 & 128 \\
56 & 64 & 120 & 168 & 120 & 72 & 64 & 200 \\
112 & 56 & 40 & 56 & 136 & 120 & 48 & 192 \\
168 & 128 & 40 & 56 & 136 & 120 & 48 & 192 \\
176 & 40 & 112 & 24 & 88 & 120 & 64 & 160 \\
152 & 48 & 168 & 48 & 16 & 120 & 8 & 8 \\
48 & 176 & 160 & 128 & 16 & 32 & 96 & 104
\end{bmatrix}
\]

Subtract the constant:

\[
f[n, m] - 128 =
\begin{bmatrix}
32 & 56 & -32 & 32 & 64 & 8 & 40 & 24 \\
48 & 24 & 32 & 16 & 32 & -40 & -48 & 0 \\
-72 & -64 & -8 & 40 & -8 & -56 & -64 & 72 \\
-16 & -72 & -88 & -72 & 8 & -8 & -80 & 64 \\
40 & 0 & -88 & -72 & 8 & -8 & -80 & 64 \\
48 & -88 & -16 & -104 & -40 & -8 & -64 & 32 \\
24 & -80 & 40 & -80 & -112 & -8 & -120 & -120 \\
-80 & 48 & 32 & 0 & -112 & -96 & -32 & -24
\end{bmatrix}
\]
Evaluate the $8 \times 8$ DCT

$$F[k, \ell] = \begin{bmatrix}
-164.0 & 40.5 & 79.7 & -56.8 & 86.0 & -31.2 & 77.4 & -32.3 \\
161.7 & -55.2 & -38.6 & 14.4 & 90.5 & -28.3 & -83.6 & -3.2 \\
72.4 & 89.6 & -70.4 & -27.5 & -107.1 & 27.4 & -71.4 & 35.3 \\
60.4 & 8.6 & 48.5 & 156.9 & 8.5 & -35.0 & 35.9 & 54.3 \\
34 & -39 & 45.1 & -14.1 & 0 & -24.3 & -105.3 & -66.5 \\
-70 & -16.2 & -16.3 & 12.4 & -14.8 & 68.9 & 46.6 & -34.4 \\
28.5 & -92.1 & 0.6 & -50.2 & 18.4 & -12.1 & -33.6 & 8.7 \\
13.6 & 0.3 & -40.6 & 19.2 & 0.7 & -43.7 & -7.2 & 9.4 \\
\end{bmatrix}$$

Normalize by the weighting matrix

$$\frac{F[k, \ell]}{N[k, \ell]} = \begin{bmatrix}
-10.3 & 3.7 & 8.0 & -3.6 & 3.6 & -0.8 & 1.5 & -0.5 \\
13.5 & -4.6 & -2.8 & 0.8 & 3.5 & -0.5 & -1.4 & -0.1 \\
5.2 & 6.9 & -4.4 & -1.1 & -2.7 & 0.5 & -1.0 & 0.6 \\
4.3 & 0.5 & 2.2 & 5.4 & 0.2 & -0.4 & 0.4 & 0.9 \\
1.9 & -1.8 & 1.2 & -0.3 & 0.0 & -0.2 & 0.1 & 0.0 \\
-2.9 & -0.5 & -0.3 & 0.2 & -0.2 & 0.7 & 0.4 & -0.4 \\
0.6 & -1.4 & 0.0 & -0.6 & 0.2 & -0.1 & -0.3 & 0.1 \\
0.2 & 0.0 & -0.4 & 0.2 & 0.0 & -0.4 & -0.1 & 0.1 \\
\end{bmatrix}$$

Round to nearest integer

$$\text{integer} \left\{ \frac{F[k, \ell]}{N[k, \ell]} \right\} = \begin{bmatrix}
-10 & 4 & 8 & -4 & 4 & -1 & 2 & -1 \\
13 & -5 & -3 & 1 & 3 & 0 & -1 & 0 \\
5 & 7 & -4 & -1 & -3 & 0 & -1 & 0 \\
4 & 1 & 2 & 5 & 0 & 0 & 0 & 1 \\
2 & -2 & 1 & 0 & 0 & 0 & -1 & -1 \\
-3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}$$
Renormalize by multiplying by $N [k, \ell]$:

$$
\hat{F} [k, \ell] = N [k, \ell] \cdot \text{integer} \left\{ \frac{F [k, \ell]}{N [k, \ell]} \right\} = \begin{bmatrix}
-160 & 44 & 80 & -64 & 96 & -40 & 102 & -61 \\
156 & -60 & -42 & 19 & 78 & 0 & -60 & 0 \\
70 & 91 & -64 & -24 & -120 & 0 & -69 & 56 \\
56 & 17 & 44 & 145 & 0 & 0 & 0 & 62 \\
36 & -44 & 37 & 0 & 0 & 0 & -103 & -77 \\
-72 & 0 & 0 & 0 & 0 & 104 & 0 & 0 \\
48 & -64 & 0 & -87 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Calculate the inverse $8 \times 8$ DCT

$$
\hat{f} [n, m] - 128 = \begin{bmatrix}
42 & 45 & -14 & 41 & 41 & -4 & 59 & 10 \\
30 & 21 & 19 & -6 & 52 & 1 & -79 & 8 \\
-56 & -72 & 21 & 45 & -47 & -47 & -86 & 96 \\
-11 & -79 & -104 & 26 & -99 & -43 & -74 & 58 \\
44 & -17 & -104 & -48 & -28 & 8 & -81 & 70 \\
39 & -65 & 6 & -83 & -36 & -8 & -81 & 32 \\
31 & -70 & 13 & -87 & -105 & -52 & -105 & -113 \\
-85 & 64 & 37 & -12 & -100 & -61 & -72 & -3
\end{bmatrix}
$$

Add back the constant:

$$
\hat{f} [n, m] = \begin{bmatrix}
170 & 173 & 114 & 169 & 169 & 124 & 187 & 138 \\
158 & 149 & 147 & 122 & 180 & 129 & 49 & 136 \\
72 & 56 & 149 & 173 & 81 & 81 & 42 & 224 \\
117 & 49 & 24 & 154 & 29 & 85 & 54 & 186 \\
172 & 111 & 24 & 80 & 100 & 136 & 47 & 198 \\
167 & 63 & 134 & 45 & 92 & 120 & 47 & 160 \\
159 & 58 & 141 & 41 & 23 & 76 & 23 & 15 \\
43 & 192 & 165 & 116 & 28 & 67 & 56 & 125
\end{bmatrix}
$$
The error is the difference.

\[ \varepsilon[n, m] = f[n, m] - \hat{f}[n, m] = \begin{bmatrix}
-10 & 11 & -18 & -9 & 23 & 12 & -19 & 14 \\
18 & 3 & 13 & 22 & -20 & -41 & 31 & -8 \\
-16 & 8 & -29 & -5 & 39 & -9 & 22 & -24 \\
-5 & 7 & 16 & -98 & 107 & 35 & -6 & 6 \\
-4 & 17 & 16 & -24 & 36 & -16 & 1 & -6 \\
9 & -23 & -22 & -21 & -4 & 0 & 17 & 0 \\
-7 & -10 & 27 & 7 & -7 & 44 & -15 & -7 \\
5 & -16 & -5 & 12 & -12 & -35 & 40 & -21 
\end{bmatrix} \]

Note that the error is much larger in several of the locations, because the high-frequency coefficients that are necessary to produce the “sharp edges” have been quantized to zero.